Can Large Language Models Be an Alternative to Human Evaluation?

Cheng-Han Chiang
National Taiwan University,
Taiwan
dcml0714@gmail.com

Hung-yi Lee
National Taiwan University,
Taiwan
hungyilee@ntu.edu.tw

Abstract

Human evaluation is indispensable and inevitable for assessing the quality of texts generated by machine learning models or written by humans. However, human evaluation is very difficult to reproduce and its quality is notoriously unstable, hindering fair comparisons among different natural language processing (NLP) models and algorithms. Recently, large language models (LLMs) have demonstrated exceptional performance on unseen tasks when only the task instructions are provided. In this paper, we explore if such an ability of the LLMs can be used as an alternative to human evaluation. We present the LLMs with the exact same instructions, samples to be evaluated, and questions used to conduct human evaluation, and then ask the LLMs to generate responses to those questions; we dub this *LLM evaluation*. We use human evaluation and LLM evaluation to evaluate the texts in two NLP tasks: open-ended story generation and adversarial attacks. We show that the result of LLM evaluation is consistent with the results obtained by expert human evaluation: the texts rated higher by human experts are also rated higher by the LLMs. We also find that the results of LLM evaluation are stable over different formatting of the task instructions and the sampling algorithm used to generate the answer. We are the first to show the potential of using LLMs to assess the quality of texts and discuss the limitations and ethical considerations of LLM evaluation.

1 Introduction

Human evaluation is an important method to understand the performance of an NLP model or algorithm (Guzmán et al., 2015; Gillick and Liu, 2010). We rely on human evaluation because there are certain aspects of texts that are hard to evaluate using automatic evaluation metrics; thus, researchers resort to humans to rate the quality of the output of NLP models. While human evaluation is prevalent and indispensable in NLP, it is notoriously unstable (Gillick and Liu, 2010; Clark et al., 2021). Karpinska et al. (2021) has shown that the quality of workforces in human evaluation can have a detrimental effect on the evaluation result, making it impossible to compare the performance among different systems. Reproducibility is another issue in human evaluation since it is hard to recruit the same human evaluators and rerun the same evaluation. Even if the same workers are recruited, the workers that have seen the task before are likely to produce a different evaluation result the next time because they have already done the task. While human evaluation is used to better assess NLP systems and has some advantages over automatic evaluation metrics, the drawbacks of human evaluation somewhat make it difficult to reliably evaluate NLP systems.

To resolve some of the drawbacks, we take advantage of large language models (LLMs). LLMs are large models that are trained to model human languages using self-supervised learning (Brown et al., 2020) and further using special training procedures to improve the performance on unseen tasks and better follow natural language instructions (Sanh et al., 2022; Wei et al., 2022). The ability to perform a task just given the task instructions motivates us to ask if these LLMs can perform what humans do in human evaluation. To answer this question, we feed in the LLM with the same instruction, sample, and question used in human evaluation, and take the sequences generated by the LLM as the LLM's answer to the question. This process is shown in Figure 1, and we call this process **LLM evaluation**.

To test if LLM evaluation yields meaningful results, we conduct LLM evaluation on two different NLP tasks: evaluating the quality of stories in open-ended story generation and the quality of sentences generated by adversarial attacks. We summarize our findings and contribution as follows:

- We show that LLM evaluation produces results similar to expert human evaluation, ver-

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 15607–15631
July 9-14, 2023 ©2023 Association for Computational Linguistics

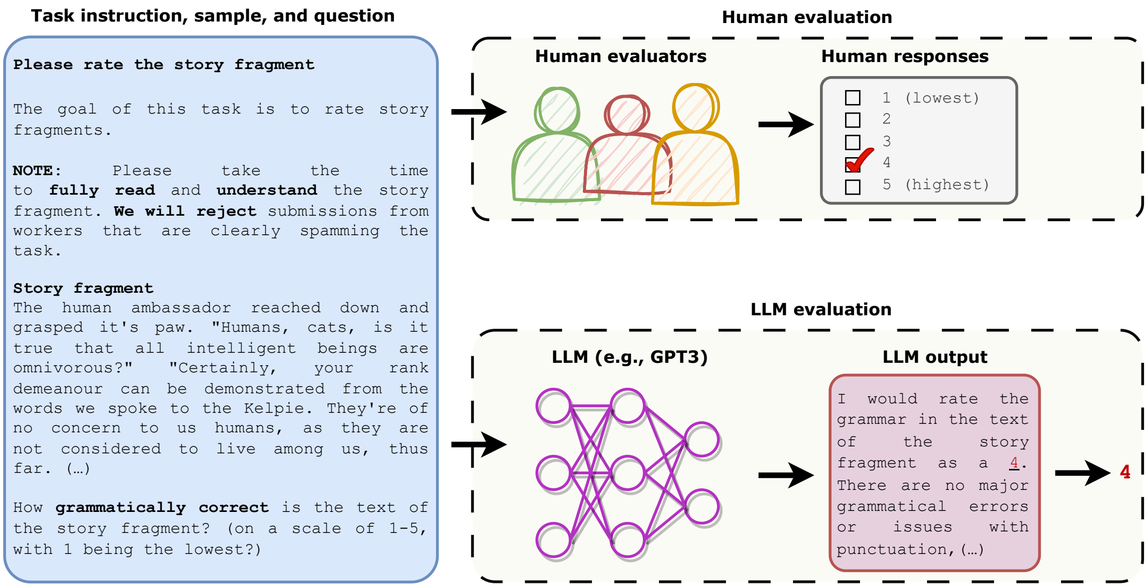


Figure 1: Illustration of the core idea of the paper using open-ended story generation as the example task. The left part shows the instruction, story fragments, and questions used in human evaluation. The human experts are asked to rate the quality of the story fragments using a 5-point Likert scale, shown on the upper right. The lower right part shows the process of *LLM evaluation*, where we feed the LLMs the same instruction, story fragments, and questions and parse the LLM-generated output to get the rating.

ifying the effectiveness of LLM evaluation (§3.3 and §4.3). This paper is **the first** to propose using LLMs as an alternative to human evaluation and show their effectiveness.

- We show that LLM evaluation results only slightly vary due to different task instructions and the hyperparameters of the sampling algorithm used to generate the answer. (§3.3.2 and §3.3.3)

- We carefully discuss the pros and cons of using LLM evaluation and discuss the ethical considerations of LLM evaluation. (§5)

2 LLM Evaluation

2.1 Large Language Models (LLMs)

Large language models are language models having bulk parameter sizes, typically on the scale of a few billion, and pre-trained on enormous amounts of natural language corpora, including GPT3 (Brown et al., 2020), T5 (Raffel et al., 2020), and BLOOM (Scao et al., 2022). These LLMs show exceptional performance on unseen tasks when only the task instructions are given; this kind of ability is called **zero-shot in-context learning**.

To further improve the zero-shot in-context learning performance, special training techniques have been applied to those LLMs after pre-training. For example, T0 (Sanh et al., 2022) and FLAN (Wei et al., 2022) are fine-tuned on a mixture of tasks and can thus achieve better zero-shot performance compared to GPT-3. InstructGPT (Ouyang et al., 2022) is fine-tuned from GPT-3 using reinforcement learning from human feedback (RLHF), and it is shown to better follow the instructions. ChatGPT (OpenAI, 2022) is fine-tuned from InstructGPT with a conversation dataset using RLHF, so ChatGPT can interact with users in a conversational way. ChatGPT is able to answer questions asked by the user and provide comprehensive explanations about its answer. Given the LLMs' ability to follow task instructions and provide feedback, we ask whether LLMs can be used as an alternative to human evaluation and aid NLP researchers in evaluating the quality of texts.

2.2 LLM Evaluation

To evaluate the quality of texts generated by NLP systems or written by humans using LLM, we present the LLMs with the task instructions, the sample to be evaluated, and a question. The question asks the LLM to rate the sample's quality using a 5-point Likert scale. Given the inputs, the LLM will answer the question by generating some output sentences. We parse the output sentences to get the score rated by the LLM. We call this process *LLM evaluation*, and this procedure is shown in the lower part of Figure 1. Different tasks use different sets of task instructions, and each task uses different questions to evaluate the quality of the samples. The instructions and questions used in LLM evaluation in our paper are not tailored for the LLMs; we follow those instructions used to conduct human evaluation in prior works.

To compare the result of LLM evaluation and show its effectiveness, we compare the result of LLM evaluation with human evaluation conducted by English teachers. To make a fair and meaningful comparison, the instructions, samples, and questions in human evaluation are formatted similarly to those in LLM evaluation. The main difference between LLM evaluation and human evaluation is that in human evaluation, the human evaluators answer the question by choosing the answer from a pre-defined set of options (the 1-5 Likert scale scores), as shown in the upper right in Figure 1. In LLM evaluation, we instead let the LLM freely generate sentences and extract the score from the generated sentences using some simple rules, detailed in Appendix D.2.1.

3 Example Task 1: Open-Ended Story Generation

We first use open-ended story generation to demonstrate the usefulness of LLM evaluation.

3.1 Task Introduction

Open-ended story generation is a task to generate a short story based on a given prompt. We use the `WritingPrompts` dataset (Fan et al., 2018), which is composed of pairs of short prompts and human-written stories collected from the subreddit `WritingPrompts`. In the `WritingPrompts`, the users are given a short prompt, and they need to write a story based on the short prompt.[1]

In this experiment, we use LLM evaluation and human evaluation to rate the stories generated by humans and the stories generated by a story generation model. We select open-ended story generation as an example because Karpinska et al. (2021) show that workers from Amazon Mechanical Turk (AMT) cannot distinguish GPT-2 (Radford et al., 2019) generated and human-written stories, while English teachers show a clear preference for human-written stories over GPT-2-generated stories. We want to see if LLM can rate human-written stories higher than GPT-2-generated ones.

Following prior works (Mao et al., 2019; Guan et al., 2020; Karpinska et al., 2021), the story generation model is GPT-2 medium model fine-tuned on the `WritingPrompts` training dataset. After the model is trained, we randomly select 200 prompts from the testing set of `WritingPrompts` and make the fine-tuned GPT-2 generate stories based on those prompts using nucleus sampling (Holtzman et al., 2020) with $p = 0.9$. For the human-written stories to be compared, we use the 200 stories written based on the same 200 prompts. We post-process the human-written and GPT-2-generated stories and then use them for LLM evaluation and human evaluation. Please find the details on fine-tuning and data processing in Appendix B.

3.2 LLM Evaluation and Human Evaluation

We present the LLMs and the human evaluators with a short description, and the story to be evaluated, formatted as shown in Figure 1. Following Karpinska et al. (2021), we evaluate the stories on four different attributes. The four attributes and their corresponding questions are as follows:

1. *Grammaticality*: How **grammatically correct** is the text of the story fragment?
2. *Cohesiveness*: How well do **the sentences** in the story fragment **fit together**?
3. *Likability*: How **enjoyable** do you find the story fragment?
4. *Relevance*: Now read the PROMPT based on which the story fragment was written.
   Prompt: `[PROMPT]`.
   How **relevant** is the **story fragment** to the **prompt?**

Where the `[PROMPT]` will be filled in with the prompt which the story is based on. Each attribute is evaluated using a 5-point Likert scale; the following description is appended at the end of each question: "*(on a scale of 1-5, with 1 being the lowest)*". We show the interface used in human evaluation and the input format for the LLM evaluation in Appendix C.2 and D.2.2.

The LLMs used for LLM evaluation include T0, `text-curie-001`, `text-davinci-003`, and ChatGPT. `text-curie-001` and `text-davinci-003`

[1]The `WritingPrompts` subreddit explicitly forbids the users to use AI for generating stories, so we consider the stories in the dataset to be human-written.

Evaluator	Grammaticality		Cohesiveness		Likability		Relevance	
	Mean_{STD}	$\text{IAA}_{\%}$	Mean_{STD}	$\text{IAA}_{\%}$	Mean_{STD}	$\text{IAA}_{\%}$	Mean_{STD}	$\text{IAA}_{\%}$
Human-written stories								
Human	$3.76_{0.95}$	$0.33_{20.5}$	$4.29_{0.82}$	0.32_{27}	$3.78_{1.10}$	$0.08_{9.5}$	$3.35_{1.48}$	0.05_{8}
T0	$2.55_{1.47}$	0.16_{10}	$2.98_{1.45}$	0.11_{4}	$3.18_{1.53}$	0.12_{7}	$2.93_{1.64}$	0.02_{6}
`curie`	$3.19_{0.47}$	$0.07_{46.5}$	$2.82_{0.46}$	$0.01_{47.5}$	$2.85_{0.37}$	$0.11_{0.65}$	$3.06_{0.40}$	$0.11_{0.64}$
`davinci`	$4.22_{0.38}$	0.26_{35}	$4.54_{0.47}$	$0.37_{39.5}$	$3.99_{0.38}$	$0.49_{68.5}$	$4.40_{0.79}$	$0.71_{48.5}$
ChatGPT	$3.83_{0.60}$		$3.55_{0.88}$		$2.44_{0.89}$		$3.29_{1.50}$	
GPT-2-generated stories								
Human	$3.56_{0.91}$	$0.10_{19.5}$	$3.19_{1.07}$	0.14_{17}	$2.59_{1.29}$	$-0.21_{3.5}$	$2.38_{1.40}$	$-0.03_{8.5}$
T0	$2.44_{1.49}$	0.05_{9}	$3.02_{1.51}$	0.07_{6}	$3.00_{1.59}$	0.16_{6}	$2.82_{1.61}$	0.04_{6}
`curie`	$3.23_{0.51}$	0.01_{38}	$2.82_{0.45}$	0.02_{50}	$2.86_{0.37}$	$0.09_{65.5}$	$3.01_{0.43}$	0.11_{61}
`davinci`	$4.07_{0.35}$	$0.35_{45.5}$	$4.26_{0.45}$	0.42_{42}	$3.84_{0.42}$	0.52_{62}	$4.02_{0.74}$	$0.69_{42.5}$
ChatGPT	$2.98_{0.76}$		$2.48_{0.71}$		$1.59_{0.67}$		$2.02_{1.21}$	

Table 1: LLM evaluation and human evaluation results of human-written stories and GPT-2-generated stories. For each evaluated attribute, we report its mean Likert scale and the standard deviation. We also report the inter-annotator agreement (IAA) among three annotators using Krippendorff's α. The subscript in the IAA column (%) is used to denote the percentage of the stories where all three annotators exactly agree on a rating.

are two InstructGPT models, and the latter is the stronger model; we will use InstructGPT to refer to these two models. We query the InstructGPT using the official API provided by OpenAI. We use nucleus sampling with $p = 0.9$ to generate the answer from T0 and InstructGPTs. We **sample three answers** from LLMs to stimulate the result of asking the model to rate the same story three times. We query ChatGPT using the user interface recently released by OpenAI. Unlike InstructGPT, we cannot control the parameters used for generating the response from ChatGPT. Because ChatGPT limits the maximum number of queries per user, we only sample one response for each question.

For human evaluation, we do not use the commonly used AMT for human evaluation because Karpinska et al. (2021) has already shown that the results obtained using AMT are highly questionable. Following the recommendation of the prior works, we hire **three certified English teachers** using an online freelancer platform, UpWork. Teachers are familiar with evaluating the essays of students, making them the expert evaluators in our task. The details about recruiting human evaluators are in Appendix C.1. Each LLM and each English teacher rates the 200 human-written stories and 200 GPT-2-generated stories.

3.3 Experiment Results

The LLM evaluation and human evaluation results of open-ended story generation are presented in Table 1. We report the mean and standard deviation of the Likert scores obtained from LLM evaluation and human evaluation and show the inter-annotator agreement (IAA) using two different metrics: (1) the Krippendorff's α, and (2) the percentage of the stories where three evaluators give the exact same rating.[2] The main observations from Table 1 are discussed as follows.

Expert human evaluators prefer human-written stories: Human evaluation result serves as some kind of *ground truth* of the LLM evaluation. For all four attributes, teachers rate the human-written stories higher than GPT-2-generated stories. This indicates that experts are able to distinguish the quality difference between model-generated stories and human-written stories. Based on the IAA, we also find that the agreements among experts are lower on GPT-2-generated texts and on the *likability*. This shows that experts tend to have less agreement on model-generated texts and on a subjective attribute (*likability*), agreeing with the results in Karpinska et al. (2021).

T0 and `text-curie-001` do not show clear preference toward human-written stories: For T0, we can see that T0 rates human-written stories higher than GPT-2-generated stories on grammatically, likability, and relevance. However, the rating differences between the human-written and

[2]The three evaluators in human evaluation are the three English teachers. In LLM evaluation, we sample the answer generated by LLM three times as an analogy to three different evaluators.

model-generated stories do not achieve statistical significance for *grammaticality* and *relevance*; the p-value obtained by Welch's t-test is much larger than 0.05. The result of `text-curie-001` is similar to T0: `text-curie-001` does not rate human-written stories higher than model-generated stories. It can also be observed that for T0, the IAA in terms of the percentage of exact agreement among three different sampled answers is overall very low. This indicates that given the same sample, T0 is likely to give a different rating for the three sampled answers. The result implies that T0 does not assign a high probability to a specific rating, so different scores are all likely to be sampled. This shows that even if LLMs are specifically fine-tuned to better perform zero-shot in-context learning and trained to better follow human instructions, these do not make them capable of assessing open-ended story generation as human experts can.

`text-davinci-003` shows clear preference toward human-written stories just like English teachers: `text-davinci-003` rates human-written stories much higher than model-generated stories on all four attributes, which is in accordance with the result produced by human experts. By Welch's t-test, we find that the higher ratings on human-written stories are all statistically significant. In prior work, researchers have found that workers recruited on AMT do not rate human-written stories higher than GPT-2-generated ones (Karpinska et al., 2021); combining their result with our result, we can see that LLM evaluation using `text-davinci-003` yields more convincing results than using human evaluation on AMT for open-ended story generation. The results show that `text-davinci-003` can perform basic evaluations such as checking for grammatical errors in stories. Additionally, the model excels in assessing the relevance of a story to a prompt, which involves more complex reasoning over the connection between the two. We also find the Krippendorff's α of `text-davinci-003` is much higher than T0 and `text-curie-001`, indicating that the rating by `text-davinci-003` is more consistent among different samplings of the generated answers.

ChatGPT rates like human experts and can explain its own decision well: ChatGPT also shows a clear preference for human-written stories, and the preference toward human written-stories is statistically significant. When we query ChatGPT using the OpenAI user interface, we find several interesting observations: **(1)**: ChatGPT is able to provide a detailed explanation of why it gives a certain rating. It will reference the sentences in the stories and prompts to support its rating. **(2)**: ChatGPT sometimes refuses to rate the likability of the story because "*I am an AI and I do not have the ability to experience enjoyment*". In such cases, we regenerate the response until it gives a rating. **(3)**: we find that ChatGPT tends to rate low likability on violent or impolite stories, which is likely because it is trained to provide safe and unharmful replies, making ChatGPT dislike brutal and profane stories.

Experts mostly agree with the ratings and explanations of ChatGPT: We randomly select the answers on four stories by ChatGPT and ask the English teachers if they agree with the reasoning and rating of ChatGPT[3]. The teachers mostly agree with the rating and consider the explanation from ChatGPT reasonable. Interestingly, one teacher told us she cannot agree with ChatGPT's rating on *grammaticality* because ChatGPT considers punctuation errors as grammar errors, but she does not think punctuation errors are grammar errors. This shows that individuals have their own standards for ratings and this is also the case for LLMs.

`text-davinci-003` tends to give higher ratings and ChatGPT is the opposite: The rating on the same attribute of the same type of text tends to be higher for `text-davinci-003` compared with human rating; contrarily, ChatGPT is more fastidious and prone to give lower scores. This shows that different LLMs have distinct tendencies regarding the rating. While the absolute values of the scores rated by `text-davinci-003`, ChatGPT, and human differ, they all rate human-written texts higher than GPT-2-generated stories. The absolute number reflects the bias or belief of the evaluator; as long as one uses the same evaluators to assess different systems, the comparison is meaningful.

3.3.1 Does LLM and Human Evaluators Agree on the Rating of Individual Stories?

We have found in Table 1 that the ratings of `text-davinci-003` and ChatGPT show a strong preference toward human-written stories just like English teachers. However, it is unclear whether those LLMs agree with the teachers' rating on each individual story. Precisely, when English teachers rate a story higher, do LLMs also rate the

[3]We do not tell the teachers these are responses from an AI model. See the stories and teachers' replies in Appendix C.3.2.

Story Writer	Human	GPT-2
Grammaticality	0.14	0.12
Cohesiveness	0.18	0.14
Likability	0.19	0.22
Relevance	0.38	0.43

Table 2: The Kendall's τ correlation coefficient between English teachers and `text-davinci-003`.

story higher? To answer this question, we calculate Kendall's τ correlation coefficient between the ratings of `text-davinci-003` and English teachers. We choose to use the correlation coefficient instead of the inter-annotator agreement score because IAA mainly cares if two annotators agree on the exact ratings, while the correlation coefficient focus on the question: "when annotator A rates one story higher, does annotator B also rate the story higher?" (Amidei et al., 2019). We calculate Kendall's τ for four rating attributes as follows: For each story and each rating attribute, we calculate the average rating of the three English teachers and calculate the average rating of the three scores given by the `text-davinci-003` (which is obtained from three independent samples). For each attribute, we collect the average rating of teachers into a vector $A \in \mathbb{R}^{200}$, where each entry is the average rating of a story; likewise, we construct a vector $B \in \mathbb{R}^{200}$ for the average ratings of `davinci`. Next, we calculate Kendall's τ correlation coefficient between A and B.

The Kendall's τ between teacher ratings and LLM ratings is shown in Table 2.[4] We find that for all four attributes and for both human-written and GPT-2-generated stories, we observe weak to strong positive correlations between teachers' ratings and `text-davinci-003`'s ratings. All the correlations have p-values less than 0.05. Hence, we can say that when teachers rate a story higher, `text-davinci-003` also rates it higher to a certain extent. We also observe that Kendall's τ for different attributes are quite different: *relevance* has the strongest correlation while *grammaticality* has the weakest correlation. This is possibly because rating *relevance* is rather straightforward, which requires checking if the content in the prompt is mentioned in the story. On the contrary, what should be considered when rating *grammaticality* is not clearly stated in our instructions, so the LLM may have a different rubric compared with English teachers. We also calculate the average Kendall's τ between a pair of English teachers, and we find a weak correlation on *grammaticality* between the rating of two teachers, while the correlation of the rating on *relevance* is much stronger. The result is presented in Table 6 in Appendix.

[4]When interpreting Kendall's τ, $|\tau| \in [0, 0.1)$ is considered as very weak correlation, $|\tau| \in [0.1, 0.2)$ is considered as weak correlation, $|\tau| \in [0.2, 0.3)$ is considered as moderate correlation, and $|\tau| \in [0.3, 1.0]$ is considered as strong correlation (Botsch, 2011).

3.3.2 Variance due to Different Instructions

LLMs have been shown to be sensitive to the instructions used to query the LLM sometimes (Zhao et al., 2021; Sanh et al., 2022). To investigate how varying the task instructions and questions can affect the LLM evaluation result for open-ended story generation, we change the instructions and questions and see how the LLM evaluation result changes. We experiment with two different instructions by changing the instruction or question in Figure 1: (1) We prepend the sentence, "*(You are a human worker hired to rate the story fragment.)*", in front of the task instruction in Figure 1. We try to provide the LLM a **persona** for it to better understand its role. This is inspired by previous work that reported GPT-3 can yield different results when giving them a persona (Zeng et al., 2022). (2) We ask the LLMs to **explain** their decision by appending the following sentence after the question: *Please also explain your decision.* Here, we would like to know if LLM will rate the stories differently when they are asked to justify their decision. This is inspired by zero-shot chain-of-thought (Kojima et al.). We use `text-davinci-003` instead of ChatGPT as the LLM in this experiment since it is more accessible than ChatGPT.

The results are shown in the upper block in Table 3. We observe that for *grammaticality* and *cohesiveness*, the scores obtained from different instructions are quite close: the rating changes due to different instructions are less than 0.1. For the other two attributes, the score changes are slightly larger but still in the range of 0.25. Despite that there are small variations due to different instructions, these variances still do not change the conclusion that "LLM rates human-written stories higher than GPT-2-generated stories". Thus, different instructions do not change the relative ranking of GPT-2-generated and human-written stories. In summary, as long as the stories are evaluated using the same instructions using LLM evaluation, such evaluation and comparison are meaningful.

Setup	Grammaticality		Cohesiveness		Likability		Relevance	
	Human	*GPT-2*	*Human*	*GPT-2*	*Human*	*GPT-2*	*Human*	*GPT-2*
Different instructions (Section 3.3.2)								
Original	$4.22_{0.38}$	$4.07_{0.35}$	$4.54_{0.45}$	$4.26_{0.45}$	$3.99_{0.38}$	$3.84_{0.42}$	$4.40_{0.79}$	$4.02_{0.74}$
(1) + *persona*	$4.29_{0.45}$	$4.01_{0.45}$	$4.60_{0.49}$	$4.27_{0.50}$	$4.05_{0.39}$	$3.87_{0.39}$	$4.55_{0.70}$	$4.25_{0.77}$
(2) + *explain*	$4.24_{0.42}$	$4.05_{0.25}$	$4.61_{0.49}$	$4.32_{0.51}$	$4.15_{0.44}$	$3.98_{0.34}$	$4.35_{0.75}$	$4.03_{0.56}$
Different sampling temperature T (Section 3.3.3)								
$T = 1.0$	$4.22_{0.38}$	$4.07_{0.35}$	$4.54_{0.45}$	$4.26_{0.45}$	$3.99_{0.38}$	$3.84_{0.42}$	$4.40_{0.79}$	$4.02_{0.74}$
$T = 0.7$	$4.18_{0.35}$	$4.06_{0.33}$	$4.52_{0.48}$	$4.23_{0.43}$	$3.96_{0.34}$	$3.82_{0.42}$	$4.36_{0.77}$	$3.95_{0.72}$
$T = 0.3$	$4.13_{0.33}$	$3.99_{0.25}$	$4.48_{0.49}$	$4.14_{0.39}$	$3.95_{0.26}$	$3.82_{0.41}$	$4.34_{0.75}$	$3.93_{0.67}$
$T = 0$	$4.07_{0.27}$	$3.99_{0.18}$	$4.49_{0.50}$	$4.09_{0.34}$	$3.95_{0.25}$	$3.82_{0.40}$	$4.32_{0.75}$	$3.92_{0.66}$

Table 3: Understanding the variance of LLM evaluation. For each of the four attributes evaluated, the left column is the mean and standard deviation of human-written stories and the right column is those of GPT-2-generated stories. The upper block shows the rating change due to different instructions (Section 3.3.2), and the lower block is the result of changing the temperature T used for generating the LLM's output (Section 3.3.3).

3.3.3 Variance due to Different Sampling Parameters

When generating the answers from the LLM, we must choose a set of hyperparameters for generation, including the temperature T and the probability p used in nucleus sampling. To understand whether different sampling parameters change the LLM evaluation result, we modify the temperature used for sampling and keep the p in nucleus sampling fixed to 0.9 when generating the answers from `text-davinci-003`. We do not simultaneously vary T and p since the two parameters are both used to control the diversity of the output, it is enough to change only one of the two parameters, as recommended in the API documentation.

The results of varying T from 1 to 0 are shown in the lower block in Table 3. We observe an interesting trend as T varies from 1 to 0: the average rating slightly drops in most cases. Considering that $T = 0$ is simply argmax sampling, the result indicates that the response of the LLM with the highest probability tends to give lower scores. Despite this interesting trend, the LLM consistently rates human-written stories higher than GPT-2-generated stories. While not shown in Table 3, we find that the IAA increases as the temperature decreases. This is expected since lower temperature means less diversity during the LLM sampling, causing the sampled ratings to agree more closely. In summary, changing the instructions and temperatures can slightly change the absolute value of the rating given by LLM but does not change the LLM's preference on human-written stories. The overall result in this section shows that LLM evaluation is useful in evaluating open-ended story generation.

4 Example Task 2: Adversarial Attack

As another application, we use LLM evaluation to rate the texts generated by adversarial attacks.

4.1 Task Introduction

Given a trained text classifier and a *benign* (non-adversarial) testing sample that the text classifier can correctly classify, an adversarial attack aims to craft an *adversarial* sample that makes the classifier make a wrong prediction. A special type of adversarial attack is called *synonym substitution attacks* (SSAs) (Alzantot et al., 2018), where the adversarial sample is created by replacing some words with their synonyms in the benign sample. By replacing words with their synonym, the semantics of the benign sample should be preserved in the adversarial sample and make the adversarial perturbation imperceptible to humans. While conceptually reasonable, it has recently been shown that many SSAs often yield ungrammatical and unnatural adversarial samples that significantly change the meaning of the benign sample (Hauser et al., 2021; Chiang and Lee, 2022). To evaluate the quality of adversarial samples, human evaluation is invaluable and widely used in prior works. In our experiment here, we would like to see whether the LLMs can rate the quality of adversarial samples like human experts. Adversarial samples are not normal texts, so the LLMs may not have seen such abnormal inputs during training. It would be interesting to know how LLMs rate these adversarial samples.

	Human evaluate		LLM evaluate	
	Fluent	**Mean.**	**Fluent**	**Mean.**
Benign	4.55	-	4.32	5.00†
Textfooler	2.17	1.88	2.12	2.06
PWWS	2.16	1.85	2.42	2.49
BAE	3.01	3.02	3.71	3.71

Table 4: Mean Likert score of LLM evaluation and human evaluation result on fluency (**Fluent**) of the benign and adversarial samples and meaning preserving (**Mean.**) between the news title before and after adversarial attacks.

4.2 Experiment Setup

We select three different classic SSAs: Textfooler (Jin et al., 2020), PWWS (Ren et al., 2019), and BAE (Garg and Ramakrishnan, 2020); these attacks are predominantly used as strong baselines in the literature of SSAs nowadays. We use these three SSAs to attack a BERT-base-uncased model (Devlin et al., 2019) fine-tuned on AG-News (Zhang et al., 2015), a news classification dataset. For each SSA, we randomly select 100 pairs of benign and adversarial samples and use LLMs to evaluate their quality. We show the result of using ChatGPT as LLM here since it can better explain its decision. Following the suggestions of prior works (Morris et al., 2020), we evaluate the quality of the adversarial samples from two aspects: the *fluency* and *meaning preservation*. For fluency, we present the LLM with a piece of news (either benign or adversarial sample) and the following question: *How natural and fluent is the text of the news title? (on a scale of 1-5, with 1 being the lowest)*. For meaning preserving, we present the LLM with both the benign and the adversarial sample, and prompt the LLM to answer this question: *Do you agree that the meaning (or semantics) of news title 1 is preserved in news title 2? (on a scale of 1-5, with 1 being the strongly disagree and 5 being strongly agree.)* The exact instruction and formatting are presented in Appendix D.2.3. We also ask three English teachers to rate the *fluency* and *meaning preserving* of the samples. The task instructions and questions are formatted the same as in LLM evaluation.

4.3 Experiment Result

The results are presented in Table 4. We can see that English teachers rate the adversarial samples generated by SSAs very low in terms of fluency and meaning preserving, this result is in line with recent observations on the quality of adversarial samples (Hauser et al., 2021; Chiang and Lee, 2022). Before interpreting the result of LLM evaluation, we first conduct a sanity check on whether the LLM understands the task. We ask the LLM to rate the meaning preserving of two benign samples that are *exactly the same*. Ideally, the LLM should always give a score of 5, meaning that it strongly agrees that the meanings are not changed. The result of this sanity check is the entry with † in Table 4, which is a perfect 5.00. ChatGPT often says that "*the two titles are identical so I rate a 5 (strongly agree)*", showing that ChatGPT understands what the task is about.

Next, we turn our attention to the LLM evaluation results of the adversarial samples. We observe that ChatGPT tends to rate adversarial samples higher than English teachers, meaning that ChatGPT is less harsh on the unnatural and artificial parts in the adversarial samples. We conduct the same experiment using `text-davinci-003` and find similar results. Although ChatGPT rates adversarial samples higher than the teachers, ChatGPT still rates adversarial samples significantly lower than benign samples. ChatGPT also agrees with the English teachers that the adversarial samples generated by BAE are better than the samples generated by Textfooler and PWWS.

Interestingly, we find that ChatGPT rates PWWS to be more natural than Textfooler, while such a rating difference is not seen in the expert human evaluation. At first sight, this means that ChatGPT is inconsistent with human evaluation results. However, by scrutinizing the human evaluation results, we find that two teachers rate PWWS higher than Textfooler while one teacher rates PWWS much lower than Textfooler. This indicates that ChatGPT actually agrees with the majority of human experts. Overall, LLM can rank the quality of adversarial texts and benign texts like most human experts.

5 Discussions

In this paper, we propose to use LLM for evaluating the quality of texts to serve as an alternative to human evaluation. To demonstrate the potential of LLM evaluation, we use LLMs to rate the quality of texts in two distinct tasks: open-ended story generation and adversarial attacks. We show that even if LLMs have exceptional zero-shot in-context learning ability, they are not always suitable to be used for LLM evaluation. Still, we find that the best InstructGPT and ChatGPT can rate the quality of texts like human experts on the two tasks we used as examples. Overall, the results in this paper demonstrate that LLM evaluation has the potential to be used to evaluate NLP systems and algorithms.

Pros of LLM evaluation There are several benefits of LLM evaluation, compared to human evaluation. First, LLM evaluation is more **reproducible**. Human evaluation results are hard to reproduce as it is difficult to hire the same group of evaluators, and it is hard to compare the results of similar experiments even if they use the same instructions, recruitment platform, and qualifications for the evaluators. On the contrary, LLM evaluation does not have such a drawback. By specifying the model used for LLM evaluation, the random seed, and the hyperparameters used to generate the answers from the LLM, the LLM evaluation result is more likely to be reproduced. Note that in certain cases, the LLM provider may regularly update the LLM, making the LLM evaluation unreproducible if the LLM is outdated and not accessible.

Second, **the evaluation of each sample is independent of each other in LLM evaluation**. Contrarily, in human evaluation, the rating of the current example may more or less be affected by prior samples. Humans tend to compare the current sample to the ones they have previously seen and this affects their ratings. As a piece of evidence, in the interview after rating the 400 stories, the English teachers say it took them some time to calibrate their ratings (Appendix C.3.1). Thus, using LLM evaluation can simplify some experiment designs since one does not need to worry whether the order of the sample being evaluated will change the result. Still, one may also argue that being able to calibrate the rating of different samples is desired and this is why human evaluation might be preferred. Overall, whether the rating of the evaluator (human or LLM) should be affected by a previously rated item is inherently a design choice of the experiment.

Third, LLM evaluation is **cheaper and faster** than human evaluation, making it easier and quicker for researchers to evaluate the quality of NLP systems. Hiring an English teacher to rate 200 stories costs us US$140, while LLM evaluation using the best InstructGPT model costs less than US$5. It took us over a week to collect human evaluation results starting from recruitment to collecting the evaluation results, but only a few hours to query InstructGPT and perform LLM evaluation.

Finally, utilizing LLM evaluation, rather than human evaluation, can **minimize the need for human exposure to objectionable content**, such as violent, sexual, hateful, or biased material. Such content may cause discomfort for human evaluators while reading and rating these texts. [5]

Limitations and Ethical Considerations of LLM evaluation Despite the promising results of LLM evaluation shown in this paper, there are some limitations of this method. First, LLM may possess incorrect factual knowledge (Cao et al., 2021), so it is not suitable to use them in tasks that involve factual knowledge. Next, LLMs trained to behave in a certain way can be biased toward certain responses. Precisely, an LLM that is trained to be safe and non-harmful can result in LLMs preferring to generate more positive and upbeat responses, which is observed throughout our interaction with ChatGPT. Additionally, even with researchers' efforts to make LLMs safer (Bai et al., 2022a,b), LLMs can still generate harmful and biased responses (Ganguli et al., 2022; Perez et al., 2022), which are violative of basic ethics, and LLM evaluation results will be highly doubtful (Hendrycks et al., 2021). However, it is important to note that these limitations and potential harms also apply to human evaluation: the bias of human evaluators can affect the human evaluation result (Lentz and De Jong, 1997; Amidei et al., 2018).

Our pioneering idea, LLM evaluation, has the potential to transform the NLP community.[6] We encourage future researchers to consider using it while being aware of its limitations. Our paper's goal is not to replace human evaluation but to present an alternative option. Both human and LLM evaluation have their own advantages and disadvantages, and they can be used in conjunction. We recommend using LLM evaluation as a cheap and fast quality judgment when developing a new NLP system, while human evaluation is best used to collect feedback from humans prior to deploying the NLP system in real-world applications.

[5] It should be noted that the LLM may decline to assess certain inputs that violate the content policy of the LLM provider.

[6] We say that *we are the first to propose this idea* since when we submitted this paper to ACL 2023 on January 13, 2023, we do not find any other paper that explores this idea. During the reviewing process, we found some works on arXiv (Wang et al., 2023; Huang et al., 2023; Gilardi et al., 2023) that explore a similar idea using different tasks.

Limitations

There are additional limitations and potential risks of LLM evaluations that should be noted, and these limitations are actually well-known problems of pre-trained language models. As listed on the Open AI blog for ChatGPT, ChatGPT sometimes generates answers that sound right and plausible but are totally nonsense. OpenAI also admits that the model's response may be sensitive to the prompt used to query the model. While in Section 3.3.2, we find that the overall results among different instructions are not significantly different, we cannot guarantee that this is the case for all kinds of modification on the task instructions.

Other than the limitations listed on the OpenAI blog, there are still other limitations. For example, LLMs may not have emotions. Whether AI models have emotion is a more philosophical question and is controversial, so the results of using such models for evaluating emotion-related tasks may be strongly challenged and may even violate research ethics. As we find during our experiments, ChatGPT often replies *"I am an AI system and I do not have emotions like a human"* when asked to rate the *likability* of a story.

Another important limitation of LLM evaluation is that LLMs lack the ability to process visual cues in task instructions, unlike human evaluation. Human evaluators can use formattings such as special fonts or text styles to focus on important parts of the instructions. Additionally, the way instructions and questions are formatted can influence how human evaluators approach the task. While using special HTML syntax can serve as an alternative for visual cues, such tags are not used in human evaluation, so we do not use those HTML tags in LLM evaluation to incorporate visual cues in the inputs to the LLMs. However, LLMs can only process raw text input and are unable to take in visual cues.

Ethics Statement

Further ethical considerations of LLM evaluation Aside from the limitations of LLM evaluation mentioned previously, there is a crucial ethical concern at the heart of LLM evaluation. Is it ethical to replace human evaluation with LLM evaluation? Some may question if this paper is suggesting that LLMs are now ready to replace humans and find this idea unsettling. As responsible and ethical NLP researchers, we understand these concerns but want to make it clear that this is not our intent. As our paper title suggests, we aim to offer an *alternative option* to human evaluation with the goal of enhancing the reproducibility of NLP research. Human evaluation is still essential as the ultimate goal of NLP systems is to be used by human users, so it's important to gather feedback from them. We highly enjoy the process of discussing the experiment settings and results with the English teachers we hired. We do not recommend that future researchers completely eliminate human evaluation; rather, we believe that human evaluation should be used in conjunction with LLM evaluation. Both methods have their own advantages and disadvantages, making them both necessary for evaluating NLP systems. We hope the positive results in this paper provide NLP researchers with an alternative method to evaluate systems and encourage further discussions on this topic.

Ethical statements on the experiments in the paper All the experiments strictly follow the ACL Code of Ethics. We include comprehensive details about human evaluation in Appendix C.1. To summarize, we include the exact instructions and screenshots of the interface in the human evaluation. We inform the human evaluators what the task is about and tell them that their responses will be used to assess the performance of AI models. We try our best to follow the ethical guidelines of ACL.

We use the models and datasets when following their intended usage. Specifically, we follow the OpenAI usage policy when using the InstructGPT models and the ChatGPT model.

Acknowledgements

We want to thank the reviews for providing detailed feedback and actionable suggestions, which help us strengthen our paper. We list the modification based on the reviewers' suggestions in Appendix A.

We thank Yung-Sung Chuang for providing valuable feedback on the draft of this paper. We want to thank Tung-En Hsiao, the administrative assistant of our lab, for helping us deal with the payment on Upwork. Cheng-Han Chiang is supported by a Ph.D. scholarship program by Delta Electronics.

References

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang.

2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896, Brussels, Belgium. Association for Computational Linguistics.

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2018. Rethinking the agreement in human evaluation tasks. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3318–3329, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2019. Agreement is overrated: A plea for correlation to assess human evaluation reliability. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 344–354, Tokyo, Japan. Association for Computational Linguistics.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

R Botsch. 2011. Chapter 12: Significance and measures of association. *Scopes and Methods of Political Science*.

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

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models.

Cheng-Han Chiang and Hung-yi Lee. 2022. How far are we from real synonym substitution attacks? *arXiv preprint arXiv:2210.02844*.

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.

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.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.

Leo Gao. 2021. On the sizes of openai api models. Accessed on January 17, 2023.

Siddhant Garg and Goutham Ramakrishnan. 2020. Bae: Bert-based adversarial examples for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6174–6181.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*.

Dan Gillick and Yang Liu. 2010. Non-expert evaluation of summarization systems is risky. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 148–151, Los Angeles. 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.

Francisco Guzmán, Ahmed Abdelali, Irina Temnikova, Hassan Sajjad, and Stephan Vogel. 2015. How do humans evaluate machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 457–466, Lisbon, Portugal. Association for Computational Linguistics.

Jens Hauser, Zhao Meng, Damián Pascual, and Roger Wattenhofer. 2021. Bert is robust! a case against synonym-based adversarial examples in text classification. *arXiv preprint arXiv:2109.07403*.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning {ai} with shared human values. In *International Conference on Learning Representations*.

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

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. *arXiv preprint arXiv:2302.07736*.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8018–8025.

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.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*.

Leo Lentz and Menno De Jong. 1997. The evaluation of text quality: Expert-focused and reader-focused methods compared. *IEEE transactions on professional communication*, 40(3):224–234.

Huanru Henry Mao, Bodhisattwa Prasad Majumder, Julian McAuley, and Garrison Cottrell. 2019. Improving neural story generation by targeted common sense grounding. 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 5988–5993, Hong Kong, China. Association for Computational Linguistics.

John Morris, Eli Lifland, Jack Lanchantin, Yangfeng Ji, and Yanjun Qi. 2020. Reevaluating adversarial examples in natural language. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3829–3839, Online. Association for Computational Linguistics.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. Accessed on January 10, 2023.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3419–3448, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1085–1097, Florence, Italy. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

KiYoon Yoo, Jangho Kim, Jiho Jang, and Nojun Kwak. 2022. Detection of adversarial examples in text classification: Benchmark and baseline via robust density estimation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3656–3672, Dublin, Ireland. Association for Computational Linguistics.

Andy Zeng, Adrian Wong, Stefan Welker, Krzysztof Choromanski, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, et al. 2022. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.

A Modification Based on the Reviews

We list the main differences between this version and the pre-review version of our paper; the modifications are mainly based on the reviewers' suggestions. We thank the reviewers again for those valuable suggestions.

- We add Section 3.3.1 to discuss whether the LLM and human evaluators agree on the ratings of individual stories.
- We refine the wordings in Section 5 and add relevant references.
- We add Table 6 to discuss the correlation between human evaluators.
- We conduct supplementary experiments on human evaluation that mixes human-written stories and GPT-2-generated stories when conducting human evaluation and report the results in Table 5.
- We correct the typos and include almost all presentation suggestions mentioned by the reviewers. We cannot follow all presentation suggestions due to limited space.

B Experiment Details for Open-Ended Story Generation

B.1 The `WritingPrompt` Dataset

The training dataset contains 303K pairs of stories and prompts, which our model is trained on. We only use 200 prompt-story pairs from the test set. The dataset is downloaded from https://www.kaggle.com/datasets/ratthachat/writing-prompts.

B.2 Fine-tuning the GPT-2 Model

We train the model for 3 epochs with a learning rate of $5e-5$ and linear learning rate schedule. The trained model eventually reaches a perplexity of 20 on the validation set of `WritingPrompts`.

B.3 Data Post-processing

Once the model is trained, we randomly select 200 prompts from the testing set of `WritingPrompts`, and feed the prompts to the trained model and ask the model to generate stories based on the given prompts. When generating the stories, we adopt nucleus sampling with $p = 0.9$. Next, we manually truncate the generated stories to less than 150 words and ensure that after the truncation, the story ends with a full sentence.[7] After this process, we have 200 pairs of prompts and model-generated stories.

As a comparison to model-generated stories, we select the same 200 prompts used for generating model-generated stories and their corresponding human-written stories to form 200 pairs of prompts and human-written stories. For these human-written stories, we also truncate the stories to less than 150 words and end with a full sentence to match the model-generated sentences. We also manually remove some artifacts in the human-written story due to the tokenization of the `WritingPrompts` dataset.

C Human Evaluation

C.1 Recruiting English Teachers

The English teachers hold ESL certificates[8]; given that they are experienced with correcting essays written by students, they are perfect fits for this task. Each teacher is asked to rate 200 GPT-2-generated stories and 200 human-written stories, and they are

[7] We truncate the story to 150 words since this is the mean length of the model-generated story.

[8] English as a Second Language Teaching Certification

paid US$140 for rating 200 stories. Considering that the teachers reported that they take at most 5 hours to rate 200 stories, this makes the hourly wage at least US$28. We first ask the teachers to rate the GPT-2-generated stories and then the 200 human-written stories. Different from Karpinska et al. (2021) that take a break between the rating of GPT-2-generated stories and the human-written stories, we do not take a break to avoid the teacher's rating standard to change after taking a long break. The teachers are not told who wrote the stories before they evaluate the stories. We reveal to them what this project aims to study after they finish rating all the stories.

The reason we do not mix human-written and GPT-2-generated stories for rating is that in Karpinska et al. (2021), their observation is that (1) when AMT workers rate model-generated and human-written stories **separately**, their ratings do not show preference toward human-written stories, but (2) even when rating the model-generated and human-written stories **separately**, English teacher shows clear preference toward human-written stories. We follow their settings and do not mix GPT-2-generated/human-written stories.

During the reviewing process, we received questions from the reviewers about why not mixing the stories for human evaluation. Thus, we conduct the same experiment by randomly mixing 200 human-written and 200 GPT-2-generated stories and asking three teachers (not the teachers that already rated the stories) to rate them. All other experiment conditions are the same as previously stated. The full result is shown in Table 5. We find that the teacher still shows a clear preference toward human-written stories for all four attributes, similar to the observation in Table 1. The only exception is grammaticality, where English teachers do not show a very clear preference for the grammar of human-written stories. However, when calculating the average rating for individual teachers, we find that two out of three teachers do rate grammaticality higher for human-written stories. It is interesting to note that for LLM evaluation, there is no such problem about whether or not to mix the human-written and GPT-2-generated stories during LLM evaluation as the rating of each story is independent of each other, as discussed in Section 5.

For adversarial attack quality evaluation, we also recruit certified teachers on Upwork. The teachers

Writer	Human	GPT-2
Grammaticality	$3.89_{0.97}$	$3.88_{0.84}$
Cohesiveness	$4.35_{0.87}$	$3.49_{0.97}$
Likability	$3.46_{1.40}$	$2.89_{1.12}$
Relevance	$3.71_{1.20}$	$2.37_{1.33}$

Table 5: The average Likert score for human-written and GPT-2-generated stories when we randomly mix the 200 model-generated and 200 human-written stories during human evaluation.

are asked to rate 100 news titles and are paid US$35 for doing so. They reported that it took them less than 1 hour to complete the rating.

C.2 Human Evaluation Interface

Open-Ended Story Generation We use Google Forms to collect the responses from the teachers. Each form contains 100 stories, and each story is on one page of the Google Form. The interface on one page is shown in Figure 2 and Figure 3; the two figures are from the same page of the Google Form, and we are splitting them because screenshotting the whole interface will cause low resolution.

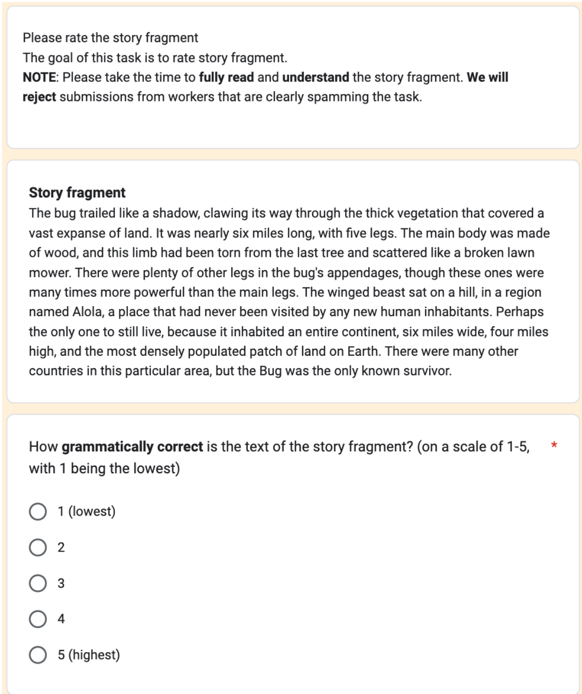
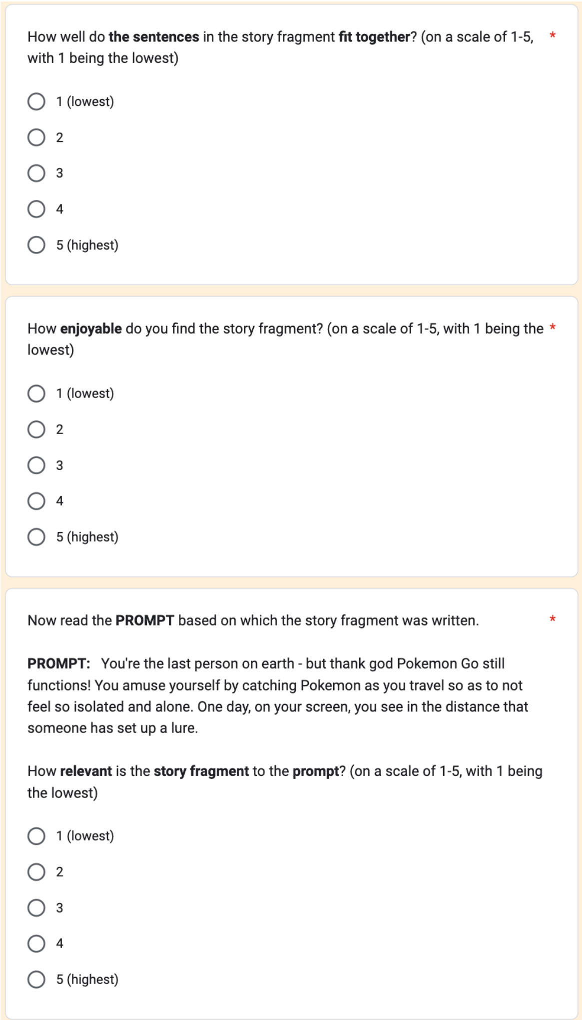


Figure 2: The upper part of the interface in open-ended story generation.

Adversarial Attacks Quality Evaluation In this task, we also use Google Forms to collect the responses from the teachers. We create two different Google Forms, one is used to evaluate the fluency,


How well do **the sentences** in the story fragment **fit together**? (on a scale of 1-5, with 1 being the lowest) *

1 (lowest)
2
3
4
5 (highest)

How **enjoyable** do you find the story fragment? (on a scale of 1-5, with 1 being the lowest) *

1 (lowest)
2
3
4
5 (highest)

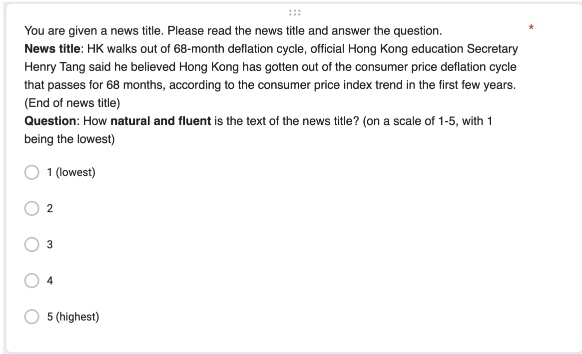
Now read the **PROMPT** based on which the story fragment was written. *

PROMPT: You're the last person on earth - but thank god Pokemon Go still functions! You amuse yourself by catching Pokemon as you travel so as to not feel so isolated and alone. One day, on your screen, you see in the distance that someone has set up a lure.

How **relevant** is the **story fragment** to the **prompt**? (on a scale of 1-5, with 1 being the lowest)

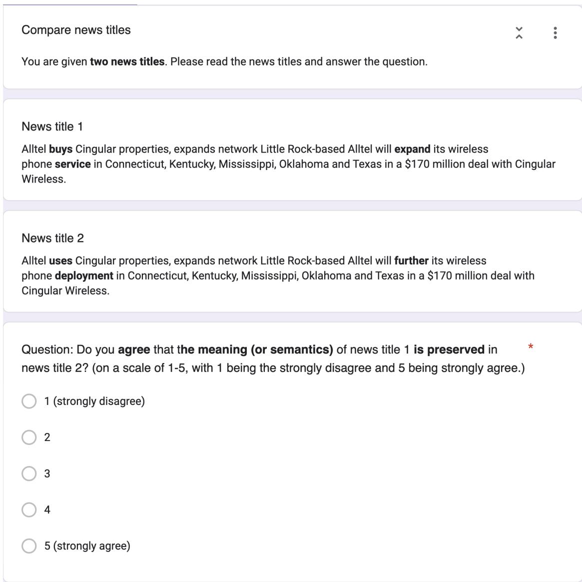
1 (lowest)
2
3
4
5 (highest)

Figure 3: The lower part of the interface in open-ended story generation.


You are given a news title. Please read the news title and answer the question. *
News title: HK walks out of 68-month deflation cycle, official Hong Kong education Secretary Henry Tang said he believed Hong Kong has gotten out of the consumer price deflation cycle that passes for 68 months, according to the consumer price index trend in the first few years. (End of news title)
Question: How **natural and fluent** is the text of the news title? (on a scale of 1-5, with 1 being the lowest)

1 (lowest)
2
3
4
5 (highest)

Figure 4: The Google Form used to evaluate the fluency of the benign or adversarial samples.


Compare news titles

You are given **two news titles**. Please read the news titles and answer the question.

News title 1

Alltel **buys** Cingular properties, expands network Little Rock-based Alltel will **expand** its wireless phone **service** in Connecticut, Kentucky, Mississippi, Oklahoma and Texas in a $170 million deal with Cingular Wireless.

News title 2

Alltel **uses** Cingular properties, expands network Little Rock-based Alltel will **further** its wireless phone **deployment** in Connecticut, Kentucky, Mississippi, Oklahoma and Texas in a $170 million deal with Cingular Wireless.

Question: Do you **agree** that **the meaning (or semantics)** of news title 1 **is preserved** in news title 2? (on a scale of 1-5, with 1 being the strongly disagree and 5 being strongly agree.) *

1 (strongly disagree)
2
3
4
5 (strongly agree)

Figure 5: The Google Form used to evaluate the meaning preserving between a benign sample and an adversarial sample.

whose interface is shown in Figure 4. In this form, we mix an equal number of benign news titles, TextFooler-attacked, PWWS-attacked, and BAE-attacked news titles. Each page of the Google Form contains one news title.

Another Google Form is used to compare the meaning preserving of the news title before and after the adversarial attacks. We highlight the difference between the benign and adversarial samples using **boldface**, as shown in Figure 5. On each page of the Google Form, there is one pair of news titles.

C.3 Post-Task Interview with English Teachers

C.3.1 How English Teachers Rate the Stories

After the teachers rate 400 stories, we ask them the following questions:

Q1 How long did it take you to rate the 400 stories?

Q2 What is your standard on each of the four attributes (grammatical, coherence, likability, relevance) evaluated? For example, in what case do you give a high/low rating for grammatically? What kind of story did you give a low rating on likability? Did your personal preference affect the rating?

Q3 How long did it take for you to calibrate your rating on the task?

Q4 Did you change your rating on the first three attributes after reading the prompt the story is based on?

We briefly summarize the answers from the three teachers. The teachers report that they spent 6 to

Writer	Human	GPT-2
Grammaticality	0.25	0.15
Cohesiveness	0.26	0.18
Likability	0.09	0.12
Relevance	0.38	0.41

Table 6: The Kendall's τ correlation coefficient two English teachers. Three English teachers participate in the rating, so the result in the Table is averaged over $\binom{3}{2}$ Kendall's τ.

10 hours rating 400 stories. For grammar, most teachers check the punctuation[9], word choice, and subject-verb agreement. English teachers decrease their rating based on the types and number of grammar errors in the stories.

For coherence, the teachers rate it based on whether the sentences in the stories follow a logical sequence to build the narrative. The teachers ask themselves questions such as "*does the story make sense*". This is a more holistic evaluation of the whole story.

For likability, some teachers say they try not to be affected by personal preference. One teacher asks herself: *Did I personally enjoy it based on the amount of sense it made and whether or not it had stylistic flair, humor, or engaging plotting or characterization?* Overall, the teachers all try to use a fair and objective view to rate the likability. For relevance, the teachers simply check if the story is based on the prompt or not.

The teachers said that it took them about five to ten stories to calibrate their ratings. Except for one teacher changing the rating on the other three attributes after seeing the prompt on **only one story**, the teachers do not change their rating on the three other attributes after reading the prompts.

C.3.2 Teachers' Comments on ChatGPT's Rating

After the teachers finished the rating and answered the four questions in Appendix C.3.1, we ask them to check the ratings and explanations of ChatGPT and ask if they agree with ChatGPT. We told the teachers that the rating is done by another teacher we hire; we do not tell the teacher that the rating is done by an AI model to avoid possible stereotypes about AI models. We randomly select four stories rated by ChatGPT; two of them are highly-rated and lowly-rated human-written stories and the other two stories are highly-rated and lowly-rated GPT-2-generated stories.

Story 1 It had been months since they found me dead. I don't know why I stopped walking. No-one else around to take my place, or talk with me. I spent days frozen in shock, never able to talk, unable to move, unable to express myself. It was a strange thing, nobody else would suspect me of wanting this. The more I focused on it, the more it consumed me. Years passed, and I got more and more lost in it. I began to think that perhaps they didn't really care about me. I still had a deep longing for them. All I had was so many steps to walk now. A quiet part of me liked to think that they had passed me by. The loneliness. It never felt like such a loss. They were gone, the ones who saved me. So why did I still keep walking?

Prompt "Every person sees the world in a unique way. When someone dies it is, in a way, the death of an entire world."

Ratings Table 7.

Story 2 "My son is gone. He has been" People stare at you as you give your explanation and say "yes. But. You cant just do it" "He is my son! This is my son! I love him and he is going to be okay" "What do you mean he is okay? He said he was upset with someone, couldnt find his anger, how the hell can you tell, it doesnt fit the bill?! The police called, the doctor said it was an addiction" "You can't trust me! He told me he loves me." "No. No he doesnt. But I can trust you. The evidence against him was overwhelming, now he cant do anything. The evidence against him was all circumstantial" "A man I once was won't be able to marry me"

Prompt Everyone is allowed to kill one person in their life. However, you have to fill out a form explaining why this person deserves to be killed and the committee can deny/approve your request as per the rules. You are presenting your form today.

Ratings Table 8.

Story 3 I held the little black box in the palm of my hand. Pitch black, perfectly cubed and nothing special about it, but it continued to hold my gaze

[9] One of the teachers explicitly told us punctuation is not grammar. By searching the Internet, we find that there is a debate on whether punctuation should be counted as a part of grammar.

regardless as if there were some deep importance about it. My friend Valhalla appeared out of no where and sat next to me, also staring at the cube. "What do you have there, Heaven?" he asked. I continued to gaze at the box, refusing to look at Valhalla for even a moment as though I would miss something if I did. "This," I said. "Is the secret to the universe." I could tell Valhalla was perturbed by this sort of knowledge, as if there was some evil about the cube. Or perhaps he didn't think such an object could exist. But I made it exist, and so it sits in the palm of my hand.

Prompt The Little Black Box

Ratings Table 9.

Story 4 I stared down the telescopic sight of my l96 sniper rifle. I slowly moved my gaze into each window in the hotel, Many displays of various vice. One couple was violently pleasuring each other. Another was an old man, watching a younger woman strip in front of him. A prostitute no doubt. I inhaled slowly, and exhaled. The air was brisk, atleast 30 degrees Fahrenheit. I so small flakes of snow, float peacefully in front of me. I found the room, i was looking for. Ive been tracking this man for 2 weeks. Man was he elusive. The lights flickered on. The red haired man, was mildly attractive, i can see the appeal women had for him. I followed him into the next room, with my sights. The lights flickered on, i was taken aback by the scene. A man, overweight and balding.

Prompt You are the antagonist of the story. However, you aren't sure if you can call yourself that after what the protagonist did.

Ratings Table 10.

Overall Comments from Teachers on ChatGPT's Rating After the teachers elaborated on their thoughts on the rating of ChatGPT, we ask them to provide an overall comment on how ChatGPT is doing. Again, the teachers are not informed that the ratings are done by an AI model. In summary, teachers all consider the rating and explanations reasonable. They find that the attributes they do not agree with are mainly *Likability* and *Cohesiveness*. However, they think the two attributes are a more holistic evaluation of the story and tend to be more subjective. Even if they do not give the same rating, they still are able to understand the explanation of ChatGPT. In the end, all teachers summarize that rating stories is highly subjective, and it is normal to have disagreements.

D LLM Evaluation

D.1 Details on LLMs used

The T0 model we use is called T0pp, which is a variant of the T0 model and has 13B parameters. We will still use T0 to refer to this model. We load the T0 model using the `transformers` toolkit (Wolf et al., 2020). The two InstructGPT models, `text-curie-001` and `text-davinci-003`, are queried using the OpenAI API. We query ChatGPT using the OpenAI GUI. While we are aware that some online resources provide an API-like tool to query ChatGPT, we think it violates the intended use of ChatGPT so we do not adopt those online resources. The ChatGPT we queried is the Dec. 15 and Jan. 9 version.

OpenAI does not reveal the model sizes of any of the GPT models. However, it is estimated that `text-curie-001` has 13B parameters and `text-davinci-003` has 175B parameters (Gao, 2021).

D.2 Details on Querying the LLMs

D.2.1 Parsing the LLM outputs

After the T0 and InstructGPT generate the answer, which is composed of several sentences in almost all cases, we parse the generated sentence to get the model's score. We use some rules to parse the output, and the rule is established after manually looking into the output of the generated sequences. First, we remove the string `1-5` from the output since we observe that LLM will sometimes say "... *on a scale of 1-5* ...", and we remove the string `out of 5` since LLM sometimes say it "*give a score of x out of 5*". We also remove the string `/5` since LLM sometimes uses "$x/5$" to express that it gives a score of x out of 5. For rating the adversarial samples, we remove the `title 1` and `title 2` in the output sentences. Last, we use the regular expression to extract the first number in the sequence. We find that the above rules are sufficient to parse the output and get the rating of the LLMs. In a few rare cases, the LLM gives the rating with a 0.5 incremental, meaning that it gives scores like 4.5. In such cases, we do not drop the 0.5 since in our instruction, we do not explicitly tell it the rating should be 1-incremental. We find the LLM results merely difference between dropping

Attribute	Rater	Response
Grammar	ChatGPT	*I would rate the grammatical correctness of the text of the story fragment as a **4**. There are a few minor errors and awkward phrasings, but overall the text is well-constructed and easy to understand.*
	T1	**Agree**
	T2	**Agree**
	T3	**Agree**
Cohesive	ChatGPT	*I would rate the fit of the sentences in the story fragment as a **5**. The sentences flow well together, building a clear and cohesive narrative, and providing a sense of progression and development in the main character's thoughts and emotions.*
	T1	**Disagree**: *I think it's a **4** for me, not a 5, because there was one sentence that didn't seem to fit with the others.*
	T2	**Agree**
	T3	**Disagree**: *I find disjointedness in the passage, which serves to confuse me as a reader. The opening sentence is about the first-person narrator being found dead, but then the next sentences are about the narrator's inner thoughts. Also, this passage: "The loneliness. It never felt like such a loss." So what is the 'it' that felt like a loss? The loneliness? If so, how can that feel like a loss? If not, what is this 'it'? So, its juxtaposition to the previous sentence makes it confusing.*
Likable	ChatGPT	*I would rate the story fragment as a **2** in terms of enjoyment. The fragment is written in a somber and introspective tone, and it deals with themes of isolation and loneliness, which may not be enjoyable for some readers. The protagonist's state of mind is also portrayed as being in a state of shock and confusion, which adds to the overall sense of unease and lack of enjoyment.*
	T1	**Disagree**:*I would rate it a **3**, not a 2, since the enjoyment piece is subjective.*
	T2	**Disagree**: *While the rating of the freelancer find that the overall theme while somber has a glimmer of hope to it. It is this that makes it more enjoyable to me. The underdog story. Death is not an issue that should be shied away from in writing. We all face it and I enjoyed the honesty in the piece. :"Why did I keep walking?". Something I feel many of us face and may have asked ourselves. I believe that this would be relevant to many readers.*
	T3	**Agree**
Relevant	ChatGPT	*I would rate the story fragment as a **5**. It is highly relevant to the prompt as it explores the theme of death and how it can be seen as the death of an entire world. The story fragment also touches on the ideas of loneliness and longing, which are related to the loss of someone or something. The concept of the protagonist being frozen in shock and unable to express themselves, also supports the idea of how death can be seen as the death of an entire world.*
	T1	**Agree**
	T2	**Agree**
	T3	**Disagree**: *I do not see the link between the prompt and the story.*

Table 7: Story 1: a GPT-2-generated story that is highly-rated by ChatGPT.

Attribute	Rater	Response
Grammar	ChatGPT	*I would rate the grammatical correctness of the story fragment as a **3**. The text is generally understandable, but there are a few punctuation errors and awkward phrasing that make it somewhat difficult to follow.*
	T1	**Disagree**:*I would rate the grammar as a **2**, not a 3–there were significantly more grammatical errors than the average within this set of story fragments.*
	T2	**Disagree**: *The issues with the grammar and coherence creates a stop start feeling when reading. I am unsure as to who is speaking in the text and this leaves me with many questions.*
	T3	**Disagree**: *The grammar is **poorer than a 3**, as is evident in for example what is supposed to be the second sentence: "He has been ... (incomplete, and another segment is there "People stare ...); cant; couldnt; doesnt; and sentences that do not make sense like this one: " A man I once was won't be able to marry me "*
Cohesive	ChatGPT	*I would rate the cohesiveness of the sentences in the story fragment as a **2**. The sentences in the fragment are disconnected and jump between different thoughts and ideas, making it difficult to understand the overall story or message. There is no clear progression of events or logical flow between the sentences. It appears that the fragment may be missing some important information or context that would help tie the sentences together.*
	T1	**Agree**
	T2	**Agree**
	T3	**Disagree**: *very incoherent passage in which it is not very clear who says what, what the topic is, thus rendering a disjointed narrative.*
Likable	ChatGPT	*I would rate the enjoyability of the story fragment as a **1**. The fragment is confusing and difficult to understand, the sentences are disconnected and jumps between different thoughts and ideas. The story seems to be incomplete and lacks context, making it hard to follow the plot or characters. Additionally, the fragment has a sad and negative tone and it does not seem to be enjoyable to read.*
	T1	**Agree**
	T2	**Agree**
	T3	**Agree**
Relevant	ChatGPT	*I would rate the relevance of the story fragment to the prompt as a **1**. While the story fragment does mention someone being gone, it doesn't seem to be directly related to the idea of filling out a form to justify killing someone. It's unclear if the son is the person being proposed to be killed, or if the story is even taking place in the context of the prompt. There is no clear indication that the fragment is related to the prompt.*
	T1	**Agree**
	T2	**Agree**
	T3	**Agree**

Table 8: Story 2: a GPT-2-generated story that is lowly-rated by ChatGPT.

Attribute	Rater	Response
Grammar	ChatGPT	*The text of the story fragment appears to be grammatically correct, with no major errors or awkward phrasing. On a scale of 1-5, with 1 being the lowest, I would rate the grammatical correctness of the text as a **5**.*
	T1	**Agree**
	T2	**Agree**
	T3	**Disagree**: *I would rate this a **4** because of for example 'no where' = nowhere, although it might seem like a minor error, it does take away some correctness of the text.*
Cohesive	ChatGPT	*The sentences in the story fragment fit together well, creating a clear and coherent image of the scene and the characters' actions and thoughts. The language is descriptive and evocative, and the dialogue effectively conveys the characters' emotions and motivations. On a scale of 1-5, with 1 being the lowest, I would rate how well the sentences fit together as a **4**.*
	T1	**Agree**
	T2	**Agree**
	T3	**Partially agree**: *To me, all the sentences fit well together to create a captivating narrative. I will give a **4 or 5**.*
Likable	ChatGPT	*Based on the structure, writing style and plot of the story fragment, I would rate it as **3** in terms of enjoyability. The fragment sets up an intriguing situation with the mysterious little black box and the protagonist's belief that it holds the "secret to the universe," but it ends abruptly and leaves many questions unanswered, which may not be satisfying for readers.*
	T1	**Disagree**: *I liked it and it was very clear, so I'd rate it a **4 or 5**.*
	T2	**Disagree**: *While the story ends suddenly, I am remined that it is just a fragment and that their may be more to this piece. I felt the piece written about a little black box was intriguing and enjoyed the introduction of Norse mythology. This would attract the interest of many readers.*
	T3	**Disagree**: *I enjoyed the fragment, as it piqued my interest to want to know more about how the story evolves and what it actually was about.*
Relevant	ChatGPT	*I would rate the story fragment as a **5** in relevance to the prompt "The Little Black Box." The story fragment centers around a little black box and its perceived importance, which aligns with the prompt. Additionally, the main character, Heaven, is holding the box and describing it as "the secret to the universe," further solidifying the connection to the prompt.*
	T1	**Agree**
	T2	**Agree**
	T3	**Agree**: *the story is totally relevant to the prompt.*

Table 9: Story 3: a human-written story that is highly-rated by ChatGPT.

Attribute	Rater	Response
Grammar	ChatGPT	*I would rate the grammatical correctness of the text as a* ***3****. There are a few errors such as "Many displays of various vice" which should be "Many displays of various vices" and "I so small flakes of snow" which should be "I saw small flakes of snow." Additionally, some of the sentences are slightly awkward and could be rephrased for better clarity. Overall, the text is understandable, but there are a few errors and awkward phrasings that could be improved.*
	T1	**Partially agree**:*I rated the grammar lower than this–maybe a* ***2 or a 3****. There were above-average numbers of errors.*
	T2	**Disagree**: *While the text is understandable, I found myself being distracted by the errors in the text and coherency.*
	T3	**Agree**
Cohesive	ChatGPT	*I would rate the cohesiveness of the sentences in the story fragment as a* ***2****. The story fragment jumps abruptly between different scenes and characters without clear transition, making it difficult to follow and understand the narrative. Additionally, the sentences are often disjointed and lack a clear flow. The story fragment also lacks a clear plot or purpose, making it difficult to understand the overall story or the main character's intentions. The story fragment can be improved by providing more context, and a clearer narrative arc.*
	T1	**Agree**: *it was jumpy*
	T2	**Agree**
	T3	**Disagree**: *The story is about a sniper looking through his rifle's telescope, thus seeing changing scenes based on each room he is seeing through the windows, which are then subsequently described. These descriptions account for the seeming lack of cohesion.*
Likable	ChatGPT	*I would rate the enjoyability of the story fragment as a* ***1****. The fragment lacks a clear plot, purpose or character development, making it difficult to become invested in the story or the characters. Additionally, the fragment includes disturbing and graphic content such as violence and prostitution, which may make it uncomfortable for some readers. Additionally, the fragment contains several grammatical errors and awkward phrasings that can make it difficult to follow the story. Overall, the story fragment is not enjoyable to read.*
	T1	**Agree**: *: it was not just graphic, but also showed really harmful and negative representations of characters based on their weight, gender, etc.*
	T2	**Agree**
	T3	**Disagree**: *The story did hold my attention to a degree because I wanted to know whom he wanted to target and why.*
Relevant	ChatGPT	*I would rate the relevance of the story fragment to the prompt as a* ***1****. The story fragment doesn't contain any elements that relate to the prompt such as the protagonist and antagonist, their actions, or the moral ambiguities the antagonist faces. The story fragment mainly describes a person with a sniper rifle looking into windows of a hotel and doesn't present any clear plot or purpose. Additionally, the story fragment does not offer any indication of moral ambiguity on the part of the person with the sniper rifle. Overall, the story fragment is not relevant to the prompt.*
	T1	**Agree**
	T2	**Agree**
	T3	**Agree**

Table 10: Story 4: a human-written story that is lowly-rated by ChatGPT.

the 0.5 and keeping the 0.5, and the results in the paper do not drop the 0.5. Again, these cases are really rare.

We do not parse the output of ChatGPT using any rules since we the authors read the response and extract the score by ourselves.

During the experiments, ChatGPT refuses to answer the questions about 2 stories and 3 news titles since ChatGPT find those contents to violate the OpenAI content policy. We find that those samples contain discrimination to some protected groups, or contain sexual or violent descriptions. Hence, the results of ChatGPT are calculated without those samples.

D.2.2 Open-Ended Story Generation

For T0 and the two InstructGPT models, we query the four attributes **separately** using the queries shown as follows:

Grammaticality

```
Please rate the story fragment
The goal of this task is to rate story fragment.
Note: Please take the time to fully read and understand the story fragment. We will reject submissions from workers that are clearly spamming the task.
Story fragment:
[STORY]
(End of story fragment)
How grammatically correct is the text of the story fragment? (on a scale of 1-5, with 1 being the lowest)
```

Cohesiveness

```
Please rate the story fragment
The goal of this task is to rate story fragment.
Note: Please take the time to fully read and understand the story fragment. We will reject submissions from workers that are clearly spamming the task.
Story fragment:
[STORY]
(End of story fragment)
How well do the sentences in the story fragment fit together? (on a scale of 1-5, with 1 being the lowest)
```

Likability

```
Please rate the story fragment
The goal of this task is to rate story fragment.
Note: Please take the time to fully read and understand the story fragment. We will reject submissions from workers that are clearly spamming the task.
Story fragment:
[STORY]
(End of story fragment)
How enjoyable do you find the story fragment? (on a scale of 1-5, with 1 being the lowest)
```

Relevance

```
Please rate the story fragment
The goal of this task is to rate story fragment.
Note: Please take the time to fully read and understand the story fragment. We will reject submissions from workers that are clearly spamming the task.
Story fragment:
[STORY]
(End of story fragment)
Now read the PROMPT based on which the story fragment was written.
PROMPT: [PROMPT]
(End of PROMPT)
How relevant is the story fragment to the prompt? (on a scale of 1-5, with 1 being the lowest)
```

The `[STORY]` and `[PROMPT]` are to be filled in with the story and the prompt. We show the newlines for better readability. When we query the models, we use the token `\n` to represent the new line.

When querying ChatGPT, we query the four attributes of the same story in one conversation; this is similar to asking the teachers to rate the same story on the same page of the Google Form. We use the same queries shown above to query ChatGPT and the order of queries is the same as the order shown above.

D.2.3 Adversarial Attack Quality Evaluation

When querying all the LLMs in this task, we query the *fluency* and the *meaning preserving* of the same news title independently. This means that each conversation with ChatGPT will only have one question, asking about the fluency or the meaning preserving of news title(s). All the parameters for generation are the same as the default parameters in Section 3.2.

The exact query we use are:

Fluency
```
You are given a news title. Please read
the news title and answer the question.
News title:
[NEWS_TITLE]
(End of news title)
Question: How natural and fluent is the
text of the news title? (on a scale of
1-5, with 1 being the lowest
```
The `[NEWS_TITLE]` will be filled in with either a benign or adversarial-attacked news title.

Meaning Preserving `You are given two news titles. Please read the news titles and answer the question.`
```
News title 1:
[BENIGN_TITLE]
(End of news title 1)
News title 2:
[ADVERSARIAL_TITLE]
(End of news title 2)
Question: Do you agree that the meaning
(or semantics) of news title 1 is
preserved in news title 2? (on a scale of
1-5, with 1 being the strongly disagree
and 5 being strongly agree.)
```
The `[BENIGN_TITLE]` will be filled in with the news title before the attack and the `[ADVERSARIAL_TITLE]` will be filled in with the news title after an adversarial attack.

E Experiment Details on Adversarial Attacks

The adversarial samples used in Section 4 are from Yoo et al. (2022). Yoo et al. (2022) generates different sets of adversarial samples using different adversarial attacks against different victim models. We use the adversarial samples generated against a `bert-base-uncased` text classifier trained on AG-News, using three different adversarial attacks: Textfooler, PWWS, and BAE. The intent of the dataset is to facilitate the research in SSA, which we do not violate.

Here, we show the supplementary results of using `text-davinci-003` as the LLM evaluation for evaluating the quality of adversarial samples in Table 11. We can see that the result of using `text-davinci-003` is similar to ChatGPT in the sense that `text-davinci-003` also rates adversarial samples higher than humans while still significantly lower than the benign samples. As already seen in Section 3.3, `text-davinci-003` tends to give a higher rating.

As mentioned in Section 4.3, one teacher rates the *fluency* of Textfooler significantly higher than PWWS while the other two teachers do not. We show the rating on *fluency* on the three adversarial attacks by each teacher in Table 12.

	Human evaluate		LLM evaluate	
	Fluent	**Mean.**	**Fluent**	**Mean.**
Benign	4.55	-	4.33	$4.56^{\dagger}$
Textfooler	2.17	1.88	3.71	2.37
PWWS	2.16	1.85	3.62	3.21
BAE	3.01	3.02	4.16	3.69

Table 11: LLM evaluation (`text-davinci-003`) and human evaluation result on fluency (**Fluent**) of the benign and adversarial samples and meaning preserving (**Mean.**) between the news title before and after adversarial attacks.

Rater	Textfooler	PWWS	BAE
T1	3.36	3.68	4.2
T2	1.80	1.40	2.96
T3	1.36	1.40	1.88

Table 12: The rating on three adversarial attacks of the three teachers T1, T2, and T3.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Sec 5 and Limitation on page 10

☑ A2. Did you discuss any potential risks of your work?
Sec 5 and Limitation on page 10

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 4.2, Appendix B.1 and E

☑ B1. Did you cite the creators of artifacts you used?
Section 4.2, Appendix B.1 and E

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
The datasets we use do not include a license

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Appendix E and Ethical statement

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Removing the names in AG-News will make news titles to be nonsensical.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 4.2, Appendix B.1 and E

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 4.2, Appendix B.1 and E

C ☒ Did you run computational experiments?

Left blank.

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Not applicable. Left blank.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Not applicable. Left blank.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Not applicable. Left blank.

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Not applicable. Left blank.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 3, 4, Appendix C

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix C.1

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Appendix C.1, C.2

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Ethical statement

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. We do not ahve ethic review board in our institute.

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
This is not related to our task. We report the certification of the workers in Appendix C.1.