On the probability–quality paradox in language generation

Clara Meister[🤯] Gian Wiher[🤯] Tiago Pimentel[🥱] Ryan Cotterell[🤯]
[🤯]ETH Zürich [🥱]University of Cambridge
clara.meister@inf.ethz.ch gian.wiher@inf.ethz.ch
tp472@cam.ac.uk ryan.cotterell@inf.ethz.ch

Abstract

When generating natural language from neural probabilistic models, high probability does not always coincide with high quality: It has often been observed that mode-seeking decoding methods, i.e., those that produce high-probability text under the model, lead to unnatural language. On the other hand, the lower-probability text generated by stochastic methods is perceived as more human-like. In this note, we offer an explanation for this phenomenon by analyzing language generation through an information-theoretic lens. Specifically, we posit that human-like language should contain an amount of information (quantified as negative log-probability) that is close to the entropy of the distribution over natural strings. Further, we posit that language with substantially more (or less) information is undesirable. We provide preliminary empirical evidence in favor of this hypothesis; quality ratings of both human and machine-generated text—covering multiple tasks and common decoding strategies—suggest high-quality text has an information content significantly closer to the entropy than we would expect by chance.

1 Introduction

Today's probabilistic neural language models are often trained on millions—if not billions—of lines of human text; thus, at least at an intuitive level, we would expect high-probability generations to be human-like. Yet the high-quality[1] texts these models have become famous for producing (Brown et al., 2020; Clark et al., 2021) are usually not those assigned the highest probability by the model (Fan et al., 2018; Holtzman et al., 2020; Basu et al., 2021; DeLucia et al., 2021). Rather, the relationship between probability and quality appears to have an inflection point,[2] i.e., quality and probability are positively correlated only until a certain threshold, after which the correlation becomes negative. While the existence of such a trend has received informal explanations (see, e.g., Ippolito et al. (2019) and Zhang et al. (2021) for a qualitative discussion about the trade-off between diversity and quality), it lacks a more fundamental understanding. Why does the lower probability text produced by stochastic decoding methods—such as nucleus or top-k sampling—outperform text generated using probability-maximizing approaches? In this note, we take an information-theoretic approach in an attempt to answer this question.

In information theory, probability has another interpretation: its negative log quantifies **information content**. In the context of natural language, the notion of information content is intuitive; humans use strings as a means to convey information. Further, less predictable text, i.e., text which would be harder for us to anticipate, conveys *more* information. If we assume that the goal of human communication is to transmit messages efficiently and reliably (Gibson et al., 2019), we may predict that these strings' information content should concentrate inside a specific interval. At one extreme, strings with more-than-expected information may be hard to process, and thus ought to be disfavored when producing language.[3] At the other extreme, low-information strings may be seen as boring and uninformative.

Collectively, these concepts lead us to propose the **expected information hypothesis**: Text perceived as human-like should have an information content within a small interval around the expected information—i.e., the entropy—of natural language strings. Such a hypothesis offers

[1] We assume that "human-like" is a (necessary but not sufficient) prerequisite for "high-quality" in the context of natural language strings.

[2] The inflection point is empirically demonstrated in our App. B or in Fig. 1 of Zhang et al. (2021).

[3] Many works in psycholinguistics have shown a direct relationship between information content and processing effort (Smith and Levy, 2013; Wilcox et al., 2020, *inter alia*).

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 36 - 45
May 22-27, 2022 ©2022 Association for Computational Linguistics

an intuitive explanation for the trends observed in natural language generation (NLG), i.e., why desirable text seems to exist not always at the high end of the probability spectrum but around a certain inflection point.[4] Moreover, it also gives us a *testable* hypothesis: given a language generation model q whose entropy we can empirically estimate, we can evaluate whether high-quality text indeed has an information content that falls within an interval around this quantity.

To test our hypothesis, we perform an analysis comparing human and model-generated text, investigating multiple common decoding strategies and NLG tasks. Specifically, our analysis focuses exclusively on English text. We indeed observe that the information content of highly ranked text (as judged by humans) often falls within a standard deviation of model entropy; there is statistically significant evidence that this is not due to chance. Further, the best-performing decoding methods appear to select strings with an information content within this interval. We take these observations as empirical support for our hypothesis, helping to explain the probability–quality paradox observed in language generation.

2 Probabilistic Language Generators

In this work, we focus on probabilistic models for language generation tasks. Formally, these models are probability distributions q over natural language strings $\mathbf{y} \in \mathcal{Y}$, where $\mathcal{Y}$ is the (countably infinite) set consisting of all possible strings that can be constructed from a set vocabulary $\mathcal{V}$:

$$\mathcal{Y} \overset{\text{def}}{=} \{\text{BOS} \circ \mathbf{v} \circ \text{EOS} \mid \mathbf{v} \in \mathcal{V}^*\} \tag{1}$$

Here, BOS and EOS stand for special reserved beginning- and end-of-string tokens, respectively, and $\mathcal{V}^*$ denotes the Kleene closure of $\mathcal{V}$. In practice, we limit the set of strings we consider to $\mathcal{Y}_N \subset \mathcal{Y}$ for some maximum sequence length N.

Note that q may be a conditional model. For instance, we may model $q(\cdot \mid \mathbf{x})$ where $\mathbf{x}$ is an input text, as in the case of machine translation, or an input image, as in the case of image captioning. However, for notational brevity, we omit this explicit dependence in most of our subsequent analyses. In order to estimate q, it is standard practice to maximize the log-probability of a training corpus $\mathcal{C}$ under the model with respect to the model's parameters $\boldsymbol{\theta}$. This is equivalent to minimizing its negative log-probability:

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

There are many different decision rules one can employ for generating natural language strings from a model q; such sets of rules are generally referred to as decoding strategies; see Wiher et al. (2022) for an in-depth review. Given the probabilistic nature of the models we consider, an intuitive strategy for decoding would be to choose the string with the highest probability under q, an approach referred to as maximum-a-posteriori (MAP) decoding.[5] Yet recent research has shown that solutions to MAP decoding—or, even more generally, to heuristic mode-seeking methods such as beam search—are often not high-quality, even in state-of-the-art NLG 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). Similarly, in the domain of open-ended generation, mode-seeking methods produce dull and generic text (Holtzman et al., 2020).

Where maximization has failed, authors have turned to stochastic methods, taking random samples from q. While the resulting text is often assigned much lower probability than the mode, it can be qualitatively much better. This peculiarity has puzzled the language generation community for the last few years, with only qualitative intuitions being offered as explanation. This paper in turn offers a quantitative explanation.

3 Language as Communication

While many aspects of natural language may not perfectly adhere to Shannon's mathematical theory of communication, there are several characteristics of human language that *can* fruitfully be described using an information-theoretic framework.[6] Here we employ this framework for explaining recent phenomena observed in probabilistic NLG.

[4] Similar ideas have been used to improve language models and language generation before (Meister et al., 2020; Wei et al., 2021).

[5] Note that MAP decoding is somewhat of a misnomer since we are not maximizing over a Bayesian posterior. Nonetheless, the term has become commonplace in the language generation literature.

[6] A large body of work has explored the extent to which attributes of human languages—such as word lengths or phoneme distributions—can be explained as information-theoretic design features (Gibson et al., 2019). Surprisal theory, for instance, directly relates human language processing difficulty to information content (Hale, 2001).

3.1 Measuring Information

We can precisely compute the information content of a string given the *true* (perhaps conditional) probability distribution p over natural language strings. Fortunately, this is the exact distribution our language generation models in §2 are trained to approximate.[7] Assuming q approximates p well (as quantified by metrics such as perplexity), we may thus use it to estimate such attributes of natural language strings. In this work, we will measure the amount of information a specific realization $\mathbf{y}$ contains, which we denote $\mathrm{I}(\mathbf{y}) \overset{\text{def}}{=} -\log q(\mathbf{y})$, as well as the *expected* amount of information a random $\mathbf{y} \in \mathcal{Y}_N$ drawn from q contains, also termed the entropy of q:

$$\mathbb{E}_q\left[\mathrm{I}(\mathbf{y})\right] = \mathrm{H}(q) = -\sum_{\mathbf{y}\in\mathcal{Y}_N} q(\mathbf{y}) \log q(\mathbf{y}) \quad (3)$$

Note that Pimentel et al. (2021b, Theorem 2) prove that, as long as the probability of EOS under q is bounded below by some $\epsilon > 0$, then the entropy of q is finite. In our case we restrict q to a finite subset $\mathcal{Y}_N$ of $\mathcal{Y}$, which also implies that Eq. (3) is finite.

3.2 The Expected Information Hypothesis

Language is used as a means for transferring information. This property of language has in fact motivated several theories of language evolution; many have posited, for instance, that natural language has developed to optimize for reliable and efficient data communication, subject to cognitive resources (Zipf, 1949; Hockett, 1960; Hawkins, 2004; Piantadosi et al., 2011). The above theories arguably imply that humans tend to produce natural language strings with a certain amount of information; they also imply that, on the receiving end of communication, humans would expect similar strings. We argue that this amount is intuitively close to the language's entropy, i.e., close to the average string's information content.

Expected Information Hypothesis. *Text perceived as human-like typically encodes an amount of information close to the expected information content of natural language strings, i.e., in the interval* $[\mathrm{H}(p) - \varepsilon,\ \mathrm{H}(p) + \varepsilon]$ *for a natural language string distribution* p *and some* ε.[8] *Text that falls outside of this region is likely perceived as unnatural.*

This viewpoint can be applied to the problem of decoding neural text generators. In the context of a model q of the distribution p, this implies that—when q is a good approximation—human-like text should typically have a negative log-probability close to the entropy of q. In §4, we provide empirical evidence for this hypothesis.

Relationship to the typical set. The set of strings that we discuss has an intuitive relationship to the typical set (Shannon, 1948), an information-theoretic concept defined for stationary ergodic stochastic processes. However, generation from standard neural probabilistic language models cannot be framed as such a process.[9] While we cannot utilize the formal mathematical underpinnings of typicality, the connection can still be useful for understanding why strings with a given information content exhibit certain characteristics. An overview of the concept is in App. A for the interested reader; also see Dieleman (2020) for further insights on typicality in the context of generative models.

4 Experiments

Our experiments present an analysis of the distribution of information content in text generated by both humans and probabilistic models. Specifically, we look at the relationship between information content and quality—as measured by human judgments. We perform experiments on two natural language generation tasks: abstractive summarization and story generation. We present the results for story generation here, while the results for summarization can be found in App. B due to space constraints. A recreation of the probability versus quality plots of Zhang et al. (2021) can also be found in App. B.

We use the following Monte Carlo estimator for the entropy, i.e., expected information content, of

[7] To see this, recall that minimizing the objective in Eq. (2) is (up to an additive constant) equivalent to minimizing the Kullback–Leibler divergence—an information-theoretic quantity that measures the amount of information lost when approximating one probability distribution with another—between the empirical distribution p and our model q.

[8] While we do not offer a concrete explanation of why distributions over natural language strings have a particular entropy, we posit that it is determined by cognitive constraints, as observed with other phenomena in natural language (Coupé et al., 2019; Pimentel et al., 2021a).

[9] Specifically, most neural language models are neither stationary (due to their ability to encode arbitrarily long sequences; Welleck et al. 2020) nor ergodic (because of the absorbing nature of the EOS state). This implies that we cannot guarantee the existence of an entropy rate, which is necessary to define the typical set.

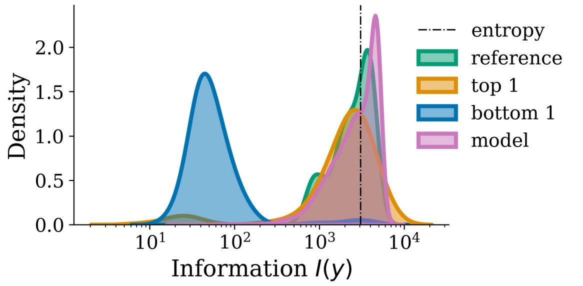


Figure 1: The distribution over information $\mathrm{I}(\mathbf{y})$ values of: MODEL, the model, as estimated using samples from q; REFERENCE, the reference strings; TOP 1 and BOTTOM 1, model-generated strings ranked first and last (respectively) among all decoding strategies by human annotators. The latter 3 are all w.r.t. a held-out test set. Same graph is reproduced for individual decoding strategies in App. B.

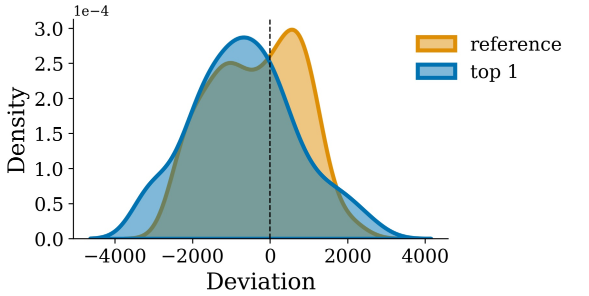


Figure 2: The distribution of the difference in total information content for (1) test-set references and (2) top-ranked model-generated strings from the (conditional) entropy of the model from which they were generated.

our model q:

$$\widehat{\mathrm{H}}(q) = \frac{1}{M}\sum_{m=1}^{M} -\log q(\mathbf{y}^{(m)}) \tag{4}$$

where we sample $\mathbf{y}^{(m)} \overset{\text{i.i.d.}}{\sim} q$. Algorithmically, taking these samples may be done in linear time using ancestral sampling. All computations are performed with the test sets of respective datasets. Note that for both abstractive summarization and story generation, where we condition on some input $\mathbf{x}$, we must compute the *conditional* entropy for each input, i.e., using $q(\cdot \mid \mathbf{x})$ instead of $q(\cdot)$. For each $\mathbf{x}$, we take $M = 100$ to estimate $\widehat{\mathrm{H}}(q(\cdot \mid \mathbf{x}))$.

4.1 Setup

Models and Data. We only conduct experiments on the English language. For story generation, we fine-tune GPT-2 (medium) (Radford et al., 2019) (checkpoint made available by OpenAI) on the WRITINGPROMPTS dataset (Fan et al., 2018). For abstractive summarization, we use BART (Lewis et al., 2020), fine-tuned on the CNN/DAILYMAIL dataset (Nallapati et al., 2016). We rely on the open-sourced code-base from the HuggingFace framework (Wolf et al., 2020) for reproducibility.

Decoding Strategies. We explore text generated according to a number of different decoding strategies. Unless otherwise stated, we use the implementation provided by Hugging Face for each of the decoding algorithms. Along with standard ancestral sampling, we experiment with the following six decoding strategies:

- **greedy search**;
- **beam search** with beam sizes $k = 5$ and $k = 10$;
- **diverse beam search** (Vijayakumar et al., 2016) with Hamming distance as a dissimilarity function and $\lambda = 0.7$ and $G = k = 5$;[10]
- **ancestral sampling**;
- **top-k sampling** (Fan et al., 2018) with $k = 30$;
- **nucleus sampling** (Holtzman et al., 2020) with $p = 0.85$;[11]
- **minimum Bayes risk decoding** (MBR; Eikema and Aziz 2020)[12] with 32 Monte Carlo samples[13] from q and BEER (Stanojević and Sima'an, 2014) as the utility function.

Human Evaluations. We use the *prolific* platform to obtain human judgments of text quality (according to 2 criteria per task) from 5 different annotators on 200 examples per decoding strategy–per task. This gives us a total of > 3000 annotated examples. We largely follow the guidelines recommended by van der Lee et al. (2021) in setting up our evaluations: For abstractive summarization, we ask annotators to rate *quality* and *accuracy* while for story generation, annotators rate *fluency* and *naturalness*. More details on our setup can be found in App. B.1.

4.2 Results

In Fig. 1, we plot the distribution of information content assigned by q to four different sets of strings: our reference (human-generated) text, the

[10] The choice of dissimilarity function and hyperparameters (λ, G, k) is based on the recommendations from the original work.

[11] This choice is based on experiments in (DeLucia et al., 2021) that suggest a parameter range $p \in [0.7, 0.9]$.

[12] We use the github.com/Roxot/mbr-nmt framework.

[13] The number of Monte Carlo samples was chosen based on the batch size constraint.

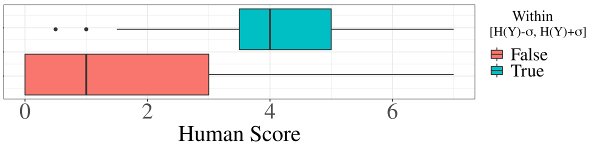


Figure 3: Human scores for strings (including both reference text and model-generated text) within 1 std of model entropy and outside of this interval. There is a statistically significant difference in means ($p < 0.001$).

top and bottom ranked (according to human annotators) strings generated from q via our different decoding strategies,[14] and strings sampled i.i.d. from q. Note that the latter should represent the distribution of negative log-probabilities assigned to strings by the model. We see that both the references and the top-ranked model-generated strings—both of which we assume are of relatively high quality—contain an amount of information clustered around the (estimated) model entropy. On the other hand, the distribution of the information content of poorly rated strings is skewed towards much lower values. The same trends hold when looking at information normalized by string length, i.e., $\mathrm{I}(\mathbf{y})/|\mathbf{y}|$ (see App. B), demonstrating these trends are not purely an artifact of string length. We note that in our human evaluations, the reference string was ranked first in 47% of cases and it was tied for first in an additional 16% of the cases. This suggests that the quality of the reference strings is on par with—if not higher than—the set of "top 1" model-generated strings.

Fig. 2 shows the distribution of deviations of strings' information content from the model entropy;[15] results are shown for both reference strings and top-ranked model-generated strings. Because these values are distributed quite evenly around 0, we take this as additional evidence that high-quality text usually has information content close to $\mathrm{H}(q)$. Further, the shapes of these curves motivate us to perform our next set of tests using $\varepsilon = \sigma$, the standard deviation of information values under q.[16]

We employ statistical hypothesis testing to see if the percentage of high-quality strings whose information content falls in the interval $[\mathrm{H}(q) - \sigma, \mathrm{H}(q) + \sigma]$ is greater than chance. For each input $\mathbf{x}$ (i.e., either a story prompt or article), we compute the information content of the reference and top-3 human-ranked strings. We then compute the percentage of items (among these four) that fall within $[\mathrm{H}(q(\cdot \mid \mathbf{x})) - \sigma, \mathrm{H}(q(\cdot \mid \mathbf{x})) + \sigma]$. We compare this percentage to the percentage of strings sampled directly from $q(\cdot \mid \mathbf{x})$ that falls within this interval. The former should (in expectation) be greater than the latter if the probability of high-quality strings having information content within this interval is greater than chance. Specifically, we test this using a paired, unequal-variance t-test, where samples with the same input are paired. At significance level $\alpha = 0.01$, we reject our null hypothesis—i.e., we reject that the percentage of highly rated strings (reference plus top-3 human-ranked strings) that fall within this interval is equal to (or less than) what we should expect by chance. Further, using a simple unpaired t-test, we find that the mean human score of strings (across all decoding strategies) within this region is significantly higher than those outside of this region. This characteristic is visualized in Fig. 3, where we plot the distributions of human quality ratings for strings inside and outside of this interval. We include a version of Fig. 3 further broken down by whether strings fall *above* or *below* this interval in App. B.

Additional plots reinforcing these observations can be found in App. B. Also see Meister et al. (2022) for follow-up experiments to this work.

5 Conclusion

In this work, we present the **expected information hypothesis**, which states that human-like strings typically have negative log-probability close to the expected information content of the probabilistic model from which they were generated. We use this hypothesis to explain why high-quality text seems to exist not necessarily at the high end of the probability spectrum but, rather, close to the entropy of the model. We provide empirical evidence in support of our hypothesis in an analysis of both human and machine-generated text, demonstrating that, overwhelmingly, high-quality text indeed has information content in the proposed region.

Ethics Statement

In order to complete our human evaluation, we used a crowdsourcing platform. For each task, we

[14] Specifically, for each input, we generate a single string according to each decoding strategy. We then rank these strings according to scores from human annotators.

[15] Note that this is not simply Fig. 1 shifted by a constant, as deviations are computed w.r.t. input-dependent conditional entropy estimates, i.e., $\widehat{\mathrm{H}}(q(\cdot \mid \mathbf{x}))$.

[16] Similarly to our estimation of $\mathrm{H}(q)$ in Eq. (3), σ can be estimated from the distribution of values of $\mathrm{I}(\mathbf{y})$ sampled from the model.

estimated the amount of time we expected the task to take and made sure that the crowdworkers would be paid (at minimum) a wage of $15 per hour. A further ethical consideration of this work is in the context of the use of language models for text generation. Language models have been used for the generation of malicious text, e.g., fake news and triggering content. The results in this work may provide insights for those using language models for such purposes as to how generations can be chosen to seem more "human-like."

Acknowledgments

The authors would like to thank the anonymous reviewers for their helpful recommendations, as well as Sander Dieleman for feedback on a preliminary draft of this paper.

References

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*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. 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 7282–7296, Online. Association for Computational Linguistics.

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):eaaw2594.

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*, pages 166–185, Online. Association for Computational Linguistics.

Sander Dieleman. 2020. Musings on typicality.

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*, pages 4506–4520. 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.

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*.

John T. Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics and Language Technologies*, pages 1–8.

John A. Hawkins. 2004. *Efficiency and Complexity in Grammars*. Oxford University Press, Oxford.

Charles F. Hockett. 1960. The origin of speech. *Scientific American*, 203(3):88–97.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *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.

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.

Clara Meister, Ryan Cotterell, and Tim Vieira. 2020. If beam search is the answer, what was the question? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2173–2185, Online. Association for Computational Linguistics.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. 2022. Typical decoding for natural language generation. *CoRR*, abs/2202.00666.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, 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.

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.

Tiago Pimentel, Clara Meister, Elizabeth Salesky, Simone Teufel, Damián Blasi, and Ryan Cotterell. 2021a. 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, Irene Nikkarinen, Kyle Mahowald, Ryan Cotterell, and Damián Blasi. 2021b. How (non-)optimal is the lexicon? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4426–4438, Online. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Claude E. Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27:623–656.

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

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*, pages 3356–3362, Hong Kong, China. Association for Computational Linguistics.

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.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151.

Ashwin K. Vijayakumar, Michael Cogswell, Ramprasaath R. Selvaraju, Qing Sun, Stefan Lee, David J. Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *CoRR*, abs/1610.02424.

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, Jaedeok Kim, Richard Yuanzhe Pang, and Kyunghyun Cho. 2020. Consistency of a recurrent language model with respect to incomplete decoding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 5553–5568, Online. Association for Computational Linguistics.

Gian Wiher, Clara Meister, and Ryan Cotterell. 2022. On decoding strategies for neural text generators. *CoRR*, abs/2203.15721.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the Cognitive Science Society*.

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.

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*, pages 25–33, Online. Association for Computational Linguistics.

George Kingsley Zipf. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley Press, Oxford, UK.

A The Typical Set

Let us imagine flipping N biased coins; specifically, let $X \sim p$ be an indicator random variable that takes values H and T. Take $p(X = \mathsf{H}) = 0.6$ and $p(X = \mathsf{T}) = 0.4$. Flipping N biased coins is then equivalent to taking N i.i.d. samples $x_n \sim p$. For reasonably large N, what might you expect the sequence $x_1, \ldots, x_N$ to look like? Few people would answer "all heads," even though this is technically the highest probability sequence. Rather, intuition tells you: an expected sequence would be one comprised of approximately 60% heads and 40% tails.

The samples that fall into the latter category have a distinctive characteristic: they contain a near-average amount of information w.r.t the support of the distribution over $X_1, \ldots, X_N$, where the information content of a realization $x_1, \ldots, x_N$ is defined as its negative log-probability. More formally, the (weakly) **(ε, N)-typical set** $A_\varepsilon^{(N)}$ for a chosen $\varepsilon > 0$ is the set of assignments $x_1, \ldots, x_N$ to random variables $\overrightarrow{X} = X_1, \ldots, X_N$ such that

$$2^{-N(\mathrm{H}(p)+\varepsilon)} \leqslant p(x_1, \ldots, x_N) \leqslant 2^{-N(\mathrm{H}(p)-\varepsilon)}$$

where $\mathrm{H}(p) \overset{\text{def}}{=} -\sum_x p(x) \log p(x)$ is the entropy—or equivalently, the expected value of the information content—of the random variable X. Under this definition we can prove that, for every $\varepsilon > 0$, there exists an N_0 such that for all $N > N_0$, we have that the (ε, N)-typical set contains at least $(1 - \varepsilon)$ of the probability mass of the joint distribution over $\overrightarrow{X}$. The concept of the typical set also generalizes to stochastic processes when we can actually compute their average information rate—or equivalently, their entropy rate.

B Experimental Design

B.1 Human Evaluations

For story generation and abstractive summarization, the raters are first presented with a news article/prompt. Next, they are presented, in random order, with the corresponding reference and the summaries/stories generated by different decoders. For each of two rating criteria, a score from 0 to 7 is assigned. For story generation the criteria are FLUENCY and NATURALNESS while for abstractive summarization QUALITY and ACCURACY are used. We provide the following short descriptions of the criteria to the raters:

FLUENCY: How fluent is the English text?

NATURALNESS: Does the text seem to be natural English text?

QUALITY: How high is the overall quality of the text?

ACCURACY: How well does the summary summarize the article?

After we obtain the ratings, we reject ratings that have not been filled out with care. Specifically, a rater is rejected if he assigns high scores to multiple examples that do not fulfill the specified criteria at all. If a rater has been rejected, we obtain a fresh set of ratings from a new rater.

C Additional Figures

We provide several additional results, looking further into the relationship between text information content and perceived quality. We see that in general, the distribution of information content of reference strings is quite close to that of the model. While the distribution of information content of top 1 ranked strings is also closer to the model distribution than many of the individual decoding strategies, the overlap is not as high as for reference strings.

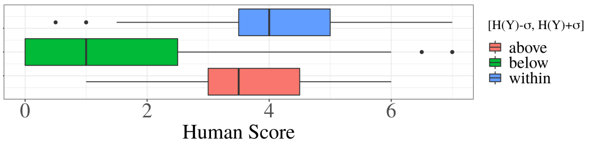


Figure 4: Human scores for strings (including both reference text and model-generated text) within 1 std of model entropy and above/below this interval. Note that "above" corresponds to text that has *lower* probability than the specified interval; due to the nature of the decoding strategies explored in this work, which all to some extent (except for ancestral sampling) disproportionately favor higher probability strings, only $< 5\%$ of all strings evaluated fall into the "above" category. Thus, we do not have a representative evaluation of this region of the probability space. However, it is often observed that extremely low-probability strings are usually incoherent or nonsensical.

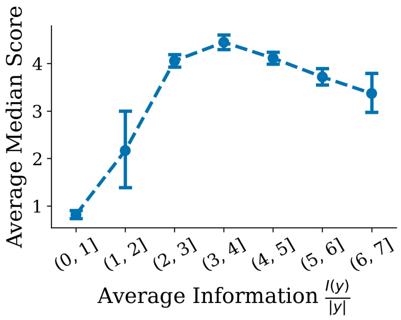


Figure 5: For story generation, median human scores (averaged across the two criterion) versus information, grouped by intervals; bars represent std. We normalize $\mathrm{I}(\mathbf{y})$ by length to mimic setup of Zhang et al. (2021), which controls for length during generation. As with Zhang et al. (2021), we see an inflection point in the relationship along the information (equivalently, negative log-probability) axis.

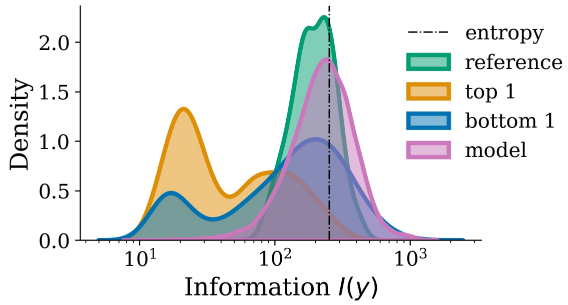


Figure 6: For abstractive summarization, the distribution over information $\mathrm{I}(\mathbf{y})$ values of: (model) the model, as estimated using samples from q; (reference) the reference strings; model-generated strings ranked (top 1) first and (bottom 1) last among all decoding strategies by human annotators. The latter 3 are all w.r.t. a held-out test set.

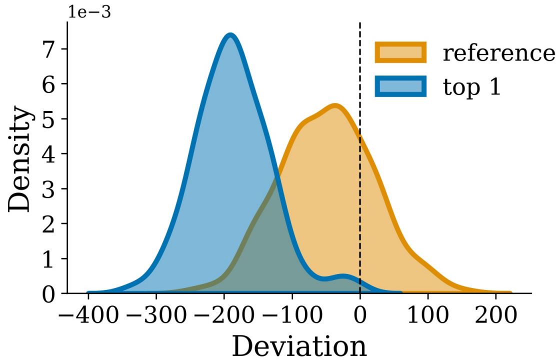


Figure 7: For abstractive summarization, the distribution of the difference in total information content for (1) test-set references and (2) top-ranked model-generated strings from the entropy of the model from which they were generated.

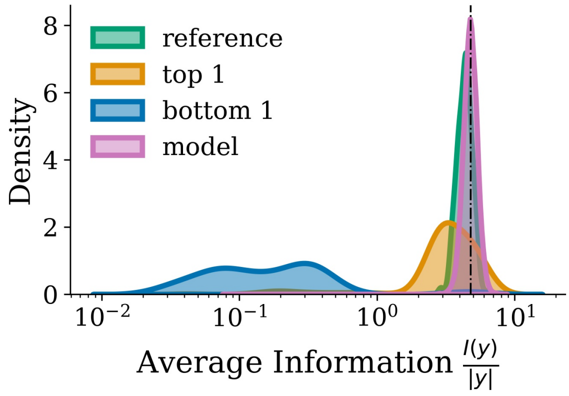
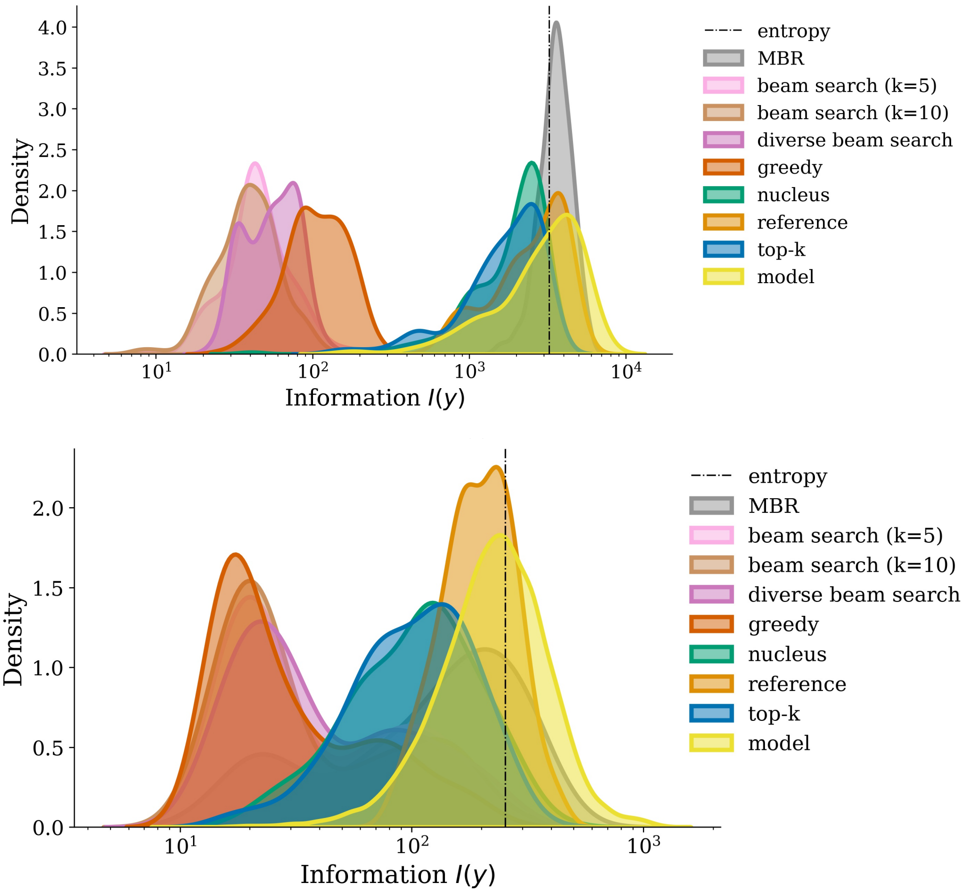


Figure 8: For story generation, the distribution over information ($\mathrm{I}(\mathbf{y})$) values normalized by length of: (model) the model, as estimated using samples from q; (reference) the reference strings; model-generated strings ranked (top 1) first and (bottom 1) last among all decoding strategies by human annotators. The latter 3 are all w.r.t. a held-out test set.



(b) a

Figure 9: The distribution over information ($\text{I}(\mathbf{y})$) values for strings generated under different decoding strategies for story generation (top) and abstractive summarization (bottom). Inputs are taken from a held-out test set.