New-Wiki Eval: An Evolving Wikipedia Multi-metric Evaluation for Large Language Models

Anonymous ACL submission

Abstract

Latest large language models (LLM) like GPT-3 are able to generate long articles that are indistinguishable from human-written ones. However, the evaluation of text generation remains challenging. While human evaluations of generated articles are shown to be expansive and slow, researchers cannot find good automatic evaluation methods because of the lack of out-of-sample reference text and the creativity of long text generation. We made a key observation that Wikipedia is constantly evolving and thus provide a good-quality out-of-sample test set for LLMs. Thus, in this paper, we propose a new evaluation framework for LLM's long text generation. We first let the LLMs do "Wikipedia generation" and then select a set of evaluation metrics to evaluate the generation from multiple perspectives. In practice, we evaluate state-of-the-art LLMs including GPT-3, BLOOM, OPT, GLM, BART, and T5 and show the evaluation results under our framework correlate with prior research.

1 Introduction

Generative language models demonstrated impressive capabilities by training with more and more parameters and corpus. In particular, GPT-3, an LLM consisting of 175 billion parameters, has demonstrated the ability to generate human-indistinguishable articles, follow instructions, and solve many traditional language tasks (Brown et al., 2020). Since then, there is a growing interest in the NLP community to make larger and better LLMs. Examples include OPT (Zhang et al., 2022), GLM (Du et al., 2021), BLOOM (BigScience, 2022), LaMDA (Thoppilan et al., 2022), and PaLM (Chowdhery et al., 2022).

While LLMs are automatically evaluated on traditional downstream tasks like question-answering and machine translation, to the best of our knowledge, there is no good automatic evaluation metric for an important task: long text generation. Brown et al. (2020) evaluate GPT-3's long text generation performance by conducting human evaluations to see if a human can distinguish the generated stories from real ones following Zellers et al. (2019). However, researchers raised many concerns about relying solely on human evaluation. First, human evaluation is expensive and slow (Sellam et al., 2020) and it's hard to be compared and reproduced because of the diverse assessment criteria (Howcroft et al., 2020). This prevents researchers from getting quick and standardized feedback of their LLMs' generations. Second, the length, fluency, and complexity of GPT-3 generated articles made it difficult for human evaluators to go beyond surface-level fluency-based quality and provide desired evaluation (Clark et al., 2021). Thus a good automatic evaluation metric for LLM's long text generation is needed along with more standardized and better-trained human evaluations.

There are two difficulties in designing an automatic metric for long text generation of LLMs. First, one needs a good reference text to evaluate text generation. But since LLMs can be trained on any corpus available on the internet, it is hard for researchers to identify reference text that LLMs are not trained on. This leads to the issue of **data contamination** that is concerned in the GPT-3 paper (Brown et al., 2020) and shown to affect the model performance substantially (Magar and Schwartz, 2022). Second, its evaluation is hard even with the reference text. Given the creativity of long text generation, the generated text can be good even not talking about the same thing as the reference text (e.g., story generation). Therefore, a generation task on a more narrowly defined topic would be easier to evaluate.

In this paper, we note that Wikipedia is constantly evolving (new Wikipedia articles are added every day) and thus provide a good quality out-of-sample reference text to evaluate long text generation. This trait of constantly evolving is a key to our research as it provides a way to separate out a test set after any chosen date so that the latest LLMs are not trained on it.

Based on this observation, we collect New-Wiki Dataset and propose a new evaluation framework for LLMs. We first propose the task "Wikipedia generation". In our task, we let LLMs generate Wikipedia-style articles given the title and first sentence of the original Wikipedia article. We expect a good language model to generate an article that is relevant, knowledge-intensive, and factually correct. So we select a suite of metrics from six aspects to evaluate the generated articles and their characteristics. Lastly, we conduct extensive experiments by using seven state-of-the-art large language models to do the Wikipedia generation task. We show that the evaluation results are highly correlated with prior research and thus the effectiveness of our evaluation framework. Aside, we also find new insights into the characteristics of different state-of-the-art language models.

Note that the main contribution of this paper is to introduce a new evaluation framework along with the New-Wiki dataset so that any metrics could be added to it and any LLM could be evaluated by it. The composite metric building primarily on the off-the-shelf metrics and the experiments all aim at showing the effectiveness of our evaluation framework.

2 Related Work

Automatic Evaluation of Long Text Generation

Many evaluation metrics for evaluating long-text generation have been proposed. They could be categorized into n-gram-based metrics and deep learning-based metrics.

N-gram based metrics including BLEU score (Papineni et al., 2002), ROUGE score (Lin, 2004) are commonly metrics for decades. However, they are sensitive to lexical differences and could not capture semantic variations from the reference text. Thus, they are criticized for not correlating well with human evaluation (Novikova et al., 2017). In the task of Wikipedia generation, having lexicon overlap is very difficult and thus n-gram-based evaluation is not effective as we show in Appendix.

Various deep learning-based metrics for NLG have been proposed recently. For example, BERTScore measures the cosine similarity between the generated text and the reference text (Zhang et al., 2019) and BLEURT designs a more robust metric (Sellam et al., 2020). Although these metrics can provide a single score for the generated text, they are black-box models that cannot explain how different aspects of the generation contribute to the scoring. This hinders researchers from understanding the characteristics of LLMs. Thus, composite metrics like Gehrmann et al. (2021) are proposed. In our research, we carefully choose our automatic evaluation metrics from six different aspects to construct a new well-rounded, fine-grained, and more explainable multi-metric evaluation framework.

The closest work to ours is the evaluation pipeline in Li et al. (2022). They also take the beginning part of the Wikipedia article as prompt, ask the LLM to complete the article, and then use the rest of the article as reference. However, the dataset they use is in-sample data for LLMs and thus raise the concern of data contamination. Our work filter Wikipedia articles by date to ensure the test set is not seen by LLMs before.

Wikipedia-related work

Wikipedia has long been studied by NLP researchers as a good source of knowledge. Wikipedia is used for question-answering (Hewlett et al., 2016), information retrieval (Lehmann et al., 2015), and text summarization (Sauper and Barzilay, 2009). The closest idea to our "Wikipedia generation" is Liu et al. (2018). In their paper, similar Wikipedia articles are first retrieved and then summarized into new Wikipedia articles, whereas we leverage the knowledge in LLMs and generate new Wikipedia directly.

Knowledge and Factuality in Language Generation

As the issue of generating fluent text is gradually solved by LLMs, researchers become more interested in generating informative and factually correct text. While LLMs are shown to memorize real-world knowledge into its parameters (Carlini et al., 2022), when it comes to open-ended generation, its generations are often hallucinating and not factually correct (Maynez et al., 2020). This is particularly a concern for grounded text generation (Honovich et al., 2021).

Various methods have been proposed to improve the informativeness, factuality, and verifiability of language generation. One line of research add the information retrieval step before language generation (RAG; Lewis et al., 2020). Another line of research performs data grounded text generation such as kb-to-text (Gardent et al., 2017), table-to-text (Parikh et al., 2020). These efforts also stress the need for an automatic evaluation metric for knowledge-intensive long text generation.

3 New-Wiki Datset

Finding a good evaluation dataset for LLMs is a challenging problem given the issue of data contamination. We choose Wikipedia for the following considerations: (1) Wikipedia and constantly evolving. This can continuously provide an out-of-sample test set that could avoid the issue of data contamination. (2) Wikipedia contains a set of factual knowledge that can be considered as "ground truth". This provides the information that is expected to show up in the generation and thus better serves as a benchmark compared to intrinsically open-ended tasks like story generation.

We also note that although the language model is not supposed to know about the future and have direct knowledge of the Wikipedia article that is created after the model's release. Since we include the title and the first sentence of the original article, the model would have enough context to infer the content. For example, given the Wikipedia title "2022 Russian invasion of Ukraine" and its first sentence, "On 24 February 2022, Russia invaded Ukraine in a major escalation of the Russo-Ukrainian War, which began in 2014", good LLMs are expected to recall the related knowledge from its memorization and generate a passage that is relevant and factually correct to some extent.

Thus we collect and publish the New-Wiki dataset consisting of Wikipedia articles created **between June, 2021 and Decemeber, 2021** as our test set.[1] While language models keep evolving and will be trained on newer Wikipedia, we keep the New-Wiki dataset updated regularly, and thus it could serve as a good test set of knowledge-intensive text generation without the issue of data contamination.

In practice, we used Wikipedia API and requested all the Wikipedia articles that are created between June 2021 and December 2021. We sample 3000 Wikipedia articles that have more than 10 revisions and longer than 500 words to do generation. The summary of the dataset is given in Table 1.

[1]We choose all articles after June, 2021 to make sure GPT-3 Davinci-002, BLOOM, OPT, GLM are not trained on them. Details about the data they are trained on can be found in Appendix.

Entity Type	Occurrence	Percentage
Human	1328	44.2%
Taxon	251	8.4%
Media	239	8.0%
Event	217	7.2%
Human Settlement	185	6.2%

Table 1: Topics covered in New-Wiki

4 Methodology

4.1 Generative Language Models

In this paper, we evaluate the following state-of-the-art generative language models: GPT-2 (Radford et al., 2019), GPT-3 (Brown et al., 2020), OPT (Zhang et al., 2022), BART (Lewis et al., 2019), T5 (Raffel et al., 2019), GLM (Du et al., 2021), and BLOOM (BigScience, 2022). For BART and T5, we fine-tune them on 2000 Wikipedia articles for 10 epochs to let them perform long text generation. Models are summarized in Table 2.

Model	# Parameters	Release date
GPT-2	1.5B	Feb. 2019
BART-base	110M	Oct. 2019
T5-base	220M	July 2020
GPT-3	175B	July 2020
OPT-66B	66B	May 2022
BLOOM	175B	June 2022
GLM	130B	Aug 2022

Table 2: The release date and parameters of SOTA large language models

4.2 Evaluation Metrics

We evaluate metrics from six different perspectives to provide a well-rounded and explainable view of the language model's performance. In practice, we choose text quality and text diversity metrics to evaluate the general quality and diversity of the generation. We choose text complexity and information density to measure the text informativeness which is the key to knowledge-intensive text. We also use the relevance metric to measure factual correctness. In addition, we measure the text repetition to assure no neural degeneration occurs. In the experiments section, we show these evaluation metrics perform as expected while being weakly correlated, making them a good composite metric.

Text Complexity

For text complexity, we adopt the Flesch–Kincaid readability score (FK) (Kincaid et al., 1975) and Gunning fog index (GF) (Wikipedia, 2022) to measure how difficult an English passage is. The formulas are given by the following, where $|\cdot|$ denotes the cardinality.

$$\text{FK} = 206.8 - 1.015 * \frac{|\text{words}|}{|\text{sentences}|} - 84.6 * \frac{|\text{syllables}|}{|\text{words}|} \quad (1)$$

$$\text{GF} = 0.4 * \frac{|\text{words}|}{|\text{sentences}|} + 100 * \frac{|\text{complex words}|}{|\text{words}|} \quad (2)$$

Text Quality

We adopt the LSTM model from Khushali Thakkar (2019) to score the generated text. The model is trained to score student essays.

Diversity

A good Wikipedia article should contain a diverse lexicon to describe the subject. To measure the lexical diversity, we use the distinct-n metric introduced in Li et al. (2015). It is given by the following formula. (In our experiments we take n equals 2.)

$$\text{Distinct-n} = \frac{|\text{unique n-grams}|}{|\text{words}|} \quad (3)$$

Repetition

Although the noxious problem of repetition is getting less prevalent as the model size grows, given the difficulty of the Wikipedia generation task, from time to time, there are still repetitions in GPT-2 and GPT-3 generated Wikipedia articles. So we include the repetition metric to assure the generated text is not repeating itself. We use the rep-n score from Welleck et al. (2019) to measure the number of repeated n-grams in the generated text. In our experiment, we take n equals 4. The formula is given by

$$\text{Rep-n} = 1.0 - \frac{|\text{unique n-grams}|}{|\text{n-grams}|} \quad (4)$$

Information Density

Given our task of generating knowledge-intensive articles like Wikipedia, evaluating whether the model could generate informative text is important. To measure informativeness, we propose the information density metric. We use spacy to do Named Entity Recognition to extract the entities and then calculate it by the following formula.

$$\text{Information Density} = \frac{|\text{entities}|}{|\text{words}|} \quad (5)$$

Relevance

The relevance between the Wikipedia articles and generated text is a crucial component of our evaluation metrics. We use the S-BERT score and entity overlap to calculate their relevance. We first purpose the Entity overlap metric which intuitively measures the number of entities mentioned both in the generated text and the reference text. It is calculated by the following formula.

$$\text{Entity Overlap} = \frac{|E_1 \cap E_2|}{|E_1 \cup E_2|} \quad (6)$$

E_1 represents the entities in the generated text and E_2 represents the entities in the Wikipedia article. We believe entities including certain terminology, people's names, locations, etc. are good indications of knowledge. Thus we use entity overlap to measure the knowledge of the model.

However, we note that the entity metrics require the and thus synonyms or different forms of the word would be overlooked. Thus we use the S-BERT score (Reimers and Gurevych, 2019) to capture the semantic similarity between the generated text and the original Wikipedia.

We are not using traditional measures of relevance like BLEU or ROUGE because getting n-gram overlap between open-ended generations is very difficult and results in a BLEU score near 0. BLEU score calculated using ScareBLEU is reported in the Appendix (under a scale of 100).

5 Experiments

For long text generation, we let each model generate 20 completions for one prompt and then we select the longest 10 generations to filter out empty and short completions. We also store the original Wikipedia text as the reference text for comparison with generated text. Finally, we apply our evaluation metrics to study the performance of generative models.

Experiment Design We conduct the following three experiments: (1) We conduct an ablation study of the GPT-2 models using different decoding methods and different parameter settings. In practice, we tried nucleus sampling with top-p = [0.5, 0.7, 0.9, 0.95, 1.0], top-k sampling with top-k = [20, 50, 100, 500], and temperature = [0.1, 0.5, 0.9]. (2) We conduct a comparison of model performance on old vs new Wikipedia articles. For the old Wikipedia generation, we randomly select 2000 articles from older Wikipedia that are longer than

400 words and went through the same generation process. We show that, across different parameter settings, the generative models tend to perform better on older Wikipedia (which they have been trained on) than on the New-Wiki dataset. (3) We experiment with different language models including GPT-3, BART, T5, OPT, GLM and BLOOM to study their performance and characteristic. We fix the model hyper-parameters to top-k = 20, top-p = 0.9, temperature = 0.9 in this set of experiments.

The objective of experiment (1) and (2) is to show the evaluation results under our framework are consistent with the design of sampling parameters and correlate well with prior research. Thus, we show the effectiveness of our evaluation framework. The objective of (3) is to apply our evaluation framework to some recently released LLMs and find insights into their performance.

Experimental Setup For GPT-3, we directly used OpenAI's text-davinci-002 API for generation and fine-tuning. The experiments cost roughly $200. For GPT-2, we run the experiments on NVIDIA RTX 3080 and RTX 6000 and each generation takes roughly 12 hours. For OPT, BLOOM, and GLM, we run them with 8*A100 GPUs on Google Cloud for about 12 hours individually.

6 Results

With 7 generative models and 7 evaluation metrics, we conduct a thorough evaluation of state-of-the-art LLMs. Full results are available in the Appendix.

To provide better visualization of the experiment results, histograms in Figures 1,2,3,5 are rescaled into 0 and 1. For mean value, we handcraft the range of the metrics and then use the min-max scaler to rescale them. We also draw the range of plus or minus one standard deviation. We set FK-Score $\in [30, 60]$, essay score $\in [4.5, 5.5]$, entity overlap $\in [0, 1]$ S-BERT $\in [0, 1]$, information density $\in [0, 0.5]$, gunning-fog $\in [10, 25]$, distinct-n $\in [0.5, 1]$. For the standard deviation of the metrics, we directly rescaled it to 0 and 1 using the min-max scaler.

6.1 Correlation with prior research

We first experiment with the different decoding mechanisms and parameters to show that our evaluation metrics would provide results that highly correlate with prior research. This validates the effectiveness of our evaluation framework.

Progress in LLMs

Table 4 shows that the progress from GPT-2 and GPT-3 is significant as Brown et al. (2020). GPT-3 has better essay scores, text complexity, relevance, information density, and notably high entity overlap. In general, the latest LLMs including BLOOM, GLM, OPT have better performance than GPT-2 from most of the metrics. (Although T5 and BART have different characteristics.) This shows the progress of LLMs by training larger models.

Old Wikipedia vs New Wikipedia

We also sample 2000 Wikipedia articles from older Wikipedia articles that the LLM might be trained on and compare the generated text with New-Wiki. As shown in Table 3, we found that when we let GPT-2 perform generation on old Wikipedia articles, across all different parameter settings, the mean value of distinct-n, essay score, and entity overlap increase while the text complexity decreases. (S-BERT and Information Density are roughly the same across all settings.) We believe this shows that as GPT-2 is trained on old Wikipedia and GPT-2 would be able to memorize some of these articles and thus generate text with higher quality. This verifies the data contamination issue and model memorization of LLM discussed in Elangovan et al. (2021) and Magar and Schwartz (2022). Thus, our New-Wiki dataset is needed as an out-of-sample test set.

Nucleus sampling

As shown in figure 1, we found that when increasing the top-p value, distinct-n, essay score, and text complexity scores would increase while the relevance score and rep-p metric would decrease. This is consistent with the design of nucleus sampling where a high top-p value leads the model to output tokens with lower probability and often harder and unexpected.

We also note that a lower top-p value leads to bad generations repetitions in the generated articles. This agrees with DeLucia et al. (2020) which argues that top-p around 0.9 is the best parameter for nucleus sampling. The issue of repetition also makes the standard deviation high since the score for repetitive articles is more extreme.

Top-k sampling

As shown in Figure 2, when increasing the top-k value, distinct-n, essay score, and text complexity scores would increase while the relevance score

Model	FK-Score (↓)	Essay scoring	Entity Overlap	S-BERT	Info density	Gunning_fog	Distinct_n	Rep_n(↓)
Bart	**30.503**	5.159	0.216	0.57	0.131	**24.129**	0.692	**0.004**
T5	57.491	5.012	0.227	0.669	0.108	18.937	0.676	0.010
GPT-2	60.456	5.048	0.3	0.702	0.106	18.563	0.679	0.021
GPT-3 Curie	49.032	5.134	0.746	0.76	0.15	22.892	0.655	0.009
OPT-66B	53.741	5.114	0.324	0.72	0.118	19.416	0.702	0.035
GLM	50.812	**5.157**	0.291	0.692	0.122	19.882	0.543	0.208
BLOOM	55.794	5.081	0.249	0.603	0.112	19.523	**0.694**	0.036
GPT-3 With Penalty	50.161	5.119	0.242	0.628	0.142	22.689	0.661	0.006
GPT-3 Davinci	47.502	5.139	**0.778**	**0.762**	**0.153**	23.503	0.639	0.008
Wikipedia	52.646	5.057	1.000	1.000	0.111	21.424	0.692	0.007

Table 3: Mean of evaluation metrics of different LLM

Model	FK-Score (↓)	Essay scoring	Entity overlap	S-BERT	Info density	Gunning fog	Distinct-n	Rep-n	Top-p	Top-k	Temp	Wiki time
GPT-2	**60.711**	5.046	0.304	0.701	0.108	**18.295**	0.685	0.035	0.9	50	0.9	new
GPT-2	61.651	**5.061**	**0.354**	0.700	0.107	18.024	**0.695**	0.035	0.9	50	0.9	old
GPT-2	**60.127**	5.050	0.296	0.698	0.106	**18.858**	0.684	0.018	0.95	50	0.9	new
GPT-2	60.987	**5.062**	**0.35**	0.699	0.105	18.607	**0.693**	0.017	0.95	50	0.9	old
GPT-2	**59.574**	5.051	0.288	0.695	0.103	**19.311**	0.683	0.01	1	50	0.9	new
GPT-2	60.274	**5.063**	**0.339**	0.694	0.103	19.175	**0.692**	0.008	1	50	0.9	old
GPT-2	**58.568**	5.053	0.279	0.688	0.102	**19.875**	0.684	0.007	1	100	0.9	new
GPT-2	59.289	**5.068**	**0.328**	0.688	0.102	19.651	**0.693**	0.005	1	100	0.9	old
GPT-2	**56.956**	5.065	0.259	0.675	0.102	**20.685**	0.689	0.005	1	500	0.9	new
GPT-2	57.756	**5.079**	**0.302**	0.672	0.101	20.503	**0.698**	0.002	1	500	0.9	old

Table 4: Mean of the evaluation metrics when changing the Wikipedia creation time

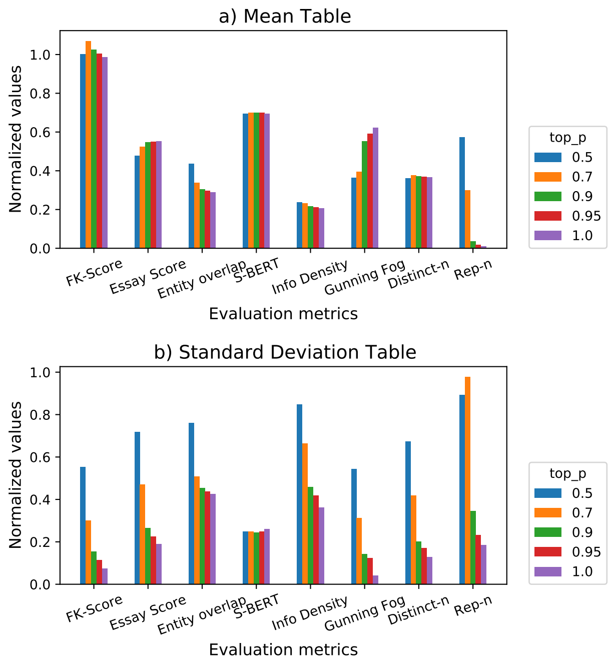

Figure 1: Mean and standard deviation of the evaluation metrics when changing the top-p value.

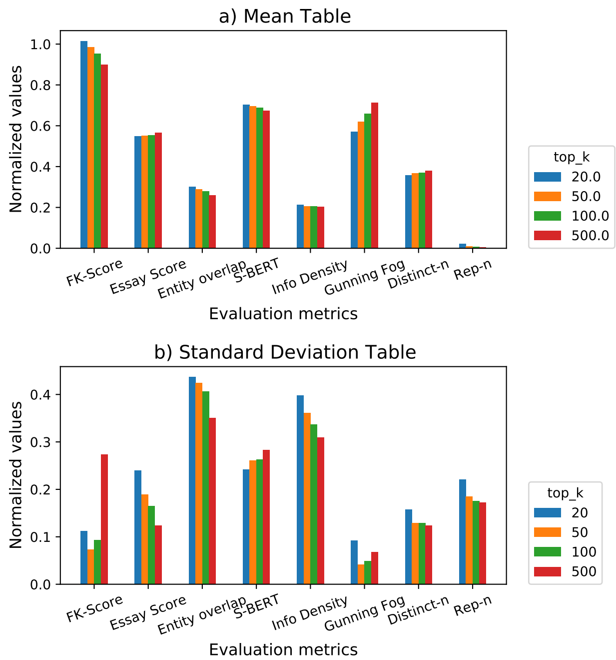

Figure 2: Mean and standard deviation of the evaluation metrics when changing the top-k value.

decreases. This is consistent with the idea of top-k sampling. We also note that the change in performance when varying top-k value is smaller than varying top-p value, which is also observed in (Holtzman et al., 2019)

Temperature

As shown in figure 3, temperature appears to be the parameter that has the most significant effect on GPT generation. When we increase temperature, the essay score and distinct-n metric increase significantly, while the relevance score decreases significantly compared to top-p and top-k sampling. This is consistent with the design of temperature where the model with high temperature is expected to be more creative and decodes tokens that are less expected tokens (and often less frequent and harder words). This is similar to the prior observation that when lowering temperature improves generation quality, it decreases the text diversity (Zhang et al., 2020).

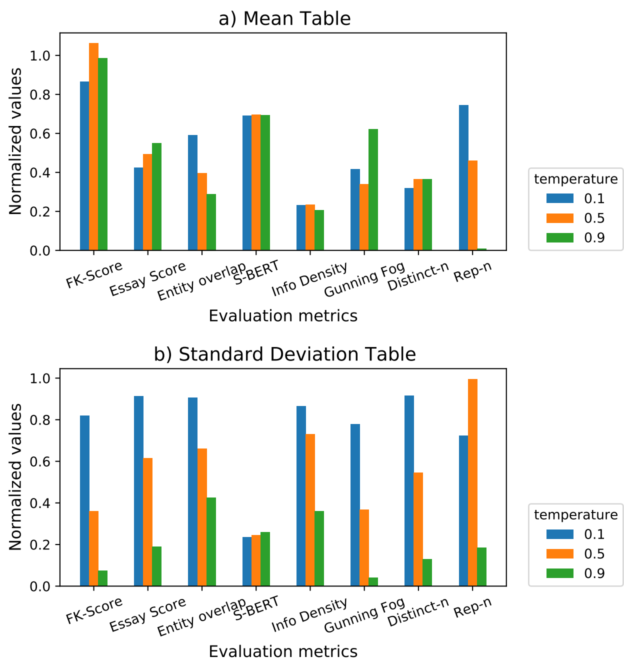

Figure 3: Mean and standard deviation of the evaluation metrics when changing the temperature

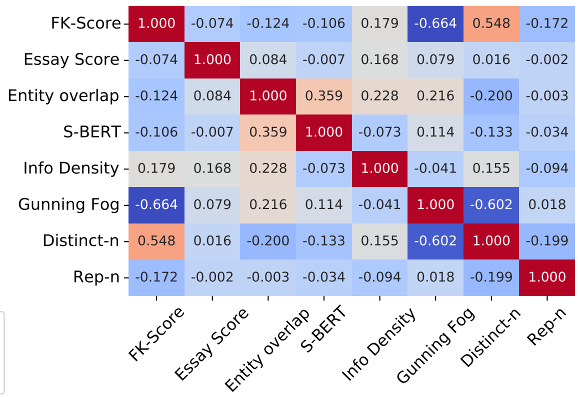

Figure 4: Correlation between evaluation metrics

6.2 Independence of Evaluation Metrics

In Figure 4, we present the correlation matrix across our metrics. We find that the majority of the metrics (from different evaluation buckets) in our evaluation framework are weakly correlated. This shows that we successfully selected evaluation metrics from different perspectives and that each metric could measure relatively independent characteristics of the LLM.

The only two sets of metrics that are highly correlated are text complexity (FK-Score and Gunning Fog Index) and text diversity (distinct-n) as both perspectives would favor harder words. The high repetition score hurts the model performance as expected since it is negatively correlated with relevance, essay score, and information density.

6.3 New insights of large language models

After showing the effectiveness of our evaluation framework, we apply it to evaluate different state-of-the-art LLMs and study their performance on long text generation. Notably, we provide evaluation on the largest LLMs as of 2022: GPT-3, OPT, GLM, and BLOOM. We provide the first set of third-party comparisons of these models. The experiment results are shown in Table 4 and discussed below.

GPT-3

GPT-3's generated text is better than all other models based on our metrics. Table 4 shows that GPT-3 has the highest Entity Overlap, S-BERT score, information density, and top 3 Essay Score and text complexity. In terms of the relevance metrics, GPT-3 achieves an extremely high entity overlap score of 0.778, meaning that the majority of the entities in the real Wikipedia are also mentioned in GPT-3's generated text. This demonstrates GPT-3's memorization ability.

BART and T5 vs. GPT-2

The BART and T5 models have very different characteristics from GPT. BART generates significantly harder words (distinct-n ↑) and harder text (gunning fog ↑). Having a higher essay score shows that these complicated words are composed together correctly but writing hallucinated passages such as with BART gives the lowest relevance score. In contrast, T5 generates simpler text (low information density and essay score) but its relevance score is significantly lower than all GPTs. Table 2 shows that BART and T5 have fewer parameters than GPT-2. This is thus support for larger language models being able to store more world knowledge.

OPT, BLOOM, GLM vs. GPT-3

OPT, BLOOM, and GLM are state-of-the-art LLMs released in 2022. Table 4 shows that their performance is significantly better than older versions of LLMs in most of the dimensions. Among these four, one can see that GPT-3 Davinci has the best overall performance, with notably higher scores in entity overlap and S-BERT score. OPT-66B and BLOOM perform reasonably well, with high text complexity and good relevance score–between GPT-2 and GPT-3. We also note that the high rep-n score for GLM indicates it is generating low-quality language. This is consistent with our manual checking, where we found sentence repetitions and trailing symbols (See Appendix).

LLM's stability

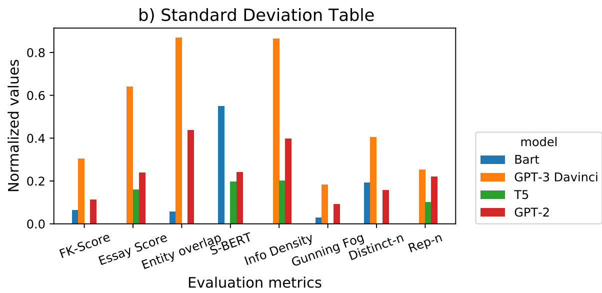

Figure 5: Standard deviation of evaluation matrix of different models

Although larger models like GPT-3 have higher scores, they also have a standard deviation, indicating that larger models are actually less stable. Figure 5 shows that the standard deviations of the metrics roughly follow this pattern: GPT-3 > GPT-2 > BART > T5. This is roughly the order of the number of parameters of these generative models. This observation leads us to hypothesize that since larger models like GPT-3 are more knowledgeable, they would have enough knowledge and thus might have the confidence to "take the risk" and output something that is more specific and risky.

Presence and Frequency Penalty

When we increase the presence and frequency penalty for GPT-3, model got penalized for generating tokens that have been used and thus force the model to change topics more frequently. Thus, we see a significant drop in the relevance between the generation and real Wikipedia (entity overlap score ↓, S-BERT score ↓) and an increase in the word diversity (distinct-n ↑). Interestingly, applying the presence and frequency penalties also hurts the quality of the generated text; essay score, information density, and gunning fog all decreased compared to GPT-3. We hypothesize that in particular in the setting of Wikipedia generation, this is because the penalty decreases the probability of generating repetitive entities and thus decreases the total number of generated entities. Thus, the generated text's complexity and informativeness would all decrease.

7 Conclusions

This work provides a new evaluation framework for LLM's long text generation ability. We first identified Wikipedia as a good-quality, constantly evolving reference text and collected the New-Wiki dataset. Then we propose task of Wikipedia generation and provided a set of automatic well-rounded metrics to help researchers evaluate their generative language models' performance from multiple perspectives. We then conducted extensive experiments on GPT-2 to show that the evaluations under our framework are consistent with the parameter design, and correlate highly with prior research. This verifies the effectiveness of our evaluation framework. Lastly, we appled our evaluation framework to a variety of state-of-the-art generative language models and found interesting characteristics of these models.

8 Limitations and Risks

Limited by computing resources, we did 10 completions for 100 prompts with OPT, GLM, and BLOOM. This may not enough data to make comparison with other models, where we did 20 completions on 2000 prompts.

While our new evaluation metric does not create any risk, we note that the Wikipedia articles may be biased and thus might lead our New-Wiki eval to be biased.

References

BigScience. 2022. Bloom. https://huggingface.co/bigscience/bloom. Accessed: 2022-08-14.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

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. *arXiv preprint arXiv:2107.00061*.

Alexandra DeLucia, Aaron Mueller, Xiang Lisa Li, and João Sedoc. 2020. Decoding methods for neural narrative generation. *arXiv preprint arXiv:2010.07375*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2021. Glm: General language model pretraining with autoregressive blank infilling.

Aparna Elangovan, Jiayuan He, and Karin Verspoor. 2021. Memorization vs. generalization : Quantifying data leakage in NLP performance evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1325–1335, Online. Association for Computational Linguistics.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The webnlg challenge: Generating text from rdf data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Aremu Anuoluwapo, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna Clinciu, Dipanjan Das, Kaustubh D Dhole, et al. 2021. The gem benchmark: Natural language generation, its evaluation and metrics. *arXiv preprint arXiv:2102.01672*.

Daniel Hewlett, Alexandre Lacoste, Llion Jones, Illia Polosukhin, Andrew Fandrianto, Jay Han, Matthew Kelcey, and David Berthelot. 2016. Wikireading: A novel large-scale language understanding task over wikipedia. *arXiv preprint arXiv:1608.03542*.

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

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. Q{2}: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. *arXiv preprint arXiv:2104.08202*.

David M Howcroft, Anja Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A Hasan, Saad Mahamood, Simon Mille, Emiel Van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: Nlg needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182.

Sankalp Jain Khushali Thakkar. 2019. Project title. https://github.com/sankalpjain99/Automatic-Essay-Scoring.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, et al. The bigscience roots corpus: A 1.6 tb composite multilingual dataset. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2):167–195.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2022. Contrastive decoding: Open-ended text generation as optimization. *arXiv preprint arXiv:2210.15097*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*.

Inbal Magar and Roy Schwartz. 2022. Data contamination: From memorization to exploitation. *arXiv preprint arXiv:2203.08242*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for nlg. *arXiv preprint arXiv:1707.06875*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Ankur P Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. Totto: A controlled table-to-text generation dataset. *arXiv preprint arXiv:2004.14373*.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Christina Sauper and Regina Barzilay. 2009. Automatically generating wikipedia articles: A structure-aware approach. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 208–216.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2019. Neural text generation with unlikelihood training. *arXiv preprint arXiv:1908.04319*.

Wikipedia. 2022. Gunning fog index — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Gunning%20fog%20index&oldid=1067780465. [Online; accessed 15-August-2022].

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *Advances in neural information processing systems*, 32.

Hugh Zhang, Daniel Duckworth, Daphne Ippolito, and Arvind Neelakantan. 2020. Trading off diversity and quality in natural language generation. *arXiv preprint arXiv:2004.10450*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

9 Appendix

9.1 LLMs and its training corpus

Below we list the training data of latest LLMs to show that GPT-3, BLOOM, OPT, and GLM are not trained on our New-Wiki dataset which contains the Wikipedia between July 2021 and December 2021.
1. GPT-3: OpenAI does not release the training data of GPT-3. However, OpenAI confirmed that GPT-3 Davinci-002 was released in January, 2022 and never changed since then. So it never seen our New-Wiki dataset.
2. BLOOM: BLOOM is trained on (Laurençon et al.) and the Wikipedia articles come from (Gao et al., 2020) which is released in 2020.
3. OPT: OPT is trained on Pile (Gao et al., 2020) which is released in 2020.
4. GLM: GLMs are trained on "Wikipedia used by BERT (Devlin et al., 2018)", which is released in 2018.

9.2 Examples of Generations

Prompt: 2022 Russian invasion of Ukraine On 24 February 2022, Russia invaded Ukraine in a major escalation of the Russo-Ukrainian War, which began in 2014. Gneration: The invasion began with a Russian air and missile attack on Ukrainian military and civilian targets, which was followed by a ground invasion by Russian troops. The Ukrainian government and military were caught off guard by the attack, and were quickly overwhelmed by the Russian forces. Within a week, the Russians had captured the Ukrainian capital of Kyiv, and the war appeared to be all but over. However, the Ukrainians soon began to mount a resistance movement, and with the help of Western military aid, they were able to slowly push the Russians back. By the end of the year, the war had reached a stalemate, with neither side able to make significant gains.

Table 5: Example article generated by GPT-3

Prompt: Data for Black Lives Data for Black Lives is a movement of a collection of data scientists, mathematicians, researchers, activists, and organizers of whom have the mission of using data science to create concrete and measurable change in the lives of Black people. Gneration: The data scientists and activists are the people who are most likely to use data science to change the lives of Black people. Data scientists are the people who are most likely to use data science to change the lives of Black people. Data scientists are the people who are most likely to use data science to change the lives of Black people. Data scientists are the people who are most likely to use data science to change the lives of Black people. Data scientists are the people who are most likely to use data science to change the lives of Black people. Data scientists are the people who are most likely to use data science to change the lives of Black people. Data scientists are the people who are most likely to use data science to change the lives of Black people.

Table 7: Repeatitive articles generated by GPT-2 with low top-p value

Prompt: 2021 Central Kalimantan floods
An ongoing series of severe floods have affected the province of Central Kalimantan, Indonesia since early September 2021.
Generation: The floods have affected about 154,000 households and affected 612,000 people. About 15,875 families have been displaced to temporary shelters. According to the government, the floods have caused about US$112M USD in economic damages. http://www.un.org/apps/news/story.asp?NewsID=72413 "The floods have affected the lives of thousands of people in Kalimantan. Some 11.7 million hectares of agricultural land and 13.8 million hectares of forest land have been severely damaged".
— "" msgstr ""
#. module: event #: model:ir.model.fields,field_description:event.field_event_event_type__display_name #: model:ir.model.fields,field_description:event.field_event_registration__display_name
#: model:ir.model.fields,field_description:event.field_event_type__name #: model:ir.model.fields,field_description:event.field_event_type_event_mixin__name msgid "Name" msgstr "Naam" #. module: event #: model:ir.model.fields,field_description:event.field_event_event__note msgid "Notes" msgstr "Notas" #. module: event #. openerp-web #: code:addons/event/static/src/js/tours/event.js:43 #, python-format msgid "Note: You

Table 6: Example of degenerated articles created by GLM. Degeneration by GLM is more prevalent than other LLMs. Examples include switching languages, generating code-like text, and non-character symbols.

Prompt: Data for Black Lives
Data for Black Lives is a movement of a collection of data scientists, mathematicians, researchers, activists, and organizers of whom have the mission of using data science to create concrete and measurable change in the lives of Black people. Generation: They are concerned with how the current state of the criminal justice system, as well as the criminal justice system and the system's systemic racism, neglects the Black people. They also are concerned about how the police and the justice system is treating Black people.
Generation: Black people are at a point now that is critical for making progress towards ending the criminal justice system as a whole and to making our lives a little better. In some ways the problems we face are the result of a systemic problem. In many ways, the problem is that we are in a cycle. It's a cycle that has been going on in our country for too long and that has continued for too long. But there is still a lot of work to be done. We are in a cycle of incarceration and disenfranchisement and of criminal justice. And that's something that we need to be very clear about.
AMY GOODMAN: We turn now to the latest piece by the Black Lives Matter movement: a book by the American Civil Liberties Union. And in it, you have some of the first pieces on this.
DAVID KAMM: The civil rights movement is a movement to end the racial inequalities in America. It's a movement to end racism.

Table 8: Articles generated by GPT-2 with high top-p value

9.3 Full Experiment Result

The following table is the major experiments conducted. It evaluate 50 different models and corresponding parameter settings with our evaluation metrics. (Note BLEU score is under the scale of 100.)

Model	Flesch	Essay scoring	Entity Overlap	S-BERT	info_density	Gunning_fog	Distinct_n	Rep-n	BLEU	Top-p	Top-k	Temp	Wiki time
BLOOM	55.794	5.081	0.249	0.603	0.112	19.523	0.694	0.036	0.10	1	20	0.9	new
BART	30.503	5.159	0.216	0.57	0.131	24.129	0.692	0.004	0.08	1	20	0.9	old
GLM	50.812	5.157	0.291	0.692	0.122	19.882	0.543	0.208	0.06	1	20	0.9	old
GPT-3 Curie	49.032	5.134	0.746	0.76	0.15	22.892	0.655	0.009	0.08	1	20	0.9	new
GPT-3 Davinci	47.502	5.139	0.778	0.762	0.153	23.503	0.639	0.008	0.08	1	20	0.9	new
GPT-3 with penalty	50.161	5.119	0.242	0.628	0.142	22.689	0.661	0.006	0.09	1	20	0.9	new
OPT-66B	53.741	5.114	0.324	0.72	0.118	19.416	0.702	0.035	0.07	1	20	0.9	new
T5	57.491	5.012	0.227	0.669	0.108	18.937	0.676	0.01	0.16	1	50	0.9	new
GPT-2	53.064	4.899	0.631	0.689	0.114	17.552	0.654	0.782	0.09	0.5	50	0.1	new
GPT-2	56.235	4.929	0.566	0.69	0.116	16.094	0.662	0.726	0.10	0.5	50	0.5	new
GPT-2	60.038	4.977	0.436	0.694	0.119	15.46	0.68	0.573	0.09	0.5	50	0.9	new
GPT-2	53.637	4.931	0.651	0.691	0.116	17.447	0.656	0.797	0.11	0.5	50	0.1	old
GPT-2	57.731	4.941	0.591	0.691	0.117	15.66	0.672	0.742	0.12	0.5	50	0.5	old
GPT-2	61.521	4.993	0.479	0.694	0.118	15.275	0.69	0.598	0.12	0.5	50	0.9	old
GPT-2	54.631	4.917	0.608	0.691	0.116	16.816	0.658	0.759	0.14	0.7	50	0.1	new
GPT-2	58.781	4.946	0.521	0.692	0.117	15.259	0.671	0.685	0.12	0.7	50	0.5	new
GPT-2	62.099	5.024	0.338	0.699	0.116	15.927	0.688	0.298	0.11	0.7	50	0.9	new
GPT-2	55.794	4.937	0.628	0.692	0.117	16.637	0.659	0.775	0.10	0.7	50	0.1	old
GPT-2	55.651	4.924	0.596	0.691	0.116	16.298	0.66	0.751	0.12	0.9	50	0.1	new
GPT-2	61.066	4.97	0.457	0.693	0.118	14.906	0.676	0.596	0.11	0.9	50	0.5	new
GPT-2	60.711	5.046	0.304	0.701	0.108	18.295	0.685	0.035	0.10	0.9	50	0.9	new
GPT-2	56.835	4.938	0.613	0.693	0.118	16.213	0.662	0.764	0.09	0.9	50	0.1	old
GPT-2	61.889	4.978	0.491	0.695	0.117	14.59	0.685	0.612	0.13	0.9	50	0.5	old
GPT-2	61.651	5.061	0.354	0.7	0.107	18.024	0.695	0.035	0.14	0.9	50	0.9	old
GPT-2	55.249	4.92	0.593	0.692	0.116	16.355	0.656	0.748	0.11	0.95	50	0.1	new
GPT-2	61.277	4.979	0.432	0.694	0.118	14.892	0.679	0.55	0.11	0.95	50	0.5	new
GPT-2	60.127	5.05	0.296	0.698	0.106	18.858	0.684	0.018	0.12	0.95	50	0.9	new
GPT-2	57.292	4.941	0.614	0.694	0.116	16.132	0.663	0.762	0.08	0.95	50	0.1	old
GPT-2	62.621	4.989	0.467	0.695	0.117	14.533	0.687	0.565	0.14	0.95	50	0.5	old
GPT-2	60.987	5.062	0.350	0.699	0.105	18.607	0.693	0.017	0.11	0.95	50	0.9	old
GPT-2	56.125	4.925	0.591	0.691	0.116	16.154	0.66	0.746	0.12	1	20	0.1	new
GPT-2	61.668	4.988	0.407	0.697	0.117	15.03	0.679	0.479	0.13	1	20	0.5	new
GPT-2	60.456	5.048	0.300	0.702	0.106	18.563	0.679	0.021	0.12	1	20	0.9	new
GPT-2	57.228	4.939	0.61	0.692	0.117	15.926	0.664	0.759	0.11	1	20	0.1	old
GPT-2	56.000	4.926	0.590	0.691	0.116	16.244	0.659	0.746	0.11	1	50	0.1	new
GPT-2	61.869	4.993	0.397	0.696	0.118	15.081	0.683	0.459	0.10	1	50	0.5	new
GPT-2	59.574	5.051	0.288	0.695	0.103	19.311	0.683	0.01	0.13	1	50	0.9	new
GPT-2	57.342	4.946	0.608	0.694	0.115	16.244	0.663	0.76	0.11	1	50	0.1	old
GPT-2	62.533	5.003	0.435	0.697	0.116	14.773	0.69	0.473	0.08	1	50	0.5	old
GPT-2	60.274	5.063	0.339	0.694	0.103	19.175	0.692	0.008	0.09	1	50	0.9	old
GPT-2	55.558	4.927	0.589	0.691	0.116	16.252	0.66	0.745	0.11	1	100	0.1	new
GPT-2	61.681	4.991	0.393	0.696	0.118	15.17	0.682	0.455	0.07	1	100	0.5	new
GPT-2	58.568	5.053	0.279	0.688	0.102	19.875	0.684	0.007	0.16	1	100	0.9	new
GPT-2	57.429	4.942	0.609	0.692	0.117	15.895	0.664	0.758	0.12	1	100	0.1	old
GPT-2	62.667	5.005	0.435	0.696	0.117	14.77	0.69	0.467	0.15	1	100	0.5	old
GPT-2	59.289	5.068	0.328	0.688	0.102	19.651	0.693	0.005	0.08	1	100	0.9	old
GPT-2	55.92	4.926	0.59	0.691	0.116	16.263	0.659	0.745	0.14	1	500	0.1	new
GPT-2	61.765	4.993	0.392	0.695	0.117	15.091	0.683	0.449	0.09	1	500	0.5	new
GPT-2	56.956	5.065	0.259	0.675	0.102	20.685	0.689	0.005	0.12	1	500	0.9	new
GPT-2	56.897	4.939	0.61	0.692	0.117	16.062	0.663	0.76	0.11	1	500	0.1	old
GPT-2	57.756	5.079	0.302	0.672	0.101	20.503	0.698	0.002	0.10	1	500	0.9	old