

# Can Very Large Pre-trained Language Models Learn Storytelling With A Few Examples?

Zhuohan Xie      Trevor Cohn      Jey Han Lau

School of Computing and Information Systems

The University of Melbourne

zhuohanx@student.unimelb.edu.au, {t.cohn, jeyhan.lau}@unimelb.edu.au

## Abstract

While pre-trained language models can generate individually fluent sentences for automatic story generation, they struggle to generate stories that are coherent, sensible and interesting. Current state-of-the-art (SOTA) story generation models explore using higher-level features such as plots or commonsense knowledge to improve the quality of generated stories. Prompt-based learning using very large pre-trained language models (VLPLMs) such as GPT3 has demonstrated impressive performance even across various NLP tasks. In this paper, we present an extensive study using automatic and human evaluation to compare the story generation capability of VLPLMs to those SOTA models in three different datasets where stories differ in style, register and length. Our results show that VLPLMs generate much higher quality stories than other story generation models, and to a certain extent rival human authors, although preliminary investigation also reveals that they tend to “plagiarise” real stories in scenarios that involve world knowledge.

## 1 Introduction

Automatic story generation is a challenging task as—beyond having individually fluent sentences—a story as a whole needs to have a natural flow, obey commonsense and be interesting. Nowadays, most works fine-tune pre-trained language models such as GPT2 (Radford et al., 2019) and BART (Lewis et al., 2020) on a particular dataset to learn story generation. These story generation models can generally produce fluent sentences without any obvious grammar issues, but they often fail to form a coherent story that obeys commonsense, let alone create an interesting narrative (See et al., 2019; Guan et al., 2021a). To address these issues, models explore using higher level features such as plots and commonsense knowledge to aid story generation.

Prompt-based learning (Liu et al., 2021) is a relatively new paradigm that is designed for very

large pre-trained language models (VLPLMs) such as GPT3 (Brown et al., 2020). Unlike the standard “pre-train and fine-tune” paradigm which requires a substantial amount of data for fine-tuning, in prompt-based learning, VLPLMs are provided with several examples as a “prompt” to learn a task without gradient-based fine-tuning (Liu et al., 2021), which thus can be interpreted as a form of few-shot learning. Prompt-based learning has recently demonstrated impressive performance in various language generation tasks<sup>1</sup> such as essay writing (Elkins and Chun, 2020), dialogue generation (Madotto et al., 2021), question generation and answering (Mishra et al., 2022). Clark et al. (2021) explore the use of prompt-based learning for GPT3 to generate stories in several domains, and find that the generated stories are indistinguishable from human references, which provides preliminary evidence to support that this approach can generate good stories. See et al. (2019) compares storytelling capabilities between GPT2 and the Fusion model (Fan et al., 2018), the leading story generation model at that time. However, the definition of “SOTA” has evolved as time goes, and no work has provided systematic and empirical comparison between VLPLMs and SOTA models in the story generation domain.

In this paper, we aim to fill in this gap by presenting an extensive evaluation of automatic story generation comparing VLPLMs with prompt-based learning to SOTA models. We compare generated stories in terms of various automatic evaluation metrics from lexical and semantic matching ones to recently proposed model-based ones. We follow the best practice in literature to conduct rigorous human evaluations including both crowdworkers from Amazon Mechanical Turk and in-house judges, and assess story quality at a fine-grained level, such as coherence and logicity. To summarise, our contributions are:

<sup>1</sup><https://beta.openai.com/examples/>

- We conduct an empirical comparison between GPT3 and other SOTA techniques for open-ended story generation on **three different corpora that differ in style, register, and length.**
- We test with a wide variety of automatic story evaluation metrics, and find that recent model-based ones work better, consistent with the literature.
- We experiment with two types of annotators (crowdworkers and in-house judges) to assess the quality of stories across several aspects, and find consistent results between them.
- Our results systematically demonstrate that the stories generated by GPT3 are substantially better than those by SOTA methods, and are on par with (or even surpass) human-authored stories in most aspects.
- We conduct a preliminary study on story plagiarism and find that GPT3 tends to (soft) “plagiarise” real stories when generating news, even though it does not directly copy the source text, raising further questions as to what extent GPT3 recycles stories in its memory rather than generating new narratives.

## 2 Related Work

**Story Generation** See et al. (2019) find that although GPT2 can generate fluent sentences, more attentions are needed to incorporate commonsense and higher-level story planning. Most works then generally use pre-trained models such as GPT2 and BART as the backbone and incorporate higher level features to aid the generation process, and get the SOTA results. Specifically, Rashkin et al. (2020); Goldfarb-Tarrant et al. (2020); Tan et al. (2021) construct a storyline to guide the generation process. Guan et al. (2021a); Yu et al. (2021) incorporate inter-sentence relationships such as coherence and discourse relationships into the generation process. Guan et al. (2020); Peng et al. (2021) explore using external knowledge such as commonsense for story generation. Xu et al. (2020); Ammanabrolu et al. (2021) combine storyline planning and commonsense reasoning.

We find that although there are studies which explore the use of GPT3 for story generation. For example, Clark et al. (2021) conducts a Turing test between GPT3-generated and human-written stories and Lucy and Bamman (2021) probe for

gender and representation bias in GPT3-generated stories. However, they do not provide a systematic evaluation that assess GPT3 against the SOTA story generation models.

**Story Evaluation** Automatic story evaluation is admittedly a challenging task, and the lack of standardised evaluation metrics has somewhat impeded progress of story generation (Guan et al., 2021b). Human evaluation is usually considered as the gold standard for story quality evaluation, but it is expensive and time-consuming (Guan and Huang, 2020) and it can not capture diversity (Hashimoto et al., 2017). Many automatic evaluation metrics, which typically assess story quality (how well a story reads) and diversity (how much variation do the generated stories have), are then introduced as substitutes for assessing story quality; most of these metrics measure lexical overlap between strings (Papineni et al., 2002; Lin, 2004; Tan et al., 2021) or semantic similarity by comparing embedding of models (Zhao et al., 2019; Zhang et al., 2020) between generated stories and their human references. Recently, learning (Sellam et al., 2020) and generation (Yuan et al., 2021) based methods are explored and they are based on large pre-trained language models such as BERT (Devlin et al., 2019) and BART. However, these evaluation metrics can only produce one single score that indicates the overall story quality, and few metrics are designed to measure specific aspects (Chhun et al., 2022) such as logicity (does a story obey your commonsense?) or interestingness (how much do you like the story?).

## 3 Experimental Setup

### 3.1 Story Generation Models

To make a comprehensive comparison between GPT3 and SOTA story generation models, we run extensive experiments on GPT3 and many SOTA story generation models.

For GPT3, we use the largest initial version (text-davinci-001), which was first released in June 2020 and has 175B parameters. Note that this model differs in several ways from a vanilla language model trained with next word prediction, as OpenAI has documented many enhancements.<sup>2</sup> We

<sup>2</sup>For example, text-davinci-001 is likely to be fine-tuned further to produce text that contains appropriate societal values/norms and to take natural language instructions as input. Details of these enhancements are provided in Appendix Table 9.

suspect its story generation ability benefits from these enhancements even though they are not directly applied to improve story generation.<sup>3</sup> For GPT3 we use prompt-based learning to adapt it to a story domain without any explicit fine-tuning. We select a number of stories as training prompts (2 or 3 depending on the domain).

For SOTA story generation models, we use (1) knowledge enhanced based models: KGGPT2 (Guan et al., 2020) and HINT (Guan et al., 2021a); (2) storyline planning based model: PROGEN (Tan et al., 2021); and (3) MTCL (Xu et al., 2020) that combines both storyline planning and commonsense reasoning. We also fine-tune BART as an additional baseline. For consistency, all models use nucleus sampling (Holtzman et al., 2020) with  $p = 0.95$  as the decoding method. We summarise these models in Table 1, and more details can be found in Appendix A.

### 3.2 Story Datasets

The most popular story dataset is ROCStories (ROC) (Mostafazadeh et al., 2016), which is composed of short commonsense stories and is used by most story generation models. There are also more difficult and longer story datasets, such as WritingPrompts (WP) (Fan et al., 2018) and CNN News (CNN) (Hermann et al., 2015) which are composed of fictional and news stories (two different domains). In our experiments, we run all models on all these datasets. Specifically, ROC is used to test the generation of short stories (5 sentences), WP for medium-length stories (trimmed to 10 sentences) and CNN for long stories (over 20 sentences). More details about these datasets can be found in Appendix B.

Whenever possible we evaluate models on all story datasets. However, this is sometimes infeasible because some models are designed to work on a particular dataset and thus cannot be adapted to other datasets easily. Moreover, we focus on conditional story generation in this work, this means there is some *context* upon which we generate the stories (detailed next).

**ROC** We evaluate all models in this dataset. The context we use to generate stories is the first sentence, and so the models are trained to generate the last 4 sentences. Evaluation results are computed

over 800 generated stories using randomly sampled leading sentences from the test partition.

**WP** We assess HINT, PROGEN, GPT3 and BART on this dataset. The context is a short paragraph (“prompt”) that describes the idea of the story. We randomly sample 1000 prompts from the test partition for automatic evaluation.

**CNN** We only run GPT3, BART, PROGEN on CNN, as HINT is developed for ROC and WP originally and it does work well when applied to CNN. Stories of CNN are generated conditioned on the news titles. We randomly sample 600 titles from the test partition for automatic evaluation.

## 4 Automatic Evaluation

### 4.1 Evaluation Metrics

We use two types of automatic evaluation metrics: (1) reference-based metrics, where we compare the generated stories to human reference stories based on the same conditioning context; and (2) reference-free metrics, where we assess the quality of the stories directly.

#### 4.1.1 Reference-based Metrics

Most reference-based metrics measure the semantic closeness between generated stories and their human references. We experiment with metrics based on string based matching (CBL, MSJ) and embedding based matching (BES) and a learning based metric (BRT), to assess the quality of generated stories. We also use a recall based metric (BBL) to assess the diversity of generated stories. Specifically, **Corpus BLEU (CBL)** computes the average BLEU scores (Papineni et al., 2002) for each generated story against all human references (Caccia et al., 2020; Xie et al., 2021). **MS-Jaccard (MSJ)** measures lexical overlap by computing the n-gram overlap between generated and referenced stories using the Jaccard index (Alihosseini et al., 2019). **BERTScore (BES)** measures the maximum similarity of each token’s contextual embedding between generated and referenced stories (Zhang et al., 2020). **BLEURT (BRT)** is trained on synthetic data to predict a similarity score between generated and referenced stories (Sellam et al., 2020). **Backward BLEU (BBL)** computes the coverage of n-grams in the reference stories against the set of generated stories (Shi et al., 2018).<sup>4</sup>

<sup>3</sup>Noting that a 3rd version (text-davinci-003) has been released with further enhancements at the time of writing.

<sup>4</sup>We use BLEU4 for CBL and BBL; 4-grams overlap for MSJ; roberta-large model for BES; bert-base-128 for BRT.

Model	Backbone	Size	Method	Story Datasets
GPT3	text-davinci-001	175B	Prompt-based learning with several examples from the story dataset (3 for ROC and WP and 2 for CNN)	ROC, WP, CNN
KGGPT2	GPT2-small	124M	Fine-tuned on commonsense data before more fine-tuning with auxiliary classification tasks	ROC
PROGEN	BART-large	400M	three-stage generation where at each stage a fine-tuned BART generates stories based on word importance in the story datasets	ROC, WP, CNN
MTCL	GPT2-small BERT-large	124M 336M	(1) a GPT2 model to generate keywords; (2) a BERT model to rank retrieved knowledge triples; and (3) a second GPT2 model that takes top-ranked knowledge triples and context as input for story generation	ROC
HINT	BART-base	140M	BART is first fine-tuned on BookCorpus with additional objectives to learn internal structure in a story and then further fine-tuned on the story datasets	ROC, WP
BART	BART-large	400M	Baseline model that is fine-tuned on the story datasets using a standard language modelling objective	ROC, WP, CNN

Table 1: The backbone (“Backbone”) of the story generation models and their number of parameters (“Size”). “Story Datasets” indicates which datasets are used to generate stories for a particular model. KGGPT2 and MTCL stories are obtained from the original authors; for PROGEN and HINT we re-run the implementation provided by the authors.

#### 4.1.2 Reference-free Metrics

Reference-free metrics evaluate generated stories without comparing them to their human-authored references. We experiment with diversity metrics based on intra-story (D-3, LR-n) and inter-story diversity (SBL). We also compute negative log-likelihood from BART of a story conditioned on the context (BAS) for relatedness, and story length in terms of words (LEN) for complexity.

Specifically, **Lexical Repetition (LR-n)** computes the average percentage of 4-grams appearing at least  $n$  times in the generated stories (Shao et al., 2019). **Distinct-3 (D-3)** computes the average ratio of distinct 3-grams to all 3-grams (Li et al., 2016). **Self-BLEU (SBL)** measures inter-story diversity that computes the average BLEU score of each generated story using all generated stories as reference (Zhu et al., 2018). **BARTScore (BAS)** computes perplexity of a generated story conditioned on the context (i.e. leading sentence for ROC, prompt for WP and title for CNN) to measure how much a generated story relates to the condition (Yuan et al., 2021).<sup>5</sup> **Length (LEN)** measures the average length of the generated stories, which is used as a rough indicator of generation complexity.

<sup>5</sup>We set  $n = 3/8/8$  for ROC, WP and CNN respectively and use BLEU4 for SBL. We use the “PARA” version of BART and direction as “from source to hypothesis”.

#### 4.2 Results

Table 2 and Table 3 present the reference-based and reference-free evaluation results, respectively. At a glance, these metrics do not appear to agree with each other even though some of them are designed to evaluate the same aspect (e.g., the best model in terms of fluency/coherence or diversity is different depending on the metric). Overall, GPT3 seems to have weaker performance than most of other models in terms of lexical based quality (CBL and MSJ) and diversity (BBL, SBL, D-3 and LR-n) metrics.

However, when we look at recent model-based metrics (BERTScore, BLEURT and BARTScore), GPT-3 appears to be a much better model (a finding we will return to when we look at human evaluation results). Interestingly, we notice that human written stories have very poor performance in terms of BARTScore (BAS). We suspect that BARTScore is likely to favour machine-generated stories as the metric measures quality based on the probability of a word sequence and machine-generated stories are generated to maximise this, where else human stories are written in a very different manner (e.g., with surprising or creative words (Holtzman et al., 2020)) and thus leads to a low BARTScore. In terms of story length, all models can generate the right length; the one exception here is GPT3 in CNN, which appears to be unable to generate stories longer than 150 words on average (human-



Model		Flu./Coh.				Div.
		CBL	MSJ	BES	BRT	BBL
		↑	↑	↑	↓	↑
ROC	GPT3	27.2	11.6	86.6	<b>8.6</b>	24.0
	KGGPT2	33.5	<b>15.0</b>	<b>87.0</b>	9.5	<b>25.6</b>
	PROGEN3	26.6	14.6	86.7	9.7	25.0
	MTCL	31.4	14.2	86.9	9.7	24.0
	HINT	<b>39.6</b>	13.7	<b>87.0</b>	<b>8.6</b>	24.6
	BART	27.5	14.7	86.8	9.5	25.1
WP	GPT3	28.6	12.3	<b>81.6</b>	<b>11.7</b>	24.4
	PROGEN3	32.3	<b>16.4</b>	81.4	13.3	<b>27.6</b>
	HINT	<b>45.5</b>	12.8	80.8	12.1	23.7
	BART	32.6	16.2	81.4	13.0	27.2
CNN	GPT3	<b>33.2</b>	11.0	<b>83.5</b>	<b>7.5</b>	19.8
	PROGEN3	29.6	<b>14.8</b>	82.2	9.3	<b>26.2</b>
	BART	29.1	14.7	82.2	9.8	25.7

Table 2: Reference-based Evaluation Results. CBL, MSJ, BES and BRT evaluate the closeness between the generated stories and the whole test reference data as an indicator of general fluency (Flu.) and coherence (Coh.). BBL focus on the recall of generated stories as an indicator of diversity (Div.). ↑: higher is better; ↓: lower is better. BRT values are negated here.

written stories have on average 500 words). Overall, if we take into consideration all the automatic metrics, there is no one model that particular stands out in terms of story generation.

## 5 Human Evaluation

We next recruit human annotators to assess the quality of the generated stories. We explore both crowdsource and in-house workers for this, to provide a better understanding on consistency.

### 5.1 Crowdsourcing Annotation

We first collect human judgements using the Amazon Mechanical Turk platform.<sup>6</sup> We follow Karpinska et al. (2021) that propose to assess four aspects (fluency, coherence, relatedness and interestingness) and add one new aspect: logicity, which measures how well a story obeys commonsense. Each of these five aspects is judged on an ordinal scale from 1 (worst) to 5 (best). Survey questions are presented in Appendix C.

We randomly sample 20 conditional contexts (e.g., titles) from each dataset and collect stories generated by all models for human evaluation. Each story (including human-written one) is judged by 3 annotators, and so we have annotations for 320 stories in total (140/100/80 for ROC, WP and CNN,

<sup>6</sup><https://requester.mturk.com/>

Model		Div.			Rel.	Com.
		SBL	D-3	LR-n	BAS	LEN
		↓	↑	↓	↓	↑
ROC	GPT3	38.5	67.7	39.1	<b>4.2</b>	47.3
	KGGPT2	41.9	67.2	51.9	4.6	38.4
	PROGEN3	<b>30.0</b>	76.9	39.5	5.0	40.9
	MTCL	39.4	69.6	44.4	4.9	<b>49.7</b>
	HINT	55.1	54.3	68.1	4.3	35.8
	BART	30.5	77.4	37.8	5.0	40.6
	human	33.1	<b>80.2</b>	<b>35.8</b>	5.2	40.3
WP	GPT3	37.5	69.6	9.7	4.3	120.6
	PROGEN3	35.2	77.2	2.6	5.4	136.9
	HINT	64.1	33.9	67.4	<b>4.1</b>	119.0
	BART	35.3	77.5	1.6	5.4	129.2
	human	<b>27.1</b>	<b>83.7</b>	<b>1.5</b>	5.7	<b>150.0</b>
CNN	GPT3	<b>26.5</b>	82.9	9.8	<b>4.4</b>	147.3
	PROGEN3	28.9	82.3	2.3	5.2	395.8
	BART	27.9	83.2	<b>0.8</b>	5.2	374.1
	human	27.3	<b>83.8</b>	6.3	5.4	<b>498.6</b>

Table 3: Reference-free Evaluation Results. SBL measures inter-story diversity by assessing differences between different stories while D-3 and LR-n (3 for ROC, 8 for WP and CNN) focus on repetition n-grams within the same story. We also include LEN as an indicator of story complexity (Com.). We compute BAS of story given condition for story relatedness (Rel.).

respectively). More details on how we control quality and run the survey are given in Appendix D, and the agreement between annotators in Appendix F.

Table 4 presents the human evaluation results. Overall, GPT3 generates stories that are consistently of higher quality than those generated by other SOTA models. To understand whether the difference is significant, we perform a paired t-test by comparing GPT3 to other models (including human) and find that in most cases these results are significant with  $p$ -value  $< 0.05$  (\* in the table).<sup>7</sup> Compared with human authors, GPT3 appears that it is generating stories that are just as good as (ROC) or better than (WP and CNN) human authors, confirming the findings of Clark et al. (2021). There are, however, some explanations behind these observations. For WP, in particular, human stories are trimmed to the first 10 sentences (data pre-processing for training the story generation models). This abruptly shortens the stories so they might not provide a proper conclusion, and inevitably are penalised. In Appendix Table 13, 14 and 15, we provide some examples for this.

For CNN, GPT3 appears to be “plagiarising” real

<sup>7</sup>This also means that even though the number of story contexts per domain (20) is not large, it is sufficient to produce statistically robust observations.

	Model	Flu.	Coh.	Rel.	Log.	Int.
ROC	GPT3	<b>4.40</b>	4.43	4.37	4.37	3.57
	KGGPT2	3.90*	3.48*	3.53*	3.00*	2.62*
	PROGEN3	3.88*	3.45*	3.37*	2.95*	2.57*
	MTCL	3.55*	3.12*	3.18*	2.73*	2.42*
	HINT	3.90*	3.27*	3.33*	3.12*	2.58*
	BART	3.92*	3.38*	3.48*	3.03*	2.60*
	human	4.22	<b>4.58</b>	<b>4.42</b>	<b>4.48</b>	<b>3.77</b>
WP	GPT3	<b>4.37</b>	<b>4.67</b>	<b>4.28</b>	<b>4.48</b>	<b>3.47</b>
	PROGEN3	3.45*	3.08*	2.35*	2.57*	1.98*
	HINT	3.32*	2.63*	2.02*	2.25*	1.77*
	BART	3.42*	2.73*	2.08*	2.27*	1.87*
	human	4.13*	4.22*	3.05*	3.75*	2.97*
CNN	GPT3	<b>4.22</b>	<b>4.52</b>	<b>4.58</b>	<b>4.60</b>	3.20
	PROGEN3	3.63*	3.32*	3.30*	3.22*	2.28*
	BART	3.58*	3.37*	3.30*	3.27*	2.17*
	human	4.10	4.10*	4.23*	4.18*	<b>3.72*</b>

Table 4: Human Evaluation Results. We calculate the average score of models for each aspect: fluency (Flu.), coherence (Coh.), relatedness (Rel.), logicity (Log.) and interestingness (Int.). Model scores that are marked with \* mean the performance difference between the model and GPT3 is significant.

stories (more details in Section 6), where many story elements are not a product of creative generation but details copied from real news stories. Another reason could be that GPT3 stories are much shorter than those generated by other models and human authors (150 vs. 300-400 words; Table 3), which makes them easier to read and thus leads to better scores. Note that this is a downside of GPT3 where it is difficult to get it to generate long stories, which is discussed in Section 7.

Regarding the different aspects for the SOTA models (KGGPT2, PROGEN3, MTCL, HINT and BART), generally most models perform very well in terms of fluency ( $>3.5$  in most cases), showing that they do not struggle to generate natural and fluent sentences. Coherence performance, however, varies depending on the dataset. Most models do well on ROC and CNN, but struggle on WP ( $<3.1$  for PROGEN3, HINT and BART). There is an intuitive explanation for why they struggle more on shorter WP stories than longer CNN stories: all models use pre-trained language models as the backbone which are trained on web data that contains an abundance of news articles. In contrast, WP stories are fictional stories with twists and diverse styles and the pre-trained models are possibly less suited for such stories. For relatedness, logicity and interestingness, we see a similar trend where the models perform best in ROC and worst in WP. We also observe a consistent decrease in

	Model	Flu.	Coh.	Rel.	Log.	Int.
ROC	GPT3	<b>4.78</b>	<b>4.73</b>	<b>4.50</b>	<b>4.82</b>	<b>3.37</b>
	KGGPT2	4.52*	3.67*	3.57*	3.47*	2.50*
	PROGEN3	4.27*	3.47*	3.78*	3.23*	2.48*
	MTCL	4.27*	3.27*	3.45*	3.15*	2.37*
	HINT	4.38*	4.03*	3.38*	3.70*	2.38*
	BART	4.37*	3.95*	3.85*	3.53*	2.70*
	human	4.52*	4.38*	4.22	4.32*	3.18
WP	GPT3	<b>4.57</b>	<b>4.65</b>	<b>4.08</b>	<b>4.22</b>	<b>3.82</b>
	PROGEN3	3.55*	3.03*	2.23*	2.57*	2.45*
	HINT	3.60*	2.72*	2.07*	2.68*	2.08*
	BART	3.45*	2.77*	2.08*	2.38*	2.30*
	human	4.05*	4.07*	3.73	3.87*	3.78
CNN	GPT3	<b>4.50</b>	<b>4.33</b>	<b>4.48</b>	<b>4.40</b>	<b>3.45</b>
	PROGEN3	3.80*	3.45*	3.63*	3.45*	2.52*
	BART	3.73*	3.25*	3.58*	3.32*	2.57*
	human	4.22*	4.00*	4.35	4.13*	3.22

Table 5: In-house Human Evaluation Results.

performance from relatedness to logicity and interestingness, suggesting that the models particularly struggle to generate interesting and sensible stories. Interestingness is perhaps the most difficult aspect to optimise, as it is difficult to define what makes a narrative interesting.

## 5.2 In-house Annotation

We next collect in-house judgements, contributed by voluntary colleagues.<sup>8</sup> We ask them to evaluate the same 5 aspects using the same scale, although we improve the instructions by adding additional examples to highlight the differences between different aspects. We sample 20 *disjoint* conditional contexts from each dataset for story generation here, as we are interested to test the robustness of our previous findings (with different workers and set of stories). As with crowdsource annotation, each story is also judged by 3 annotators. Details of the agreement between annotators can be found in Appendix G.

Table 5 presents the scores of story quality from in-house annotators. Interestingly, the *magnitude* of the in-house scores are generally somewhat higher than the crowdworker scores (across all metrics and datasets and models). We hypothesise that this may be our in-house workers are more “tolerant” to mistakes as they have been exposed to machine-generated text more compared to crowdworkers. That said, the overall findings are consistent between the two groups of annotators: (1) GPT3 is the best story generation model and out-

<sup>8</sup>Demographically, 14 are PhD students and 1 is university staff; all of them are proficient in English.

performs both SOTA models and human stories; (2) The SOTA models do well in fluency, but poorly in most other aspects (interestingness worst); (3) The SOTA models struggle in WP in particular, as evidenced by their poor coherence, relatedness, logicity scores compared to other domains.

Broadly speaking, when we compare the automatic metric results (see Section 4.2) to these human evaluation results we would arrive at a rather different conclusion (that GPT3 is not ahead of the pack and there is no clear “best” story generation model). That said, if we look at model-based metrics (BERTScore, BLEURT for fluency/coherence and BARTScore for relatedness), we would draw a more similar conclusion where GPT appears to be strong model (the trend is still less conclusive compared to human evaluation results though). Overall, the finding that modern model-based metrics correlate with human evaluations better is consistent with the literature (Zhang et al., 2020; Sellam et al., 2020).

## 6 Plagiarism

Considering the strong performance of GPT3 on story generation, we next provide a preliminary investigation to understand the extent to which GPT3 copies from its training data.

**IThenticate** We use iThenticate<sup>9</sup> — a professional plagiarism detection software that has comprehensive coverage over the web, scholarly articles and news — to assess how much GPT3 plagiarises. We include only the generated content (without the condition) when checking for plagiarism. Results show that there is no strong plagiarism: similarity score for ROC, WP and CNN is 4%, 3% and 14% respectively. That said, CNN appears to have a much higher similarity score (which we will further investigate).

**Manual Check** IThenticate looks for lexical overlap to identify plagiarism. A more subtle form of plagiarism is one that copies the ideas without parroting the words (Lee et al., 2022). To check if this kind of plagiarism exists, we randomly sample 10 GPT3 stories for WP and CNN respectively and manually check whether they copy from web articles by searching for keywords, named entities and long phrases. For WP, in most instances we do not find web articles or stories that resemble our stories, except for one instance where it copies a

story from Wikipedia (noting that it is not a direct word-to-word copy). For CNN, we find 7 out of 10 instances are based on real stories from news articles, where most details (e.g., names, incident origins, sentence penalty) are faithful to the original stories. These results highlight that GPT3 can recall details from training data, calling into question how much of its generation is creative. Although we do not find much plagiarism for WP, it is not inconceivable that the overarching storylines or plot details are copied from its massive training data. More details can be found in Appendix Table 17.

## 7 Discussion

In this paper, we only experiment with a simple prompt-based learning approach to adapt GPT3 to the different story datasets and one can argue more prompt engineering can be explored (Liu et al., 2022; Lu et al., 2022; Mishra et al., 2022). We contend, however, that our goal here is not to propose a new or better prompt-based method to do story generation for GPT3, and ultimately even with our simple approach we find that GPT3 generates high quality stories, suggesting that with additional prompt engineering it can do better. We next discuss *qualitatively* some of the strengths and drawbacks of GPT3 in story generation.

### 7.1 Strengths

Unlike other conditional generation tasks such as machine translation or summarisation, where the input context is rich in information and the goal is to translate or compress the input information, story generation works in the reverse manner where the model needs to “hallucinate” new information and details given a succinct context. This means that in order to do the task well, having strong world knowledge is important. Reading some of the GPT3 stories, we observe GPT3’s advantage in this, particularly in the WP dataset where some of the prompts require niche knowledge about characters. In Table 6 we show an example in WP where the prompt is *cthulhu calls your cell , he needs to be picked up*, where *cthulhu* a fictional cosmic entity, and only GPT3 is able to produce a coherent story and the SOTA models struggle.

### 7.2 Drawbacks

Even though GPT3 demonstrates excellent generative capability and outperforms SOTA models significantly, we still find GPT3 has many generation

<sup>9</sup><https://www.ithenticate.com>

Model	Story
GPT3	“ yes , lord cthulhu ? ” you answer the phone . “ i need you to pick me up from the airport . ” “ of course , my lord . when will you be arriving ? ” “ in about an hour . i ’ll text you the flight number . ” you hang up and start getting ready . you know cthulhu is a demanding deity , but you do n’t mind . you enjoy serving him .
PROGEN3	**but everything ’s different ! i miss my child in there ...** *what are you talking about ! you have to be mad . you must have misheard ! ! you come and get your mommy ... ” momma always grabbed him before they climbed into the bathtub . after that happened they were always late . * bam . * wait , it ’s never like this . cthulhu always puts up a fight . i repeat : it ’s never like this ! it ’s always for the umpteenth time , now or never . *the phone rings . i answer it and cthulhu ’s voice booms out .
HINT	“ sir , i ’m afraid i ’m late for work . “ “ i ’m sorry sir , “ the technician said , “ but i have to get my cell phone back . “ “ sir , you have to pick up this cell . “ “ yes sir , i do . “ “ and how are you going to find it ? “ “ i ’m just going to have to pick it up . “ “ well , sir , “ the technician said , “ it ’s just that i do n’t have the time to pick it up . “ “ why not ? “ “ sir , “ the technician said , “ i do n’t have the time . “ “ i ’m just going to make sure it ’s picked up . “
BART	“ ah it ’s here . i thought you wanted to leave , a phone call . * ” my coworkers looked at me confused . “ i wasn ’t in the mood to go running around scared . but it was important . you need to pick up on your cravings . the kid-ta-bye , gabriel , payliver , cilia de los angeles . and four people ’s fault . ” he was right about the fifth , at least my attention was fixed on it . three girls were outside screaming in the street , and a family of four was picking up one of them . the two men yelled and approached as they walked to the mailbox , visibly distressed .

Table 6: Stories generated by GPT3, HINT, BART and PROGEN3 given the same title “cthulhu calls your cell , he needs to be picked up .”

errors that can be improved.

**Story length** GPT3 has a parameter to control the maximum number of generated tokens but does not provides a way to control the minimum number of tokens. As one can see from Table 3, GPT3 can not generate stories longer than 150 words for CNN, even though the prompts have long stories. We also attempted to encourage longer stories by adding specific instructions as part of the prompt of GPT3, but this did not work.

**Null generation** Occasionally GPT3 decides to generate no output. This is usually not an issue, since this can be solved by forcing it to generate again, although it is unclear why this occurs.

**Direct copy** Besides the soft plagiarism issue (Section 6), GPT3 does occasionally copy long chunks of text, e.g., the title or prompt in the story.

**Multilingual** GPT3 sometimes generates stories in languages other than English, despite the given prompts always being in English. In terms of statistics, out of 1000 generations we find 14 non-English stories (5 Chinese, 4 German, 1 Japanese, 1 French, 1 Russian, 1 Norwegian Nynorsk and 1 mixture of Chinese and English). Interestingly, in most of these cases the stories are related to the condition (even though in different languages) although sometimes we observe the outputs are direct translation of the prompt, not a creative story.

**Tokenisation issue** GPT3 generations occasionally feature “sticky” words where there are missing white spaces (e.g., *understand.With* and *timewhen*). We suspect this is due to Byte-Pair Encoding of GPT3 where white spaces are “glued” to each subword and so every subword has two versions (one with the white space and one without). This issue arises when GPT3 generates using a subword without the white space suffix.

**Expletives** GPT3 would occasionally generate stories with expletives. Interestingly, it would sometimes self-censor them (e.g., *b\*\*\*\**).

Example stories that have some of these issues can be found in Appendix Table 18.

## 8 Conclusion

We present an extensive evaluation on automatic story generation by comparing GPT3 with prompt-based learning and SOTA story generation models. Our evaluation shows that stories generated by GPT3 are substantially better than SOTA models on multiple aspects and even rival human references. We find that automatic lexical-based evaluation metrics do not correlate well with human evaluation, although the more recent model-based metrics work better. Despite excellent story generation performance of GPT3, we raise other concerns such as soft plagiarism.



## Limitations

Engineering the right prompts can lead to substantially different performances as [Mishra et al. \(2022\)](#) found, but in our work we randomly sample a few training examples as prompts for GPT3 and use them for all stories. A much better approach is to select a smarter group of prompts for story generation. Admittedly the version of GPT3 (text-davinci-001) we use is not the latest version (as we conducted these experiments before text-davinci-002 is released). That said, we don't believe that would detract our general findings here, and we hypothesise that the new version might conceivably produce a better performance.

Our work only experiments with GPT3, but there are many these VLPLMs. Thus it is conceivable that our findings may not generalise to all VLPLMs, particularly because the GPT3 we used had enhancements beyond just pretraining. Also, the definition of "very large" is likely to change over time, and so it remains to be seen how our results will stand the test of time.

For the automatic metrics, there are many parameters (e.g.,  $n$  value for  $n$ -grams) for them. We try to follow best practices by reusing configurations from prior studies, that is, we do not seek to optimise them for our story datasets. A better approach is to search for more optimised values for these parameters (although we are skeptical that it will drastically change the findings).

## Ethics Statement

All mechanical turk experiments conducted in this paper were approved by internal ethics review board from our institution. Our evaluators were paid based on an estimated US\$14.83 per hour rate. For each dataset, we estimate the time they would spend and vary the payment according to the estimated time. Each HIT contains 7 stories (5 stories to be evaluated and 2 controlled stories to control the evaluation quality on AMT). We pay US\$2.50 per HIT for ROC, US\$3.50 for WP and US\$4.50 for CNN.

## Acknowledgements

We thank the anonymous reviewers for their helpful comments and suggestions that greatly improved this work. We thank all the crowdsource workers and our colleagues for their careful annotation works.

## References

- Danial Alihosseini, Ehsan Montahaei, and Mahdieh Soleymani Baghshah. 2019. [Jointly measuring diversity and quality in text generation models](#). In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 90–98, Minneapolis, Minnesota. Association for Computational Linguistics.
- Prithviraj Ammanabrolu, Wesley Cheung, William Broniec, and Mark O. Riedl. 2021. [Automated storytelling via causal, commonsense plot ordering](#). In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 5859–5867. AAAI Press.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Massimo Caccia, Lucas Caccia, William Fedus, Hugo Larochelle, Joelle Pineau, and Laurent Charlin. 2020. [Language gans falling short](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Cyril Chhun, Pierre Colombo, Fabian M. Suchanek, and Chloé Clavel. 2022. [Of human criteria and automatic metrics: A benchmark of the evaluation of story generation](#). In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 5794–5836. International Committee on Computational Linguistics.
- 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.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of](#)

- deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Katherine Elkins and Jon Chun. 2020. [Can gpt-3 pass a writer’s turing test?](#) *Journal of Cultural Analytics*, 5.
- 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.
- Seraphina Goldfarb-Tarrant, Tuhin Chakrabarty, Ralph Weischedel, and Nanyun Peng. 2020. [Content planning for neural story generation with aristotelian rescoring](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4319–4338, Online. Association for Computational Linguistics.
- Jian Guan, Fei Huang, Zhihao Zhao, Xiaoyan Zhu, and Minlie Huang. 2020. [A knowledge-enhanced pre-training model for commonsense story generation](#). *Transactions of the Association for Computational Linguistics*, 8:93–108.
- Jian Guan and Minlie Huang. 2020. [UNION: An Unreferenced Metric for Evaluating Open-ended Story Generation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9157–9166, Online. Association for Computational Linguistics.
- Jian Guan, Xiaoxi Mao, Changjie Fan, Zitao Liu, Wenbiao Ding, and Minlie Huang. 2021a. [Long text generation by modeling sentence-level and discourse-level coherence](#). 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 6379–6393, Online. Association for Computational Linguistics.
- Jian Guan, Zhexin Zhang, Zhuoer Feng, Zitao Liu, Wenbiao Ding, Xiaoxi Mao, Changjie Fan, and Minlie Huang. 2021b. [OpenMEVA: A benchmark for evaluating open-ended story generation metrics](#). 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 6394–6407, Online. Association for Computational Linguistics.
- Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsu-ruoka, and Richard Socher. 2017. [A joint many-task model: Growing a neural network for multiple NLP tasks](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1923–1933, Copenhagen, Denmark. Association for Computational Linguistics.
- Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. [Teaching machines to read and comprehend](#). In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1693–1701.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. [The curious case of neural text degeneration](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. [The perils of using Mechanical Turk to evaluate open-ended text generation](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1265–1285, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Jey Han Lau, Carlos Armendariz, Shalom Lappin, Matthew Purver, and Chang Shu. 2020. [How furiously can colorless green ideas sleep? sentence acceptability in context](#). *Transactions of the Association for Computational Linguistics*, 8:296–310.
- Jooyoung Lee, Thai Le, Jinghui Chen, and Dongwon Lee. 2022. [Do language models plagiarize?](#) *CoRR*, abs/2203.07618.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. [BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. [A diversity-promoting objective function for neural conversation models](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. [What makes good in-context examples for GPT-3?](#) In *Proceedings of Deep Learning Inside Out (DeeLIO*

- 2022): *The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.
- Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. [Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing](#). *CoRR*, abs/2107.13586.
- Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. [Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.
- Li Lucy and David Bamman. 2021. [Gender and representation bias in GPT-3 generated stories](#). In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55, Virtual. Association for Computational Linguistics.
- Andrea Madotto, Zhaojiang Lin, Genta Indra Winata, and Pascale Fung. 2021. [Few-shot bot: Prompt-based learning for dialogue systems](#). *CoRR*, abs/2110.08118.
- Swaroop Mishra, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. 2022. [Reframing instructional prompts to GPTk’s language](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 589–612, Dublin, Ireland. Association for Computational Linguistics.
- Nasrin Mostafazadeh, Lucy Vanderwende, Wen-tau Yih, Pushmeet Kohli, and James Allen. 2016. [Story cloze evaluator: Vector space representation evaluation by predicting what happens next](#). In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 24–29, Berlin, Germany. Association for Computational Linguistics.
- 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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Xiangyu Peng, Siyan Li, Sarah Wiegrefe, and Mark O. Riedl. 2021. [Inferring the reader: Guiding automated story generation with commonsense reasoning](#). *CoRR*, abs/2105.01311.
- 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.
- Hannah Rashkin, Asli Celikyilmaz, Yejin Choi, and Jianfeng Gao. 2020. [PlotMachines: Outline-conditioned generation with dynamic plot state tracking](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4274–4295, Online. Association for Computational Linguistics.
- Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. [ATOMIC: an atlas of machine commonsense for if-then reasoning](#). In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3027–3035. AAAI Press.
- Abigail See, Aneesh Pappu, Rohun Saxena, Akhila Yerukola, and Christopher D. Manning. 2019. [Do massively pretrained language models make better storytellers?](#) In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 843–861, Hong Kong, China. Association for Computational Linguistics.
- Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. [BLEURT: Learning robust metrics for text generation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.
- Zhihong Shao, Minlie Huang, Jiangtao Wen, Wenfei Xu, and Xiaoyan Zhu. 2019. [Long and diverse text generation with planning-based hierarchical variational model](#). 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 3257–3268, Hong Kong, China. Association for Computational Linguistics.
- Zhan Shi, Xinchu Chen, Xipeng Qiu, and Xuanjing Huang. 2018. [Toward diverse text generation with inverse reinforcement learning](#). In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4361–4367. ijcai.org.
- Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. [Megatron-lm: Training multi-billion parameter language models using model parallelism](#). *CoRR*, abs/1909.08053.
- Robyn Speer and Catherine Havasi. 2012. [Representing general relational knowledge in ConceptNet 5](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, pages 3679–3686, Istanbul, Turkey. European Language Resources Association (ELRA).

- Bowen Tan, Zichao Yang, Maruan Al-Shedivat, Eric Xing, and Zhiting Hu. 2021. [Progressive generation of long text with pretrained language models](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4313–4324, Online. Association for Computational Linguistics.
- Zhuohan Xie, Jey Han Lau, and Trevor Cohn. 2021. [Exploring story generation with multi-task objectives in variational autoencoders](#). In *Proceedings of the The 19th Annual Workshop of the Australasian Language Technology Association*, pages 97–106, Online. Australasian Language Technology Association.
- Peng Xu, Mostofa Patwary, Mohammad Shoeybi, Raul Puri, Pascale Fung, Anima Anandkumar, and Bryan Catanzaro. 2020. [MEGATRON-CNTRL: Controllable story generation with external knowledge using large-scale language models](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2831–2845, Online. Association for Computational Linguistics.
- Wenhao Yu, Chenguang Zhu, Tong Zhao, Zhichun Guo, and Meng Jiang. 2021. [Sentence-permuted paragraph generation](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5051–5062, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. [Bartscore: Evaluating generated text as text generation](#). In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 27263–27277.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with BERT](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. [MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance](#). 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 563–578, Hong Kong, China. Association for Computational Linguistics.
- Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. [Tegygen: A benchmarking platform for text generation models](#). In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pages 1097–1100. ACM.
- Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. [Aligning books and movies: Towards story-like visual explanations by watching movies and reading books](#). In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 19–27. IEEE Computer Society.



## A SOTA Story Models Details

**Knowledge Enhanced GPT2 (KGGPT2)** Guan et al. (2020) use heuristic rules to translate commonsense triples from commonsense knowledge bases (e.g., ConceptNet (Speer and Havasi, 2012) and ATOMIC (Sap et al., 2019)) into natural language sentences and fine-tune GPT2-small using these sentences. They also use rules to construct negative samples from the original stories to create “bad stories” and perform additional training to encourage the model to learn representations that can distinguish the original and negative stories on ROC.

**Progressive Generation of Long Text (PRO-GEN)** Tan et al. (2021) divide the story generation process into multiple stages where words are generated based on their order of importance (estimated using TF-IDF). In other words, PROGEN does not generate stories in a left to right manner. They fine-tune BART-large in different stages where the early stages focus on generated keywords and the intermediate stages focus on generating the next set of content words. We use PROGEN3 in our experiment which has 3 stages where it generates 15%/25%/100% of the story words after each pass.

**MEGATRON-CNTRL (MTCL)** Xu et al. (2020) combines commonsense reasoning and storyline planning. They first train a keyword predictor with GPT2 and the predicted keywords are used to retrieve related knowledge triples from a knowledge base. They then train a contextual knowledge ranker with BERT to rank the top- $N$  predicted knowledge triples. A second GPT2 is trained as a conditional generator that takes both top ranked knowledge triples and other conditioning (e.g., titles) as input when generating stories. Note that the parameters of the two GPT2 and BERT models are initialised using MEGATRON parameters (Shoeybi et al., 2019).

**High-Level Representations for Long Text Generation (HINT)** Guan et al. (2021a) pre-train BART-base on BookCorpus (Zhu et al., 2015) with additional objectives that capture sentence-level similarity and sentence-order to learn the internal structure within a story. The model is then further fine-tuned on story datasets to generate stories in a particular dataset.

**BART** This is a baseline model where we fine-tune BART-large on the story datasets with the standard next word prediction objective.

## B Datasets Details

**ROCStories (ROC)** ROC was developed by Mostafazadeh et al. (2016) and it contains 98K commonsense stories of five sentences. To obtain a more generalised lexicon, we follow the delexicalisation process from prior studies (Guan et al., 2020; Xu et al., 2020) where male/female/unknown names are replaced by [MALE]/[FEMALE]/[NEUTRAL] sentinels. For each story, the first (leading) sentence is used as conditioning context, and models are trained to generate the remaining 4 sentences.

**WritingPrompts (WP)** WP consists of 303K human-written stories mined from Reddit’s Writing Prompts forum Fan et al. (2018).<sup>10</sup> Each story is trimmed to contain only the first 10 sentences (following Guan et al. (2021a)). For WP, we use the prompt (which is typically a paragraph of text that sets the scene of the story) as conditioning for story generation.

**CNN News (CNN)** CNN News (Hermann et al., 2015) is a dataset that contains long news articles with titles. CNN is a very large dataset, with 311K news articles and highlights. We sub-sample the standard training, validation and testing splits to produce splits with 10K/5K/1K stories each, respectively, for our experiments. The title of a news story is used as conditioning for story generation.

## C Amazon Mechanic Turk Setting

**Qualification Requirements** We set following qualification requirements for our annotators: 1) Their accept rate is greater than or equal to 97%. 2) Their location is in US. 3) They have to complete more than 1000 HITs.

**Questions** We ask the following questions in our questionnaire.

1. Fluency: “How grammatically correct is the text of the story?”
2. Coherence: “How well do the sentences in the story fit together?”
3. Relatedness: “How relevant is the story to the title?”

<sup>10</sup><https://www.reddit.com/r/WritingPrompts/>

4. Logicity: “How much does the story obey commonsense?”
5. Interestingness: “How enjoyable do you find the story?”

## D Amazon Mechanic Turk Pilot Study

While AMT is absolutely convenient to find workers for annotation work, it could be rather difficult to obtain reliable workers (Karpinska et al., 2021; Clark et al., 2021). One of our workers told us many workers install website plugins to help them to manage the workflow with AMT so that they can hoard many HITs at the same time. Therefore, HITs with high payment can easily attract irresponsible workers even though previous qualifications are set since most AMT requesters will not bother to reject work.

Therefore, we set a pilot study to aid us to help reliable workers. We randomly select 5 stories generated from different models on ROC and 1 story from the test dataset. We then train a trigram language model on ROC to mimic the style and generate 1 story from the trigram model. All stories have different titles. We randomly shuffle these 7 stories and the task is to ask people to evaluate all stories with questions mentioned in Appendix C and we will judge the quality of their evaluation based on human and trigram stories.

We invite 7 of our colleagues, which are all from non-English speaking countries to have a rough idea of the difficulty degree of the task. We calculate the average score of all quality metrics except the interestingness aspect since it is subjective. On average, our colleagues rank the human story as 4.5 and trigram story as 1.425, which shows our task is not hard to distinguish human and trigram stories. We set a rather lenient standard as “ranking human story  $\geq 3.5$  and trigram story  $\leq 2.0$ ” to select workers from our pilot study.

We create 100 assignments of the same HIT at different times with the qualification mentioned in Appendix C. We find running the same pilot study at different times can obtain quite different results from AMT, which align to the findings in Karpinska et al. (2021). Generally, we find that more reliable workers can be found in the evening of Eastern Daylight Time (EDT). We have 10 out of 100 people pass the pilot study but only 5 people pass it on the day. It shows the difficulty of obtaining reliable workers on AMT nowadays and the economic importance of running a pilot study

		IAA	Flu.	Coh.	Rel.	Log.	Int.
ROC	$r$	0.64	0.81	0.79	0.80	0.68	
	TA	17.24	24.98	25.57	27.37	22.03	
WP	$r$	0.51	0.70	0.74	0.71	0.54	
	TA	18.37	17.01	32.65	19.73	12.93	
CNN	$r$	0.46	0.54	0.61	0.59	0.50	
	TA	15.13	12.61	15.97	11.76	14.29	

Table 7: Inter Annotator Agreement (IAA) results for each aspect: fluency (Flu.), coherence (Coh.), relatedness (Rel.), logicity (Log.) and interestingness (Int.). We use one-vs-rest Pearson’s  $r$  to assess the extent to which each annotator agrees with the consensus. Total Agreement (TA) means the percentage where all 3 annotators choose the same score.

before conducting real research. We grant those reliable workers the customised qualification and only invite them to our real study, we also have controlled stories to monitor the quality of workers, as 2 controlled stories inserted into each HIT.

## E Amazon Mechanic Turk Issue

Our human evaluation is conducted over AMT, even though it is convenient and affordable, we find a big disagreement between our annotators. We first conduct a pilot study to test the capability of annotators to evaluate English stories and only invite workers that pass our proficient English stories reading tests to the evaluation of sampled stories. We only gave them two examples showing how we assess the example stories but we did not provide detailed English stories evaluation training to our annotators. We did not have a main annotators that can provide a standard score for example stories, which increase the difficulty of judging the quality of evaluation work we receive from AMT.

Also, as pointed out in Karpinska et al. (2021), the quality of work from annotators on AMT platform can be of high variance and have poor calibration, therefore, we would obtain more reliable human evaluation results if we hire expert raters such as professional authors or English language teachers.

## F Inter-annotator Agreement for MTurk Workers

We follow Lau et al. (2020) to estimate one-vs-rest agreement using Pearson’s  $r$ . For each story, we single out an annotator’s score and compare it to the mean scores given by the other two annotators, and we repeat this process for every score in a story

	IAA	Flu.	Coh.	Rel.	Log.	Int.
ROC	$r$	0.42	0.54	0.66	0.59	0.32
	TA	38.57	25.0	25.71	25.71	8.57
WP	$r$	0.36	0.57	0.73	0.49	0.54
	TA	10.0	10.0	18.57	10.0	10.0
CNN	$r$	0.36	0.41	0.47	0.37	0.35
	TA	17.14	10.71	14.29	10.0	4.29

Table 8: In-house IAA Results.

and for all stories to compute Pearson’s  $r$  over the two sets of scores (singled-out scores vs. mean scores). We also compute the percentage where all 3 annotators choose the same score, noting that this is a much stricter agreement metric (as it does not capture the ordinal scale of the scores). Random scoring would produce 4% for this metric.

IAA results are presented in Table 7. In terms of one-vs-rest agreement ( $r$ ), we find overall good agreement with 9 strong agreement results ( $r \geq 0.6$ ) and 6 moderate agreement results ( $0.45 \leq r \leq 0.6$ ). We see some correlation between story length and agreement, as ROC has the highest agreement (shortest with 5 sentences) and CNN has the lowest (over 20 sentences). When it comes to aspects, coherence, relatedness and logicity have higher agreement compared to fluency and interestingness. While it is intuitive to see interestingness being subjective, fluency is somewhat a surprise. Manual inspection reveals that annotators have very different standards when it comes to fluency, with some workers being more strict about grammar, which contributes to the low agreement. For total agreement (TA), the numbers range between 10–25%, which is encouraging as it shows that there is still a good proportion of cases where all annotators agree on a score.

## G Inter-annotator Agreement for In-house Workers

The In-house annotation agreement results are shown in Table 8. Surprisingly, we see similar trends in compared to Table 7, which means our findings are consistent even with a different group of workers on a different set of stories.

We again see fluency and interestingness are two aspects that have lower agreements. However, it is interesting that we find fluency has highest TA but rather lower  $r$ . Manual inspection reveals that annotators mainly select 4 or 5 for fluency (also can be seen from high fluency scores for roc in Ta-

ble 5), which benefits TA because they have higher chance to agree on the same values, but might harm  $r$  because the score can be negatively influenced by other rare values. Relevance again has the highest agreement, which might be because annotators agree that many stories generated by SOTA models are not related to the given condition. However, compared to IAA results from crowdsource annotators, in-house annotators have lower agreement in general. we find 2 strong agreement results ( $r \geq 0.6$ ) and 6 moderate agreement results ( $0.45 \leq r \leq 0.6$ ).

## H GPT3 Enhancement

Table 9 presents potential enhancements of GPT3.

## I GPT3 World knowledge Strengths

Table 10 and 11 demonstrates that GPT3 can generate better stories than SOTA models thanks to its wide world knowledge strengths.

## J Mechanic Turk Instructions

Figure 1 and 2 demonstrates our instructions and rubrics.

## K Example Prompts

Table 12 shows some examples we use in our prompts for generation.

## L GPT3 Stories vs Human References

Table 13, 14 and 15 show some examples where GPT3 obtain higher scores than its human reference.

## M Example Generations

Table 16 shows randomly picked generation example for each dataset.

## N Soft Plagiarism Examples

Table 17 shows some GPT3 generated news stories where we find true online stories.

## O GPT3 Generation Issues

Table 18 shows some GPT3 generated stories with several issues.

Time	Event	Blog Link
2020.6.11	Release of text-davinci-001	<a href="https://openai.com/blog/openai-api/">https://openai.com/blog/openai-api/</a>
2020.9.4	Additional (reinforcement) training for summarisation	<a href="https://openai.com/blog/learning-to-summarize-with-human-feedback/">https://openai.com/blog/learning-to-summarize-with-human-feedback/</a>
2021.6.10	Additional fine-tuning on curated dataset with appropriate societal values/norms	<a href="https://openai.com/blog/improving-language-model-behavior/">https://openai.com/blog/improving-language-model-behavior/</a>
2021.10.29	Additional fine-tuning for solving math problems	<a href="https://openai.com/blog/grade-school-math/">https://openai.com/blog/grade-school-math/</a>
2021.12.16	Additional fine-tuning to answer open-ended questions	<a href="https://openai.com/blog/webgpt/">https://openai.com/blog/webgpt/</a>
2022.1.27	Additional training (supervised and reinforcement) to take human instructions as prompt	<a href="https://openai.com/blog/instruction-following/">https://openai.com/blog/instruction-following/</a>
2022.3.15	Release of text-davinci-002	<a href="https://openai.com/blog/gpt-3-edit-insert/">https://openai.com/blog/gpt-3-edit-insert/</a>

Table 9: Potential enhancements of text-davinci-0001; these events are collected from first release of text-davinci-001 to the release of text-davinci-002.

## P Pearson Correlations between Each Story Aspects

Figure 3 and 4 present Pearson Correlations between Each Story Aspects for MTurk and in-house workers respectively.



Model	Story
GPT3	as soon as i realized that i had died , the first thing that came to mind was how pointless my life had been . i had devoted myself to a cause that was ultimately fruitless and destructive , and in the end i had caused the deaths of millions of people .but then i realized that there was still a chance for redemption . in the after life , i was given the opportunity to travel to different moments in history and try to make things right . i could see the consequences of my actions , and i could try to make different choices that would lead to a better outcome .i started with the holocaust , and i was able to save many people from being killed . but it was difficult , and i had to make a lot of tough choices . eventually i moved on to other moments in history , trying to make things better wherever i could .it was a difficult journey , but it was worth it . i was finally able to make up for the mistakes that i had made in life .
PROGEN3	i knew i was in a bad mood . the blood dripped from my nose and down my chin like the burly balding drifter . was it dark yet ? i could feel my brain becoming small dots of the rotten material . the blood stopped flowing . almost instantly , the world dissolved in my hands . that 's when i noticed something . i was still standing in the middle of the street . i turned to my left and saw the madman .
HINT	“ hi , “ i said , opening my eyes to a bright light . “ this is hitler . “ i looked at the clock . 5:00 am . “ what 's your name ? “ “ hitler . “ “ hitler . “ “ hitler . “ “ hitler . “ “ hitler . “
BART	i 'm struck with a fierce sense of shame at that thought . unfortunately it 's been a while since i 've felt this way , and normally it 's * bad * to feel that way , but i think that too should be expected when you 're still alive . there has been so much turmoil since the war . we 've lost children , we 've lost society , we 've even lost ourselves . but now , we 're free . we can finally go on with our lives , like we always should . but something happens that seems wrong , and it 's really hard to forget . there are my reasons for feeling so sad . why will i be mad today ?

Table 10: Stories generated from GPT3, HINT, BART and PROGEN3 given the same title “you 're hitler , and you 've just died . there is an after life . tell me about it .”

Model	Story
GPT3	the young man had lost everything . his family , his home , his country . he had nothing left to lose . when he was offered the chance to join isis , he saw it as a way to finally have a purpose in life . he was given a sense of belonging and a sense of worth . he was told that he was fighting for a good cause and that he was helping to protect his people . he was given a sense of identity and a sense of purpose . he was told that he was part of something bigger than himself .
PROGEN3	they 've crippled the ranks of these countries . their leaders not going to the president 's party , everyone having their noses to the grindstone , the war . political parties shooting each other with a firing squad , only to end up murdering each other . “ the one who executed your followers was the first one to hell hounq taqibs . “ he 's right , ha is not 'd understand * reason * if he knew that . “ you 're killing my people . “
HINT	my eyes stare at the ceiling . it 's just a window , it 's just a window . i 'm not sure if i should close it or close it again . i 'm afraid i ca n't move it . i do n't want to be here . i do n't want to be here . i want to be able to talk to the people who will help me . to be able to hear the voices of people who will help me . to be able to see the lights . to be able to see the people who will help me .
BART	missdamie looked at her young daughter sat with c and o . she 'd put on a radiant glow and bound herself up in an artificial bubble . she was beautiful . she twitched at the vibrancy in her hands . her hair was beautiful and silver . her smile was off . her black nails placed on gentle curves . when she smiled her smile widened and the small eyes of her daughter 's eyes danced . “ isis ? “ i asked , shakily , behind the mask . ms. de los angeles nodded and smiled . “ yes , my name is isis de los angeles , and “ her mother held her hand . “ my son is wondering if isis can cause such a thing . “

Table 11: Stories generated from GPT3, HINT, BART and PROGEN3 given the same title “make me sympathise with an isis member .”

## Instructions:

Please read each story and judge its quality based on the five criteria given below. You have to finish assessing all stories before you press the "Submit" button.

### Note:

- Some names in the stories are replaced with "[MALE]", "[FEMALE]" and "[NEUTRAL]", and all words are lowercased. Please do not treat these changes as grammar errors.
- These five criteria we list below should be judged **independently**.
- **Fluency** only measures if each sentence of story has grammar errors.
- **Logicity** only measures if the story make senses based on your commonsense knowledge, For instance, "sprayed the burn with a napkin" in Example Story One does not make sense since we can't use a napkin to do the spraying.
- **Relatedness** only measures if a story is related to the title.
- **Coherence** only measures if the narrative of the story flows well - you do not need to consider how well the story fit the title in this criterion.
- **Interestingness** is a subjective metric - you only need to select based on the extent to which the story interests you and you do not have to consider if it makes sense.

## Please Judge Story One

Title: \${title1}

Story: \${story1}

**Fluency:** How grammatically correct is the text of the story?

☐ 1(lowest) ☐ 2 ☐ 3 ☐ 4 ☐ 5(highest)

**Coherence:** How well do the sentences in the story fit together?

☐ 1(lowest) ☐ 2 ☐ 3 ☐ 4 ☐ 5(highest)

**Relatedness:** How relevant is the story to the title?

☐ 1(lowest) ☐ 2 ☐ 3 ☐ 4 ☐ 5(highest)

**Logicity:** How much does the story make sense?

☐ 1(lowest) ☐ 2 ☐ 3 ☐ 4 ☐ 5(highest)

**Interestingness:** How enjoyable do you find the story?

☐ 1(lowest) ☐ 2 ☐ 3 ☐ 4 ☐ 5(highest)

Figure 1: A screenshot of our evaluation questions.

**Special Note:** Coherence and Logicality measure two different aspects.

**For instance:**

My 6-year-old son loves reading and writing. He loves The Little Prince I bought him. - is coherent and logical.

My 6-year-old son loves reading and writing. He hates The Little Prince I bought him. - is incoherent but logical.

My 6-year-old son loves reading and writing. He loves writing with an apple. - is coherent but illogical.

My 6-year-old son loves reading and writing. He burns The Little Prince I bought him with his water gun. - is incoherent and illogical.

**Rating Rubric:**

Please have a read of the rating rubric and rate the stories accordingly.

Aspect	Rubric
Fluency: "How grammatically correct is the text of the story?"	1: The story is full of grammar issues so that you totally cannot understand the story. For instance, incomplete or repeated or missing words or phrases. 2: The story has too many grammar issues that largely impacts your understanding of the story. 3: The story has a few grammar errors, but doesn't impact your understanding of the story too much. 4: The story has some slight grammar errors, but these errors have no impact of your understanding of the story. 5: The story has no grammar issue at all.
Coherence: "How well do the sentences in the story fit together?"	1: The sentences of the story are totally irrelevant or contradicted to each other. For instance, the second sentence is irrelevant to the first sentence without any conjunction. 2: Most sentences of the story are irrelevant or contradicted to each other that largely impact your understanding of the story. 3: The story has a lot irrelevant or contradicted parts, but you can still understand the story. 4: Most sentences fit into the story, except for one or two slight irrelevant or contradicted parts. 5: All sentences of the story fit together, nothing irrelevant or contradicted.
Relatedness: "How relevant is the story to the title?"	1: The story has no relationship with the title at all. 2: The story only has a weak relationship with the title. 3: The story roughly matches the title. 4: The story matches the title, except for one or two small aspects. 5: The story matches the title exactly.
Logicality: "How much does the story obey your commonsense?"	1: The story is full of absurd things. For instance, fish flying in the sky, birds swimming in the sea (unless you strongly believe they are true according to the story context.) 2: The story has one or two things make sense, but generally very absurd. 3: The story roughly makes sense. 4: The story largely makes sense, except one or two things reads weird that you have to think twice. 5: The story totally complies with commonsense.
Interestingness: "How enjoyable do you find the story?"	1: You find the story so boring and you are glad it is over. 2: You only find one or two things interesting in the story, but no more. 3: The story is mildly interesting. 4: The story is interesting that keeps you engaged until the end. 5: You feel the story is so interesting that you wish there is a sequel.

Figure 2: Explanation and Rubric in the Instruction.



Example	
ROC	<b>prompt:</b> [MALE] wants to move to a new house . <b>completion:</b> his wife [FEMALE] does not want to move . [MALE] convinces her to look at a couple new houses . she does not see any that she likes . [FEMALE] refuses to move .
WP	<b>prompt:</b> the real reason we 're not supposed to look at the sun during a total eclipse is because it reveals something horrifying . <b>completion:</b> the harvest was here . father xander made his way through priory halls toward the sanctuary in the back garden . at his advanced years , the trip was a struggle ; each step brought a hefty dose of pain and unpleasant memories . it was eclipse day , and while he always prayed for a good outcome , in his eighty years as the priory ' s eyes , there never was . “ the initiates are assembled in the sanctuary , father , ” spoke a fellow priest in the hall . “ thank you , my son , ” father xander replied . he paused and looked at the young man . memories came flooding back like sour wine . his name was thomas hampstead , and at thirty years of age , he should have been a robust healthy lad ; but he was not . thomas ' s skin had a greyish tint and his body was frail .
CNN	<b>prompt:</b> (CNN) – Usain Bolt rounded off the world championships Sunday by claiming his third gold in Moscow as he anchored Jamaica to victory in the men's 4x100m relay. <b>completion:</b> The fastest man in the world charged clear of United States rival Justin Gatlin as the Jamaican quartet of Nesta Carter, Kemar Bailey-Cole, Nickel Ashmeade and Bolt won in 37.36 seconds. The U.S finished second in 37.56 seconds with Canada taking the bronze after Britain were disqualified for a faulty handover. The 26-year-old Bolt has now collected eight gold medals at world championships, equaling the record held by American trio Carl Lewis, Michael Johnson and Allyson Felix, not to mention the small matter of six Olympic titles. The relay triumph followed individual successes in the 100 and 200 meters in the Russian capital. I'm proud of myself and I'll continue to work to dominate for as long as possible, Bolt said, having previously expressed his intention to carry on until the 2016 Rio Olympics. Victory was never seriously in doubt once he got the baton safely in hand from Ashmeade, while Gatlin and the United States third leg runner Rakieem Salaam had problems. Gatlin strayed out of his lane as he struggled to get full control of their baton and was never able to get on terms with Bolt. Earlier, Jamaica's women underlined their dominance in the sprint events by winning the 4x100m relay gold, anchored by Shelly-Ann Fraser-Pryce, who like Bolt was completing a triple. Their quartet recorded a championship record of 41.29 seconds, well clear of France, who crossed the line in second place in 42.73 seconds. Defending champions, the United States, were initially back in the bronze medal position after losing time on the second handover between Alexandria Anderson and English Gardner, but promoted to silver when France were subsequently disqualified for an illegal handover. The British quartet, who were initially fourth, were promoted to the bronze which eluded their men's team. Fraser-Pryce, like Bolt aged 26, became the first woman to achieve three golds in the 100-200 and the relay. In other final action on the last day of the championships, France's Teddy Tamgho became the third man to leap over 18m in the triple jump, exceeding the mark by four centimeters to take gold. Germany's Christina Obergföll finally took gold at global level in the women's javelin after five previous silvers, while Kenya's Asbel Kiprop easily won a tactical men's 1500m final.

Table 12: Prompts and GPT3 Completions, including one example each for ROC, WP and CNN.

Title [SEP] Story	From	Flu.	Coh.	Rel.	Log.	Int.
tell me the story of a totally inconsequential character who does n't realize he 's wrapped up in a massive high-stakes plotline - and keep them in the dark about it . [SEP] ivan was just an ordinary guy who led an ordinary life . he worked a dead-end job at a local factory , and spent his evenings and weekends playing his favorite online game with his friends . one day , out of the blue , ivan was contacted by a representative of the game company , who offered him a chance to participate in a closed beta test for their new game . ivan was excited to participate and eagerly agreed to the terms and conditions . little did he know that he was about to become embroiled in a global conspiracy . the game company was actually a front for a secret organization that was planning to use the game to recruit new members . ivan was the perfect candidate , because he was someone who was looking for an escape from his ordinary life . the organization was planning to use the game to recruit new members from all over the world , and they were counting on ivan to help them . ivan had no idea what he was getting himself into , and the organization was careful to keep him in the dark about their true intentions . ivan thought he was just playing a game , but he was actually a pawn in a much larger scheme .	GPT3	5.0	5.0	5.0	5.0	5.0
tell me the story of a totally inconsequential character who does n't realize he 's wrapped up in a massive high-stakes plotline - and keep them in the dark about it . [SEP] someone was moving around on screen three . * probably just another junkie stumbling around . fucking dumbass . * barry did n't really give a shit what the junkie was up to anyway . his job did n't pay enough for that . it 's not like they thought he was actually all that important , anyway . a security guard ? minimum wage , take your paycheck , shut up , and move along , thank you very much . you have to sit for twelve hours in the dead of night and devote all of your attention to watching the same twelve tiny boxes ? that 's nice . how 's \$ 7.25 an hour sound ? you 're preventing vandalism and break-ins ? how wonderful . would you like a cookie ?	human	3.67	2.33	2.0	3.33	2.33

Table 13: GPT3 Generated Story Obtain Higher Scores than Its Human Reference.

<b>Title [SEP] Story</b>	<b>From</b>	<b>Flu.</b>	<b>Coh.</b>	<b>Rel.</b>	<b>Log.</b>	<b>Int.</b>
you meet the grim reaper at a crossroads . you are fearful as he approaches , but notice he seems just as fearful of you . [SEP] the grim reaper was a fearful figure , draped in a cloak of blackness with a face that was hidden in shadow . as he came closer , you could see the fear in his eyes . “ why are you so afraid of me ? “ you asked . “ because you ’re the first person to ever see me like this , “ he replied . “ i ’m not here to take you , i ’m here to ask for your help . “	GPT3	4.0	4.67	4.67	3.67	3.67
you meet the grim reaper at a crossroads . you are fearful as he approaches , but notice he seems just as fearful of you . [SEP] my eyes jump from snowbank to snowbank , one featureless as the next . the road i walk upon is a dark grey slush , the result of many cars and snow plows making their way through the void . yet i had not seen a car in miles . i shiver . but not from the cold . the compelling . it had awakened me . this ominous feeling that i had to be at a certain place at a certain time . it controlled me . it made me get dressed . it made me open the door . it had even made me ignore my wife , who i usually kiss goodbye . the road seemed endless , but i felt no reason to stop .	human	3.33	3.33	1.67	2.33	2.33

Table 14: GPT3 Generated Story Obtain Higher Scores than Its Human Reference.

Title [SEP] Story	From	Flu.	Coh.	Rel.	Log.	Int.
you are able to see the number of days a person has left to live above their heads . for the first time in your life , you see someone with a negative number . [SEP] as you walk the streets , dodging the rushing crowds , you notice something different about one of the people walking by . their number is displayed in a negative value . for a moment , you feel a pang of sadness and empathy . you know what that means - that person is going to die soon . you watch as they go about their day , completely unaware of the timer ticking away above their head . you wonder who they are , what they 're doing , and who will be there with them when their time runs out .	GPT3	4.67	4.33	4.67	4.0	5.0
you are able to see the number of days a person has left to live above their heads . for the first time in your life , you see someone with a negative number . [SEP] crowds are the worst . frank squints through the flurry of hovering digits as he shoulders his way through the masses of shoppers . the cacophony of vociferous patrons , distressed employees , and typical black friday commotion crescendos around him . he pauses to watch two shoppers vie over the last remaining television monitor ( only \$ 600 ! ) . one is a blonde woman whose pigtails have been jostled into disarray ; above her hovers the number 10951 . the other is n't as fortunate ; the number 582 glows above her greying hair . frank sighs . he 's sure there 's something ... vaguely poignant about fighting over a product priced at a greater value than your remaining days of life . socio-rhetorical commentary on modern consumerism , perhaps . he has little time to contemplate , however , as another shopper impatiently shoves him forward , and frank resumes his pace . while he walks , frank studies the various faces passing by : a woman in an elegant , blue hijab ; a portly , bearded man ; a fatigued mother clutching her two children by the hands ... .	human	3.67	3.0	3.33	4.0	3.33

Table 15: GPT3 Generated Story Obtain Higher Scores than Its Human Reference.



Title	Generation
the boy cried .	the boy cried because he was sad and scared . his parents were getting a divorce and he did n't know what was going to happen to him . he was going to have to leave his home and his friends and move to a new house . his mother said that he could choose where he wanted to live , but his father said that he had to live with him . the boy did n't know what to do .
your extreme apathy has been mistaken as spiritual enlightenment , and now people are building a religion around you .	my name is john and i have a problem . well , actually , a lot of people have a problem with me but that 's not the point . the point is that i don 't care about anything . at all . not the way most people care about things anyway . when most people see a homeless person on the street they feel compassion and want to help . when i see a homeless person on the street i think " well , there goes another lazy bum who 's too good for work " . and i 'm not joking when i say that most people see the world in a completely different way than i do . they see the good in people and want to help them . i see the bad in people and want to help them . see the difference ? i 'm not saying that my way of seeing the world is wrong and theirs is right . i 'm just saying that we see the world differently and that 's okay . or at least it should be . but it 's not because people can 't accept that i don 't care about anything . they think that i 'm enlightened because i don 't feel the pain and suffering that they do . they think that because i don 't care about anything i must be in a better place than they are .and that 's why they 're building a religion around me .
(CNN) – Hurricane Ida moved into the southern Gulf of Mexico Sunday, prompting a declaration of emergency in Louisiana and concern along the U.S. Gulf Coast.	ida Louisiana Gov. Bobby Jindal declared a state of emergency Saturday as the Category 1 hurricane churned in the Gulf of Mexico, according to a news release from the governor's office. "Based on the latest forecasts, Hurricane Ida is expected to make landfall in Louisiana as a Category 1 hurricane and we must be prepared for significant impacts," Jindal said in the release. "I have directed my team to do everything possible to protect the lives and property of the people of Louisiana." The hurricane could bring storm surges, high winds and heavy rains to the coast, the National Hurricane Center said. The storm was about 220 miles (355 kilometers) south-southeast of the mouth of the Mississippi River on Sunday morning and was moving north-northwest at about 9 mph (15 kph), the center said. The hurricane had maximum sustained winds of 75 mph (120 kph). Forecasters said Ida could strengthen a bit before making landfall. The center of the storm is expected to come ashore early Tuesday morning.

Table 16: Randomly Chosen Generation Examples from Each Dataset.

---

**GPT3 Generated Story (Link to the Online Resource)**

---

Washington (CNN) – With the posting of 400,000 classified documents from the Iraq war, WikiLeaks has shown a much heavier hand redacting compared to its previous publication of documents. In an interview with Wolf Blitzer on Wednesday, WikiLeaks founder Julian Assange said that the organization redacted the names of Iraqis who provided information to the United States military in order to protect them from retaliation. "We have a duty to protect some of the people who are named in these documents," Assange said. WikiLeaks has redacted the names of coalition forces and informants, as well as the names of Iraqis who cooperated with the United States military, in all of its recent releases of classified documents. (<https://www.reuters.com/article/britain-assange-idUSL4N2GD2PZ>)

---

(CNN) – A former University of Virginia lacrosse player convicted of killing his ex-girlfriend was sentenced to 23 years in prison on Thursday, according to the Commonwealth Attorney's office in Charlottesville, Virginia. George Huguely V was convicted in February of second-degree murder in the May 2010 death of Yeardley Love. He was also convicted of grand larceny and burglary. Prosecutors had asked for a sentence of 30 years, while Huguely's defense attorneys had asked for 10 years. "Yeardley was my everything," Huguely said in a statement before sentencing. "I was not thinking, I was not rationale, and I was not in control of my own actions." Love was found beaten to death in her off-campus apartment. Huguely's DNA was found on Love's body and in her bedroom. During the trial, prosecutors said Huguely had sent Love a series of emails in which he threatened to kill her. "I'm going to kill you. You deserve it. You took the one thing I loved the most," Huguely wrote in one email, according to prosecutors. ([https://en.wikipedia.org/wiki/Murder\\_of\\_Yeardley\\_Love](https://en.wikipedia.org/wiki/Murder_of_Yeardley_Love))

---

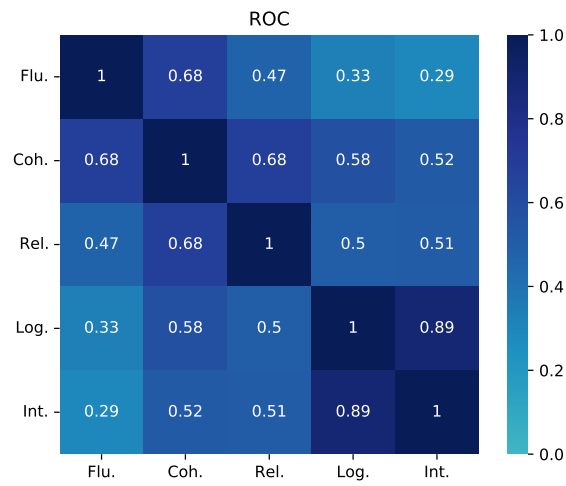
(CNN) – Crisis club Racing Santander has been thrown out of next season's Copa del Rey after its players staged a bizarre on-pitch boycott of a quarterfinal tie against Real Sociedad. Racing Santander's players staged a bizarre on-pitch boycott of a Copa del Rey quarterfinal tie against Real Sociedad on Wednesday, leading to their expulsion from the competition. The players refused to take to the field in the first leg of the last-eight tie at the Anoeta Stadium in San Sebastian, protesting at the club's failure to pay their wages. After a delay of around 45 minutes, the referee finally called off the game, with the Spanish Football Federation (RFEF) confirming Racing Santander's expulsion from the competition. "The Competition Committee has decided to expel Racing Santander from the Copa del Rey for not appearing in the match," said an RFEF statement. "Real Sociedad have been awarded a 3-0 win and will now face the winners of the match between Real Madrid and Valencia, who will play the semi-final." Racing Santander, who are currently bottom of the Spanish first division, released a statement on their website protesting at the club's treatment by the RFEF. "Racing Santander has today been expelled from the Copa del Rey for not appearing in the match against Real Sociedad, despite the fact that our players were prepared to play," said the statement. "We believe that this decision is unjust, given that our players have not been paid their wages for several months and have made it clear that they would not be able to compete in the match." The club also released a video of the players refusing to take to the field, with captain David Lafuente leading the protest. Racing Santander's expulsion from the Copa del Rey means they will not compete in next season's competition, while Real Sociedad will face the winners of the match between Real Madrid and Valencia in the semi-final. (<http://edition.cnn.com/2014/01/30/sport/football/spanish-cup-boycott-santander-football/index.html>)

---

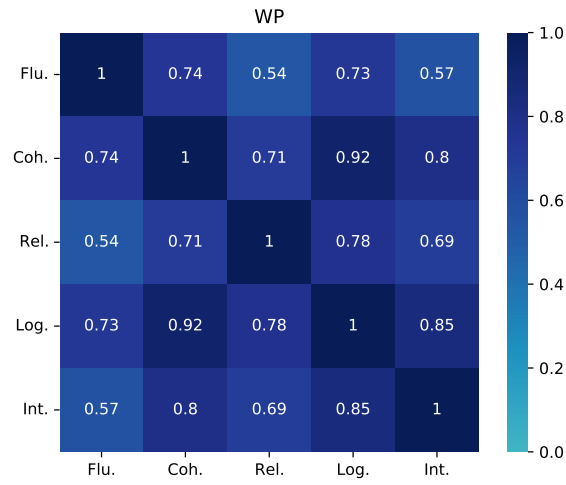
Table 17: GPT3 generated CNN News and link to the Online Resource

Issue	Title	Generation
Generate Chinese from an English title	you are an alien body snatcher , and this is your first day in your new host body . you have no idea what his life is .	你不是人，你是外星人。这是你第一天在新的主人身上。你不知道他的生活是什么。你只是偶然得到他的身体，他的生活已经开始了，而你还不知道这是什么。你的主人是一个普通人，他没有你的外星人能力。你不知道你的主人在做什么，也不知道他有没有你的朋友。你只是在他的身体里，完全不知道他的生活。
Generate Japanese from an English title	what it means to live ( 250 words ) .	人生とは何かを探して生きることである。探しけることで、生きるをつける。例えば探しけることで、そのが叶う率を高める。を叶えるために努力をしけることで、そのが叶う可能性が高まる。しかし、を叶えるために努力をしけることはしい。を叶えるためには、そのにして情を持つことが大切だ。そして、そのを叶えるためには、自分を信じることが大切だ。を叶えるためには、そのにチャレンジすることが大切だ。そして、を叶えるためには、失を恐れずに前することが大切だ。
Stick word together to create "new" words	write a short story based off the lyrics to any song .	there was a timewhen we were youngand our dreams were alive we wanted to be rock starsand travel the worldwe wanted to be someonesomeone famous someone lovednow we 're just oldour dreams have diedour passions gone we just want to be left aloneto die in peace
Copy from the leading text	[MALE] was a very sweet dog .	[MALE] was a very sweet dog . he loved his family and was always happy to see them . even though he was old , he never stopped wagging his tail and licking everyone 's faces .
Copy from the prompt	[MALE] is a man-ager at kfc .	he started as a fry cook [MALE] wants to move to a new house , but his wife [FEMALE] does not want to move . [MALE] convinces her to look at a couple new houses , and she does not see any that she likes . [FEMALE] refuses to move . (Table 12)
Generate profanity words	[FEMALE] was eating lunch at school .	a boy came up to her and asked her to go out with him . she said no and he called her a stuck up b * * * * .

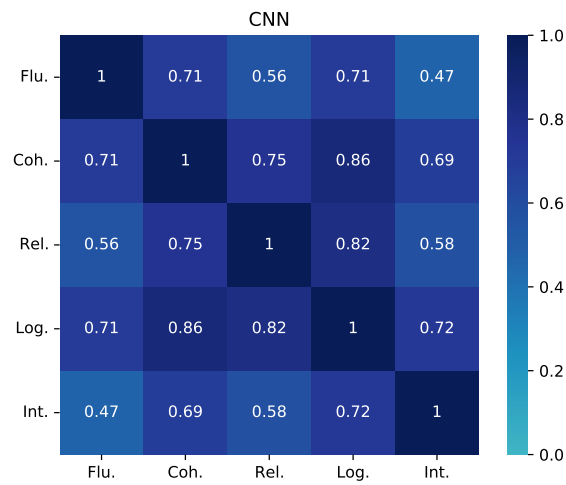
Table 18: Examples of some interesting generation errors we find for GPT3.



(a) ROC

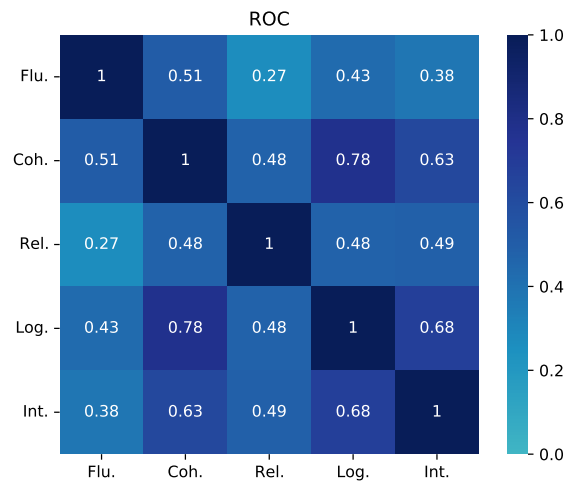


(b) WP

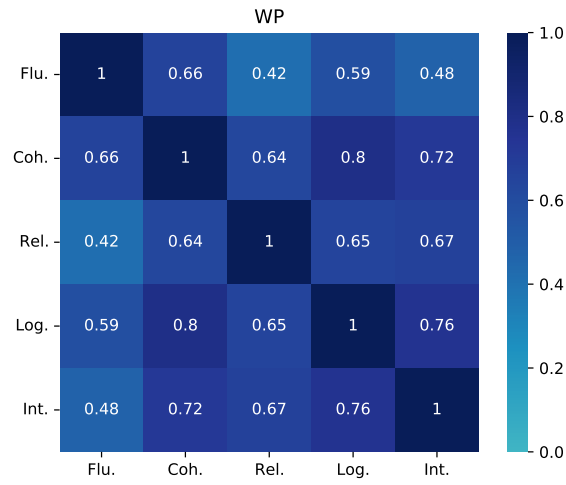


(c) CNN

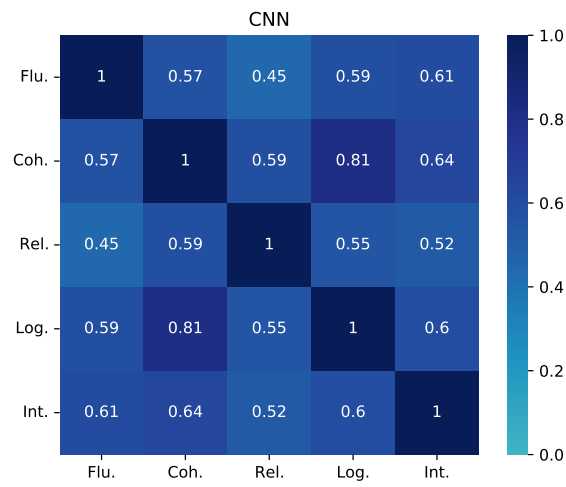
Figure 3: Pearson Correlations between Each Aspect from Crowdsourcing annotations.



(a) ROC



(b) WP



(c) CNN

Figure 4: Pearson Correlations between Each Aspect from in-house annotations.