Trading Off Diversity and Quality in Natural Language Generation

Hugh Zhang* Facebook AI hughzhang@fb.com

Daniel Duckworth* Google AI duckworthd@google.com

Daphne Ippolito Google AI dei@google.com

Arvind Neelakantan OpenAI arvindramanat@gmail.com

Abstract

For open-ended language generation tasks such as storytelling or dialogue, choosing the right decoding algorithm is vital for controlling the tradeoff between generation *quality* and *diversity*. However, there presently exists no consensus on which decoding procedure is best or even the criteria by which to compare them. In this paper, we cast decoding as a tradeoff between response quality and diversity, and we perform the first large-scale evaluation of decoding methods along the entire quality-diversity spectrum. Our experiments confirm the existence of the likelihood trap: the counter-intuitive observation that high likelihood sequences are often surprisingly low quality. We also find that when diversity is a priority, all methods perform similarly, but when quality is viewed as more important, nucleus sampling (Holtzman et al., 2019) outperforms all other evaluated decoding algorithms.

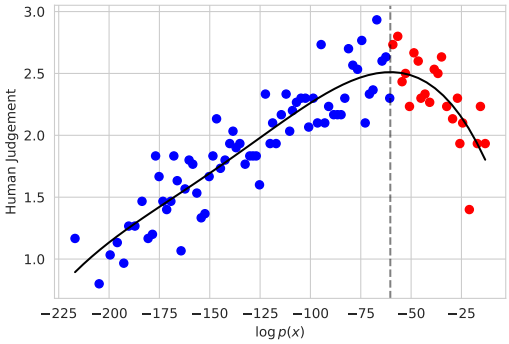


Figure 1: **The Likelihood Trap.** For a given context, we generate 100 sentences of equal length spanning a variety of model likelihoods and ask human crowdworkers to rate their quality. While model log-likelihoods are generally positively correlated with average human quality judgments, we notice an inflection point after which they become negatively correlated. Each point in the graph represents the average crowdworker rating of 5 sentences with similar model likelihoods.

1 Introduction

Generative language models are applicable for a wide variety of tasks including writing articles, composing Shakespearean sonnets, and engaging in conversation (Radford et al., 2019; Zhang et al., 2019; Fan et al., 2018). This work examines decoding methods, a critical component in language models used in open-ended generative tasks where successful models must generate a *diverse* spectrum of high *quality* answers rather than merely a single output (Ippolito et al., 2019a).

For many tasks, these two criteria of quality and diversity are not equally important. In machine translation, the most important criteria is to produce an accurate, high-quality translation of the input; generating a variety of alternative translations is also useful, but not if it comes at the cost of correctness. Meanwhile, in open domain dialogue the goal is often to sustain an enjoyable conversation with a human conversational partner and as such, a higher premium is placed on diversity. To give a concrete example for the case of dialogue, the phrase "I don't know" is typically a perfectly reasonable remark that appears quite often in the course of normal human conversation. However, a chatbot that only repeats "I don't know" makes for a very poor conversationalist. In such open-ended domains, being able to converse about a wide variety of topics with the occasional odd remark is highly preferred to merely repeating the safest possible remark over and over (Li et al., 2016).

To evaluate both of these criteria, we characterize the performance of decoding algorithms along the entire quality-diversity spectrum instead of simply at individual points. We compare a variety of commonly-used decoding algorithms in the first

* denotes equal contribution

Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval), pages 25–33
Online, April 19, 2021.

large-scale study of decoder performance, utilizing over 38,000 ratings on almost 10,000 samples. Our results indicate that when diversity is highly valued, all decoders perform similarly, but when quality is viewed as more important, the recently proposed nucleus sampling (Holtzman et al., 2019) outperforms all other evaluated decoding algorithms.

Additionally, we investigate the commonly held intuition that model likelihood is directly correlated with human quality judgments by explicitly measuring the relationship between the quality of a sentence as judged by human raters and its likelihood under a generative model. Our findings confirm the existence of a *likelihood trap*, the counter-intuitive observation that the highest likelihood sentences are of extremely low quality, despite a generally positive relationship between model likelihoods and human quality judgments. While this finding has been observed across a wide variety of models and tasks from news generation to machine translation (Cohen and Beck, 2018; Holtzman et al., 2019), to our knowledge we are the first to explicitly quantify the relationship between the two across the entire model probability space.

2 The Likelihood Trap

Sequence likelihood is commonly used as a heuristic for selecting high-quality generations. Beam search, the principal approach adopted in machine translation, encapsulates this principle by (approximately) finding the *single* most likely generation $\text{argmax} \log \text{p}_{\text{model}}(x)$.

However, prior work has suggested that this assumption of a monotonically positive relationship between sequence likelihood and sequence quality breaks down at the extremes (Section 5). We empirically quantify the relationship between sequence likelihoods and human quality judgments by subsampling a large number of context-continuation pairs representing a wide variety of model log-likelihoods. We then request human crowdworkers to rate the quality of each continuation given the context on a five-point "Terrible"-to-"High Quality" scale. Figure 1 plots these ratings as a function of $\log \text{p}_{\text{model}}$ and confirms that on average the highest quality generations are *not* the most likely. Specifically, we find that continuation quality is generally positively related with $\log \text{p}_{\text{model}}(x)$ up until an inflection point after which it becomes negatively related. Our findings suggest that while model likelihoods form a good proxy for continuation quality, naively maximizing over sentence likelihood leads to suboptimal continuation quality. We term this phenomenon the *likelihood trap*.

3 Evaluation Framework

We introduce an evaluation framework for measuring the trade off quality and diversity in language generation. We consider autoregressive language models that decompose the likelihood of a sequence $x_{1:n}$ token-by-token in a left-to-right fashion (Hamilton, 1994; Sutskever et al., 2014). Specifically, the (conditional) probability of the sequence is:

$$\text{p}_{\text{model}}(x_{1:n} \mid c) = \prod_{i=1}^{n} \text{p}_{\text{model}}(x_i | x_{1:i-1}, c)$$

where c is any additional conditioning signal, such as the previous turn of dialogue. Typically, p_{model} is not sampled from directly; it is first post-processed by a decoder to bias it towards already high-likelihood tokens.

We evaluate the quality of a single sequence $x_{1:n}$ by asking humans for a quality judgment $\text{HJ}(x)$. We can define the quality of a model $Q(p) = \mathbb{E}_{x \sim p}[\text{HJ}(x)]$ as the expected human "quality" judgment for sentences drawn from it. We measure the diversity of a model via the (conditional) Shannon entropy H (Shannon, 1948), a diversity metric widely used across many fields beyond computer science including biology, economics, chemistry, and physics. Conditional Shannon entropy is given by $H(p \mid c) = -\mathbb{E}_{x \sim p(x|c)}[\log p(x \mid c)]$. Since many metrics for measuring diversity in language generation exist in the literature, we validate our choice of entropy by measuring its correlation with other commonly used metrics of diversity based on n-gram frequency. We find the Spearman correlation with distinct-1 and distinct-2 (number of distinct unigrams and bigrams divided by total number of n-grams) to be 0.80 and 0.77 respectively over sentences generated by GPT-2.

Our choices of using the average human quality judgement to measure quality and entropy to measure diversity guarantee that the *optimal* Pareto frontier trades off monotonically between quality and diversity. Optimizing quality with no regard for diversity results in outputting only the single highest quality sentence, whereas optimizing for diversity with no regard for quality results in outputting every utterance with equal probability. Typical tasks in language generation (e.g. summarization,

machine translation, storytelling) will fall somewhere in between these two extremes.

Since our models are imperfect, each decoding algorithm will, to the best of its ability, trace out its own estimate of this frontier. As most commonly used decoding strategies offer a knob to control the diversity of the generated text, we compare the performance of decoding algorithms by plotting their performance along various positions on the quality-diversity tradeoff curve.

4 Experiments

We evaluate three commonly used decoding algorithms, sweeping across the quality-diversity curve by considering several hyperparameter settings per decoding algorithm. At the extremes of their hyperparameter ranges, these algorithms all converge to greedy and random sampling, respectively.

- **temperature**: Sample tokens with probability proportional to $p(x_i|x_{1:i-1})^{1/t}$, $t \in [0, 1]$.
- **top-k** (Fan et al., 2018): Sample tokens only from the k highest likelihood tokens in the vocabulary at each timestep, $k \in [1, \text{vocab size}]$
- **top-p** (also known as nucleus sampling) (Holtzman et al., 2019): Sample only from tokens comprising the top-p percent of probability mass at each timestep, $p \in [0, 1]$.

4.1 Setup

Due to the large monetary cost of evaluation, we evaluate each decoding algorithm on the same language model: the 774M parameter variant of GPT-2 (Radford et al., 2019), a publicly-released language model. To ground samples in a common context, we select a set of 48 examples from the GPT-2 test set to condition upon by manually filtering out examples containing explicit content or web markup. Samples are drawn by conditioning on a 'prompt' consisting of the first 20 space-delimited words of a test example. As sample quality becomes ambiguous when samples are terse (Ippolito et al., 2019a), we explicitly require all sampling methods to generate exactly 30 tokens, a length approximately equal to the prompt.

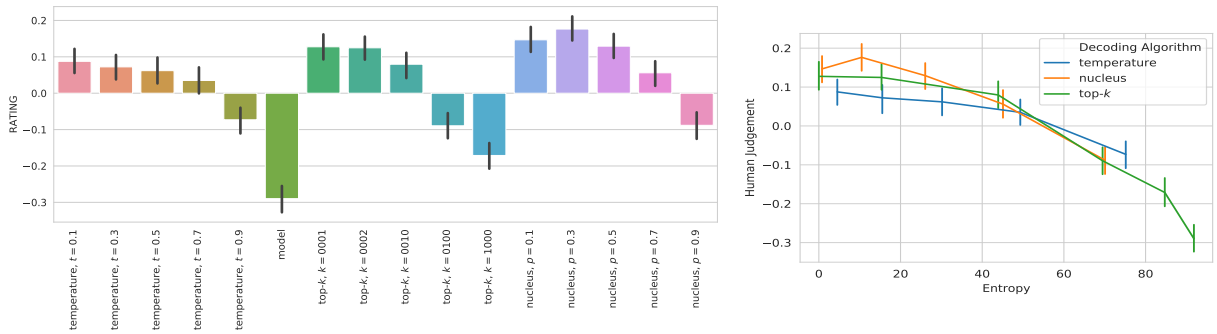
To estimate the expected Human judgment score $\mathbb{E}_p[\text{HJ}(x)]$ of the probability distributions induced by each decoding algorithm, we enlist a qualified pool of 146 Amazon Mechanical Turk (AMT) workers selected by satisfactory performance on a qualification task. Workers are presented sets of five samples, each conditioned on the same prompt and drawn from five different algorithm-hyperparameter configurations and asked to assign qualitative scores to each sample ranging from human-like to gibberish. The exact prompts as shown to crowdworkers along with thorough descriptions of our data collection process and our checks for robustness are included in the Appendix.

Prior work has found that human annotaters have significant trouble in directly separating out machine and human generated continuations when they are of similar quality, as the task of assessing sentence quality is highly subjective (Ippolito et al., 2019a). We found that constructing pairwise preference ratings by randomly pairing samples evaluated at the same time significantly reduced the variance of our results. Specifically, if one sample is rated higher than the other, one is assigned a score of +1 and the other -1. If both are rated equally, both are assigned a score of 0. The score assigned to a decoding configuration is its average score across all pairwise preference ratings.

4.2 Results

We now introduce the first large-scale study comparing decoding algorithms and their hyperparameters. Unlike all prior work (Holtzman et al., 2019; Ippolito et al., 2019b), we explicitly put decoding algorithms *on equal footing* by comparing sample quality at equal points of diversity. We consider five hyperparameter configurations per decoding algorithm for a total of fifteen configurations. For each configuration and prompt, we draw ten samples. In total, workers rate nearly 10,000 samples resulting in over 38,000 paired ratings. Our main results are summarized in Figures 2a and 2b. Reassuringly, both entropy and human quality judgements vary smoothly with decoding algorithm hyperparameter.

As expected, random sampling directly from $\text{p}_{\text{model}}(x)$ is simultaneously the highest entropy *and the lowest quality*. This is empirically consistent with the long-standing intuition that decoding algorithms are critical to improving sample quality. Why is text from random sampling such poor quality? Language models such as GPT-2 are trained to minimize the KL-divergence between a training set and the model distribution p_{model}, an objective that prioritizes recall over precision (Arjovsky et al., 2017). As a result, models tend to ensure that high quality sequences have high likelihood without insisting that all high likelihood sequences also have high quality. When we evaluate samples from the



(a) (b)

Figure 2: **(a)** Human judgment scores for each decoding algorithm and hyperparameter choice. A score of 0 represents the average human judgement rating of all the sentences evaluated. Nucleus sampling is rated the highest while random sampling ("model") performs the worst. **(b)** Decoder quality plotted as a function of entropy, with each point representing a single decoding configuration. Error bars represent 95% bootstrap confidence intervals.

model, we evaluate the latter condition.

Our second conclusion is that sample quality varies significantly with entropy for all decoding algorithms. Moreover, when aligned on entropy, sample quality between all autoregressive decoding algorithms is comparable across a wide range. It is only when entropy is low – when decoding algorithms heavily influence sampling – that sample quality between algorithms diverge. In this regime, we find that nucleus sampling outperforms top-k, which in turn outperforms temperature sampling. Observing such a difference should be unsurprising: the entropy of a distribution alone does not characterize its samples and thus its overall quality. As such, a fair comparison of decoding algorithms must not only compare at the same level of entropy but at a *range* of entropy levels.

5 Related Work

Encouraging Diversity We choose to evaluate three commonly used decoding methods: nucleus sampling (Holtzman et al., 2019), top-k sampling (Fan et al., 2018), and temperature sampling. All three methods control the relative tradeoff between quality and diversity with a single hyperparameter as described in Section 4, though many other decoding methods also exist in the literature. Ippolito et al. (2019b) compares many of these algorithmic advancements on the tasks of open-ended dialog and image captioning, concluding that quality-diversity tradeoffs make it difficult to say that any one method is ubiquitously best.

Likelihood Trap We are far from the first to observe evidence of the likelihood trap. In particular, the machine translation and image captioning communities have long known that using higher beam sizes often leads to lower BLEU scores (Vinyals et al., 2016; Yang et al., 2018; Stahlberg and Byrne, 2019; Meister et al., 2020). In open-ended generation, Holtzman et al. (2019) find similar results, observing that maximizing the likelihood generates extremely repetitive sentences. Our main contribution towards understanding the likelihood trap is the first explicit measurement of the relationship between model likelihoods and human quality judgments at all points in the model probability space, not just the endpoints.

Frameworks Our framework differs from those which ask that generative models mimic the training distribution exactly (Hashimoto et al., 2019; Kingma and Welling, 2013; Goodfellow et al., 2014). While indistinguishability is sometimes the ultimate goal, humans make errors, and a perfect model would not seek to imitate these mistakes. As we ground quality evaluations in human judgments rather than statistical measures, our framework is easily able to capture the possibility of superhuman performance.

6 Conclusion

In this paper, we propose a framework for credibly evaluating decoding algorithms and use it to conduct the first large scale evaluation of decoding algorithms by measuring their performance along the entire quality-diversity frontier. We observe that decoders can be tuned to produce higher-quality text, but that this improved quality comes at the cost of diversity. Our findings suggest that existing decoding algorithms are largely interchangeable in high diversity settings, but that nucleus sampling

performs best when quality is valued over diversity. We show that when performing a comparison of text generated from multiple decoding algorithms, it is crucial to ensure equivalent diversity to make the comparison fair, a step many evaluations fail to do. Finally, we warn against falling for the *likelihood trap*, as selecting generated text that is *too* likely results in text that humans judge to be worse.

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein gan. *arXiv preprint arXiv:1701.07875*.

Eldan Cohen and J Christopher Beck. 2018. (unconstrained) beam search is sensitive to large search discrepancies.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*.

I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*.

James D Hamilton. 1994. *Time series analysis*, volume 2. Princeton New Jersey.

Tatsunori B Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. *arXiv preprint arXiv:1904.02792*.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2019a. Human and automatic detection of generated text. *arXiv preprint arXiv:1911.00650*.

Daphne Ippolito, Reno Kriz, Maria Kustikova, João Sedoc, and Chris Callison-Burch. 2019b. Comparison of diverse decoding methods from conditional language models. *arXiv preprint arXiv:1906.06362*.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

J. Li, M. Galley, C. Brockett, J. Gao, and W. B. Dolan. 2016. A diversity-promoting objective function for neural conversation models. pages 110–119.

Clara Meister, Tim Vieira, and Ryan Cotterell. 2020. If beam search is the answer, what was the question?

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Claude E Shannon. 1948. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423.

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.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2016. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):652–663.

Yilin Yang, Liang Huang, and Mingbo Ma. 2018. Breaking the beam search curse: A study of (re-) scoring methods and stopping criteria for neural machine translation. *arXiv preprint arXiv:1808.09582*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation.

A Appendix

A.1 Experimental Design

In this section, we describe the design of experiments presented in Section 4 in greater detail.

We begin by describing the task presented to crowdsourced raters. A sample task is shown in Figure 4. Each task consists of a "context" sequence of the first 20 words in a news article.[1] We then present the rater with five continuations of 30 word-piece tokens. The rater assigns a label of "High Quality," "Decent," "Passable," "Bad" or "Terrible" to each. We note that these labels are inherently subjective, and include a description and reference example before each task to calibrate the rater. The same description and example is repeated in Figure 3.

In preliminary experiments, we found examples and instructions insufficient for achieving repeatable results. Manual inspection of rater responses revealed a failure to interpret the labels correctly as well as spammers who would always choose the same response for every prompt. As a result, we crafted a qualification exam of five continuations. Only raters which rated all five continuations correctly or nearly correctly[2] were allowed to participate in further experiments. Of the 550 crowdsourced workers surveyed, 136 met this criteria. We refer to this set of raters as the "qualified rater pool" below.

Even with a qualification exam, we found raters often disagree on the appropriate label for a given continuation. However, when asked to choose which of two continuations was higher quality quality (if any), raters were better aligned. With this in mind, we choose to analyze *pairs* of ratings given in the same task. From five absolute ratings, we construct twenty pairwise preference ratings: two per pair of continuations. If two continuations receive the same label, they are assigned a preference of 0. If the first continuation is rated higher than the second, a the pair (first, second) is assigned a score of +1 and the pair (second, first) a score of -1. All analyses comparing multiple decoding methods use this methodology.

Even with the precautions above, care is needed to ensure repeatable results. To measure this, we performed an "A/A" experiment prior to data collection. This experiment consists of having the same tasks rated by two different pools of raters. Identical analyses are performed on both rating results, and the experimental setup is deemed valid if conclusions are consistent. To achieve this, we constructed 150 tasks[3] using a subset of the context sequences and decoding methods from our primary experiment. We artificially split the qualified worker pool in two by sending the same tasks for evaluation at midnight and at noon.[4] We submit the same set of tasks to both rater pools. An analysis of results from both sets of ratings (Figure 5) reveals a statistically consistent preference of top-p over top-k and (local) temperature sampling, and a severe disapproval of random sampling from the model. These results are also consistent with the same statistics gathered in the full-scale experiment presented in the main text and another experiment described below.

To further validate the reliability of our methodology, we explicitly measure inter-rater agreement on the same set of 150 tasks in a follow-up experiment after large-scale data collection. In this experiment, we ask each task be rated by five distinct raters. We measure Fleiss's Kappa, a measure inter-rater agreement, on the resulting pairwise ratings. We obtain a score of 0.1964 – an indication that a correlation between raters exists but that the task is far from unambiguous. While this may initially appear concerning, we argue that this is an indication of the task's difficulty. Unlike image classification, for example, a universally agreeable criteria for text quality does not exist. A measure of Cohen's Kappa on the A/A experiment above produces a score of 0.19578 – nearly identical to the inter-rater agreement experiment described here. The similarity of these two statistics gives evidence that the proposed experimental design is repeatable in spite of the task's ambiguity. These reuslts underscore the importance of large-scale, repeatable studies like that presented here. We conclude by measuring rater preference between each pair of sampling method and hyperparameter on the five-raters-per-task inter-rater agreement experiment described above. Results, as shown in Figure 7, indicate that the same trends presented in the full-scale experiment (Figure 2a) hold.

[1] News articles are sourced from GPT-2's WebText dataset. https://github.com/openai/gpt-2-output-dataset

[2] Raters which incorrectly labeled at most one continuation with a label at most one level off (e.g. if the correct answer is "Bad", acceptable errors are "Passable" and "Terrible") are counted as "nearly correct".

[3] The large-scale experiment includes 1,930 tasks.

[4] All tasks within each experiment were rated within 4 hours and 1.5 hours, respectively.

Context	Continuation	$\log p(x)$	Classification
The Atlanta Falcons have started the 2015 season 4-0 under new head coach Dan Quinn. Quarterback Matt Ryan has the ...	... mental Tough O'Rourke Tough apology assessment category of virtue from Boser' Blog here. It's got letters and images on it and is utterly ...	-177	Nonsense
	... team afloat and looks closer to the 2010 Atlanta Falcons. Starting cornerback Desmond Trufant was one of the top players on the 2014 ...	-74	Reasonable
	... team in the thick of the NFC South race. The Atlanta Falcons have started the 2015 season 4-0 under new head coach Dan Quinn. Quarter...	-14	Repetition
They have changed the phone menu to try to deflect us to email, but you can still get a live ...	... answer from a female administratoria llallushoss@rahpx Sandra PJ Jenniea nightiopq HamidF daroyqg S') ...	-229	Nonsense
	... message or call on line, so I suppose they are just using that as an excuse. Yet they are still telling people to change their telephone number...	-86	Reasonable
	... link to a phone number here. They have changed the phone menu to try to deflect us to email, but you can still get a live link to...	-23	Repetition

Table 1: Examples of sentences at various model likelihoods. Sentences with very low $\log p_{\text{model}}$ generate nonsense, while sentences that have high likelihood under the model often devolve into extreme repetition. Nonsense and repetition classifications shown here are only for illustrative purposes. Crowdworkers simply rated sentences for overall quality.

Instructions

Below you will find multiple continuations to a given "context" sentence. Please rate the continuations according to their quality.

Notes: High quality continuations tend to **sound like fluent English.** Low quality continuations tend to **repeat, contradict prior statements, or look like text directly copy-pasted from a web page.** Continuations may terminate at ANY time, including in the middle of a word. **DO NOT PENALIZE for early termination.**

We will reject your HIT if you input obviously wrong answers.

Example

Example Context: **USA Today - Every offseason features trades and free-agent signings, but rarely have as many stars found new homes as they did last...**

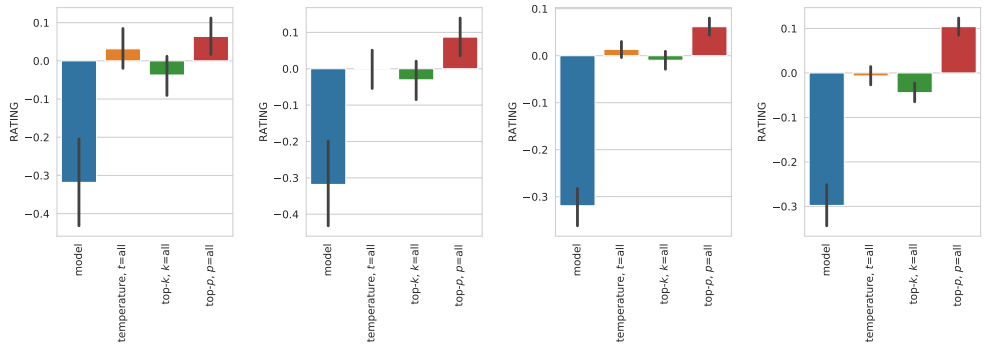
Quality	Description	Example Continuation
High Quality	Completely plausible, impossible to distinguish from human-written.	...summer. Chris Paul requested a trade and joined James Harden in Houston. Carmelo Anthony and Paul George moved to Oklahoma City. Jimmy Butler was shipped to Minnesota...
Decent	Plausible, natural-sounding text, but may contain one or two subtle contradictions or weirdnesses.	...spring. LeBron James and Tom Brady collectively decided to move to South Beach, where they joined Serena Williams in competing for the...
Passable	Looks plausible at first glance but doesn't make sense. May be entirely off-topic, repetitive, or contain serious contradictions.	...winter, where Dwight Howard of the Nets teamed up with Dwight Howard of the Nets. This super duo is set to to take the Broncos...
Bad	Not a plausible English sentence, contains extreme repetition or web fragments such as "Hide Caption 11" or "Like us on Facebook Twitter."	...a series of public public appeals. Slideshow (3 Images) FILE PHOTO https://www.washingtonpost.com/news/worldview...
Terrible	Not a valid English sentence, gibberish.	... wound' '.Breached KIA, by inter MinurtiSir Clinea490 down to Ot got herg of te-.ion (W:

Figure 3: Instructions for the crowdworker task. Each sentence continuation is labeled on a scale from "Terrible" to "High Quality". A description of each label and an example continuation that fits each each is provided before each task. Exact example used may vary.

Given the following context, please rate the next 5 continuations: **"Since Afreeca Freecs' top laner Jang "MaRin" Gyeong-hwan pulled out the splitpushing AD Kennen during the very early stages of the"**

Since Afreeca Freecs' top laner Jang "MaRin" Gyeong-hwan pulled out the splitpushing AD Kennen during the very early stages of the **2015 LCK Summer Season, Afreeca Freecs figured out how to move the weight of the team's weight when it came to mid lane...**				
○ **High Quality**	○ **Decent**	○ **Passable**	○ **Bad**	○ **Terrible**
Since Afreeca Freecs' top laner Jang "MaRin" Gyeong-hwan pulled out the splitpushing AD Kennen during the very early stages of the **spring split, two rookies stepped up to play from that spot instead. Lee "StarDust" Min Jong performed well in solo queue, and as...**				
○ **High Quality**	○ **Decent**	○ **Passable**	○ **Bad**	○ **Terrible**
Since Afreeca Freecs' top laner Jang "MaRin" Gyeong-hwan pulled out the splitpushing AD Kennen during the very early stages of the **Global Championships, a lot of Korean fans came to see what the other side of the Jungle was missing: the rookie Pawn. With his new...**				
○ **High Quality**	○ **Decent**	○ **Passable**	○ **Bad**	○ **Terrible**
Since Afreeca Freecs' top laner Jang "MaRin" Gyeong-hwan pulled out the splitpushing AD Kennen during the very early stages of the **ir match against LGD Gaming last week, we knew there'd be some upsets in the mid lane. And they have delivered so far. In...**				
○ **High Quality**	○ **Decent**	○ **Passable**	○ **Bad**	○ **Terrible**
Since Afreeca Freecs' top laner Jang "MaRin" Gyeong-hwan pulled out the splitpushing AD Kennen during the very early stages of the **season in the club's organization's first series, it has been a central issue in the relegation story. Freecs is currently tied with MK....**				
○ **High Quality**	○ **Decent**	○ **Passable**	○ **Bad**	○ **Terrible**

Figure 4: Sample crowdworker task used for the main evaluation results. Raters assign a label on a scale from "Terrible" to "High Quality" to each of five continuations sharing a common context of twenty words. Each continuation is generated by a different sampling method and hyperparameter.



(a) A/A, midnight (b) A/A, noon (c) Full-scale (d) Inter-rater

Figure 5: Average Human judgement scores for each sampling method, aggregated across sampling method hyperparameters. In spite of being collected by different raters on different sets of tasks and different points in time, rater preference remains consistent.

Experiment	Num Ratings	Kappa
A/A	2,968	0.1957 (Cohen's)
Five-Rater	14,760	0.1964 (Fleiss's)

Figure 6: Inter-rater agreement between pairwise preference ratings as measured in a preliminary A/A experiment and an explicit, five-raters-per-task inter-rater agreement experiment. While agreement is low, Kappa is strongly consistent between both experiments.

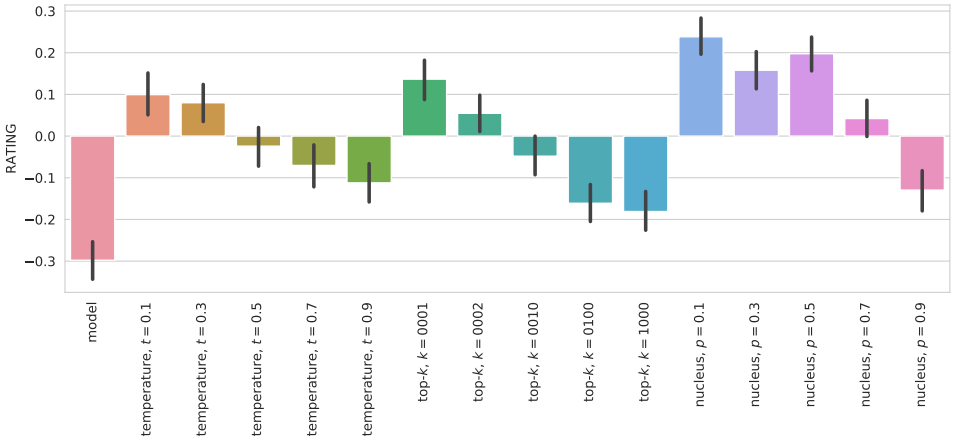


Figure 7: Human judgement scores for each decoding algorithm and hyperparameter choice, as measured in the inter-rater agreement experiment. Preference between sampling methods remains consistent with large-scale experiment shown in Figure 2a in spite of using only decodes generated by a subset of context sequences.