Exploring the Use of Large Language Models for Reference-Free Text Quality Evaluation: An Empirical Study

Yi Chen[♡♠*], **Rui Wang**[♡♠*], **Haiyun Jiang**[†], **Shuming Shi**, **Ruifeng Xu**[♡♣♠†]
[♡]Harbin Institute of Technology, Shenzhen, China
[♣]Peng Cheng Laboratory, Shenzhen, China
[♠]Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies
yichennlp@gmail.com,ruiwangnlp@outlook.com,xuruifeng@hit.edu.cn

Abstract

Evaluating the quality of generated text is a challenging task in NLP, due to the inherent complexity and diversity of text. Recently, large language models (LLMs) have garnered significant attention due to their impressive performance in various tasks. Therefore, we present this paper to investigate the effectiveness of LLMs, especially ChatGPT, and explore ways to optimize their use in assessing text quality. We compared three kinds of reference-free evaluation methods. The experimental results prove that ChatGPT is capable of evaluating text quality effectively from various perspectives without reference and demonstrates superior performance than most existing automatic metrics. In particular, the Explicit Score, which utilizes ChatGPT to generate a numeric score measuring text quality, is the most effective and reliable method among the three exploited approaches. However, directly comparing the quality of two texts may lead to suboptimal results. We believe this paper will provide valuable insights for evaluating text quality with LLMs and have released the used data[1].

1 Introduction

Automated evaluation of text generation quality has posed a long-standing challenge in the field of natural language processing (NLP). On the one hand, the diverse forms of textual expression make it impossible for reference-based methods to account for all possible situations(Zhang* et al., 2020; Yuan et al., 2021; Chen et al., 2022b). On the other hand, devising reliable metrics without reference is not a straightforward task and can also be problematic (Sun and Zhou, 2012; Niu et al., 2021; Shen et al., 2022). Furthermore, different types of text necessitate evaluation of distinct aspects, e.g. coherence, fluency, and consistency (Fabbri et al., 2021a; Mehri and Eskenazi, 2020a; Wang et al., 2023b), which makes it hard to design metrics for each type of text and dimension separately.

Nowadays, large language models (LLMs) (Brown et al., 2020; Ouyang et al., 2022; Chung et al., 2022; Chowdhery et al., 2022; Zhang et al., 2022; Touvron et al., 2023; Du et al., 2022) represented by ChatGPT[2] have revolutionized the field of NLP by achieving remarkable results in a wide range of NLP tasks (Song et al., 2023; Chen et al., 2022a). Recent studies (Fu et al., 2023; Wang et al., 2023a; Kocmi and Federmann, 2023; Ji et al., 2023) have also demonstrated the potential of LLMs in evaluating the quality of generated texts. In this paper, we present an empirical study that compares different methods for text quality evaluation using LLMs in a reference-free mode. The key insights from our empirical findings are as follows:

- **How accurately can ChatGPT assess text quality without references?** (§4.1)

It is feasible for ChatGPT to evaluate text quality without reference, and it outperforms commonly used metrics even with a simple prompt design.

- **What is the most effective approach to evaluate text quality using ChatGPT?** (§4)

Generally, using ChatGPT to generate an explicit score for text quality is the best and most stable method among the three we compared. We suggest using greedy decoding for more reliable results.

- **Why may directly comparing two texts using ChatGPT yield suboptimal results?** (§5.1)

Due to its strict standard for "high-quality" text, ChatGPT often considers most generated texts unsatisfactory. Therefore, distinguishing between two subpar texts becomes challenging for ChatGPT.

- **Why is Implicit Score generally less effective than Explicit Score?** (§5.2)

*Equal Contribution.
†Corresponding Authors.
[1]https://github.com/MilkWhite/LLMs_for_Reference_Free_Text_Quality_Evaluation
[2]https://openai.com/blog/chatgpt

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023, pages 361–374
November 1–4, 2023. ©2023 Asian Federation of Natural Language Processing

Compared to generating an Explicit Score with ChatGPT, using the confidence of text-davinci models to determine text quality (Implicit Score) is less effective due to different distribution characteristics. Implicit Score has a narrow range and peak structure, while Explicit Score allows better differentiation with its smoother distribution.

• **How can prompt design impact ChatGPT in generating an Explicit Score?** (§5.3)

When prompting ChatGPT for an Explicit Score, it would be better to avoid detailed scoring criteria if such criteria lack clear definitions for each score range. A general description of the evaluation standard is enough. Also, making ChatGPT provide justifications in a "chain-of-thought" manner before scoring can lead it to prioritize its reasoning process over the text. These justifications tend to be templated and similar across different texts, reducing the discriminative power of the final score.

2 Method

We explore two different reference-free paradigms, i.e., *Individual Score* and *Pairwise Comparison* for text evaluation using ChatGPT and text-davinci models. Individual Score assesses the quality of a single text by a numerical score, while Pairwise Comparison focuses on the relative quality of two texts and requires a direct comparison to determine which one is superior. Within the Individual Score paradigm, two methods are typically exploited: *Explicit Score*, obtained through direct text generation, and *Implicit Score*, obtained through the token probabilities outputted by the model.

2.1 Individual Score

Explicit Score Conditioned on a given input text (optional), we prompt ChatGPT to directly generate a score to measure the absolute quality of each text individually in terms of a specific aspect or the overall performance. An example prompt designed for scoring the overall quality of a storyline is shown as follows:

========== Prompt for Explicit Score ==========

Score the following storyline given the beginning of the story on a continual scale from 0 (worst) to 100 (best), where a score of 0 means "The storyline makes no sense and is totally not understandable" and a score of 100 means "The storyline is perfect-written and highly consistent with the given beginning of the story".

The beginning of the story:
`[Conditioned Text]`

Storyline:
`[Generated Text]`

Score:

Implicit Score Given the LLM's potential insensitivity to numerical values and the lack of explicit instructions for aligning score intervals with specific criteria, score fluctuations may occur across different samples. Therefore, we propose an alternative approach by framing the problem as a binary Yes or No question, where the confidence level of answering "yes" serves as the Implicit Score. An illustrative example is presented below:

========== Prompt for Implicit Score ==========

Consider the following storyline written according to the given beginning of the story:

The beginning of the story:
`[Conditioned Text]`

Storyline:
`[Generated Text]`

Question: Is the storyline well-written and consistent with the beginning of the story?

Answer:

Unfortunately, access to ChatGPT's token probabilities is currently unavailable. Text-davinci-003 is similar to ChatGPT in that they are both trained through supervised instruction tuning and Reinforcement Learning from Human Feedback (RLHF) based on GPT-3.5, and they both exhibit excellent performance in following and fulfilling human instructions. Therefore, we utilize text-davinci-003 to derive the Implicit Score as a baseline metric instead. To facilitate a more comprehensive comparison, we also obtain the Implicit Score from text-davinci-001, an earlier version of the text-davinci series model which is based on GPT-3 and has not been trained using RLHF. Due to a limitation of the OpenAI API, only the top 5 most probable tokens are returned with log probabilities. Therefore, we instead estimate the Implicit Score using the following formula:

$$
\begin{aligned}
p(\text{yes}) &= \sum_{t \in \mathcal{T}_{top5} \cap \mathcal{T}_{yes}} p(t), \\
p(\text{no}) &= \sum_{t \in \mathcal{T}_{top5} \cap \mathcal{T}_{no}} p(t), \\
\text{Implicit Score} &= \max(p(yes), 1 - p(no)).
\end{aligned} \tag{1}
$$

Here, $p(t)$ represents the probability of predicting token t immediately following the prompt "Answer:". The sets $\mathcal{T}_{yes}$ and $\mathcal{T}_{no}$ consist of the affirmative and negative response tokens, respectively, i.e., $\mathcal{T}_{yes}$ = {" Yes", "Yes", " yes", "yes"}, and $\mathcal{T}_{no}$ = {" No", "No", " no", "no"}.

2.2 Pairwise Comparison

Another paradigm to assess text quality is by directly comparing a pair of generated texts based on the same input. This method primarily focuses on the relative quality of the texts. For instance, a prompt for comparing the overall quality of two storylines written according to the same initial story beginning is shown as follows:

====== PROMPT FOR PAIRWISE COMPARISON ======

Consider the following two storylines written according to the given beginning of the story:

The beginning of the story:
`[Conditioned Text]`

Storyline-1:
`[Generated Text-1]`

Storyline-2:
`[Generated Text-2]`

Question: Which storyline is better-written and more consistent with the beginning of the story? Please answer with one of the following options.

Options:
(A) Storyline-1
(B) Storyline-2
(C) Both storylines are equally well-written and consistent with the beginning of the story.

Answer: I will choose Option

3 Experimental Setup

3.1 Tasks and Datasets

We conduct experiments on four distinct natural language generation tasks: Text Summarization, Dialogue Response Generation, Story Generation, and Paraphrase Generation.

Text Summarization aims to summarize the key points of a given long text. SummEval (Fabbri et al., 2021b) is a collection of human annotations for 16 model-generated summaries on 100 CNN/DaliyMail news over 4 dimensions: coherence (COH), fluency (FLU), consistency (CON), and relevance (REL). Due to the budget limit, we randomly sample 20 news and corresponding annotations from SummEval for evaluation.

Dialogue Response Generation aims to generate a response based on the preceding dialogue. We conduct experiments on the dialogue-level FED dataset (Mehri and Eskenazi, 2020a), which contains fine-grained human judgments for 124 conversations. The evaluation aspects include coherence (COH), error recovery (ERR), consistency (CON), diversity (DIV), topic depth (DEP), likeability (LIK), understanding (UND), flexibility (FLE), informativeness (INF), inquisitiveness (INQ) and overall performance (Overall). However, we do not include ERR in our evaluation since some annotations are missing.

Story Generation aims to automatically write a storyline based on a given beginning of the story. We employ OpenMEVA-ROC (Guan et al., 2021) for evaluation, which contains 200 story beginnings and 5 corresponding machine-generated storylines for each beginning. Each storyline is manually annotated in terms of overall quality.

Paraphrase Generation aims to rephrase a sentence in different words or forms while preserving its original meaning. We use Twitter-Para (Xu et al., 2014, 2015) for evaluation, containing 761 input sentences and each input has 9.41 paraphrase candidates on average. We adopt the test set (Shen et al., 2022) extended from Twitter-Para by adding 20% of the input sentences as candidates, denoted as Twitter (Extend).

3.2 Chosen Metrics

Following the settings of previous works, we select baseline metrics from the following widely used metrics accordingly: **ROUGE-1**, **ROUGE-2** and **ROUGE-L** (Lin, 2004); **BERTScore** (Zhang* et al., 2020); **MoverScore** (Zhao et al., 2019); **PRISM** (Thompson and Post, 2020); **BARTScore** and its enhanced versions, **BARTScore+CNN** and **BARTScore+CNN+Para** (Yuan et al., 2021); **BERT-R** (Ghazarian et al., 2019); **GPT-2** (Radford et al., 2019); **USR** (Mehri and Eskenazi, 2020b); **S-DiCoh** (Mesgar et al., 2020); **FED** (Mehri and Eskenazi, 2020a); **DynaEval** (Zhang et al., 2021); **SelfEval** (Ma et al., 2022); **PPL** (Guan et al., 2021); **iBLEU** (Sun and Zhou, 2012); **BERT-iBLEU** (Niu et al., 2021); **ParaScore** (Shen et al., 2022). Note that, Shen et al. (2022) also use a reference-free

Metrics	Spear.			
	COH	FLU	CON	REL
ROUGE-1	21.6	10.5	10.9	42.6
ROUGE-2	30.7	19.1	20.7	36.9
ROUGE-L	17.4	10.2	9.6	40.0
BERTScore	28.5	10.6	13.4	29.5
MoverScore	22.5	11.8	14.6	39.2
PRISM	23.7	17.5	35.2	16.9
BARTScore	33.4	20.9	34.8	24.8
+CNN	43.3	28.7	42.7	36.1
+CNN+Para	40.1	27.2	41.0	32.0
IMPLICIT SCORE				
text-davinci-001	-1.7	-5.6	19.7	8.4
text-davinci-003	**57.4**	**32.9**	35.2	28.1
EXPLICIT SCORE				
ChatGPT (sampling)	45.8	22.1	41.2	39.2
ChatGPT (greedy)	52.2	19.3	**43.3**	**46.0**

Table 1: Sample-level Spearman (Spear.) correlation of different aspects on SummEval.

version of BERTScore and ParaScore, denoted as **BERTScore.Free** and **ParaScore.Free**.

3.3 Meta Evaluation

Individual Score In order to assess the reliability of Individual Scores, we utilize the Spearman (Zar, 2005) and Pearson (Mukaka, 2012) correlation coefficients. As SummEval and OpenMEVA provide an equivalent number of model-generated results for each input, we present the sample-level correlations for these datasets. Whereas, for Twitter (Extend) and the dialog-level FED datasets, we report the dataset-level correlations instead.

Pairwise Comparison To avoid an excessive volume of requests when testing all permutations of pairwise comparisons in each dataset using ChatGPT, we have opted to randomly sample 200 pairs from each dataset as an approximation. To estimate the reliability of metrics for pairwise comparison, Kendall's Tau-b (Kendall, 1945) is employed to evaluate the correlation between two measured variables. A detailed explanation of Kendall's Tau-b is shown in Appendix C.

4 Main Experiments

4.1 Individual Score

Notably, as shown in Tables 1 to 4, even without providing reference or calibration details for different score ranges, ChatGPT's Explicit Score has already correlated with human scores better than most commonly used automated metrics. On Twitter (Extend), it is only outperformed by ParaScore and ParaScore.Free, which requires the use of reference or hyper-parameter adjustments on a dev set. Additionally, the performance of the Explicit Score further improves when we use greedy search instead of Top-P sampling for decoding.

It is worth noting that the Implicit Score based on text-davinci-003 also shows promising results. This suggests that LLMs' confidence level in determining whether a text meets a specific standard (yes or no) can reflect the text's quality to some extent. Besides, the Implicit Score based on text-davinci-003 performs better than that based on text-davinci-001 in most cases, perhaps due to RLHF, allowing text-davinci-003 to provide answers that align with human instructions better.

4.2 Pairwise Comparison

Scoring individual samples without providing detailed criteria for each score range may lead to inconsistent evaluation standards across different samples. Alternatively, we hypothesize that a direct comparison of quality between a pair of samples is more likely to yield reliable evaluation results from ChatGPT. However, our analysis in Tables 5 to 8 suggests that direct pairwise comparison is not as effective as expected, and eliminating the influence of sampling in decoding is not always advantageous for comparison.

We further categorize the texts for comparison into three levels of difficulty, namely hard, medium, and easy, based on the difference in human scores. The larger the score difference between a pair of texts, the easier it is to discern the better one. The performance of various metrics on distinct difficulty levels is shown in Tables 7 and 8. Overall, the metrics exhibit an increasing trend in performance as the difficulty decreases.

Moreover, our investigation indicates that the Implicit Score derived from text-davinci-003 outperforms or performs comparably to the Explicit Score based on ChatGPT when comparing hard text pairs. This finding may be attributed to the higher precision of the Implicit Score, which is based on the model's output token probability (a floating-point number), as opposed to the model's generated Explicit Score, which is limited to integer values ranging from 0 to 100.

Metrics	Spear.									
	COH	CON	DIV	DEP	LIK	UND	FLE	INF	INQ	Overall
BERT-R	22.9	16.3	19.6	19.2	28.1	19.8	25.3	21.1	33.7	24.8
GPT-2	12.3	9.1	14.7	9.7	17.9	7.0	13.4	11.6	7.1	12.3
USR	19.4	16.9	24.2	34.1	22.1	17.2	20.9	28.8	18.8	28.8
S-DiCoh	3.8	1.7	5.9	4.6	-7.0	-10.0	4.4	2.8	-5.4	-7.3
FED	25.1	11.6	44.9	52.2	26.2	30.6	40.8	33.7	29.8	44.3
DynaEval	42.3	35.2	33.2	43.9	39.8	36.1	38.9	39.6	38.8	48.2
SelfEval	43.6	34.7	26.3	32.7	39.0	40.6	31.7	31.8	42.1	43.5
IMPLICIT SCORE										
text-davinci-001	37.9	33.0	36.1	26.2	35.0	57.5	39.5	54.8	45.0	39.4
text-davinci-003	46.8	43.8	24.9	53.4	**57.3**	57.6	45.0	55.1	**59.0**	**58.0**
EXPLICIT SCORE										
ChatGPT (sampling)	57.8	**47.8**	44.5	51.5	47.2	**61.7**	49.4	61.7	42.8	55.8
ChatGPT (greedy)	**62.4**	47.5	**48.3**	**55.5**	55.4	60.0	**54.8**	**62.0**	42.3	54.2

Table 2: Dataset-level Spearman (Spear.) correlation of different aspects on dialogue-level FED.

Metrics	Spear.	Pear.
ROUGE-1	1.4	2.0
ROUGE-2	3.5	4.1
ROUGE-L	1.3	2.1
BERTScore	14.0	12.0
Perplexity	32.4	33.0
BARTScore	-6.5	-8.2
+CNN	4.9	2.6
+CNN+Para	6.4	5.0
IMPLICIT SCORE		
text-davinci-001	30.3	32.9
text-davinci-003	37.9	43.4
EXPLICIT SCORE		
ChatGPT (sampling)	47.6	49.0
ChatGPT (greedy)	**49.9**	**51.7**

Table 3: Sample-level Spearman (Spear.) and Pearson (Pear.) correlation on OpenMEVA.

Metrics	Spear.	Pear.
iBLEU	3.2	1.1
BERTScore	43.2	42.7
BERTScore.Free	41.9	31.6
BARTScore+CNN+Para	27.6	28.0
BERT-iBLEU	41.6	32.7
ParaScore	**53.0**	**52.7**
ParaScore.Free	49.5	49.6
IMPLICIT SCORE		
text-davinci-001	15.8	15.9
text-davinci-003	44.4	40.3
EXPLICIT SCORE		
ChatGPT (sampling)	45.1	44.3
ChatGPT (greedy)	46.5	45.4

Table 4: Dataset-level Spearman (Spear.) and Pearson (Pear.) correlation on Twitter (Extend).

Metrics	Kend.			
	COH	FLU	CON	REL
IMPLICIT SCORE				
text-davinci-001	-3.2	-4.3	9.3	12.9
text-davinci-003	46.9	**24.5**	35.3	29.1
EXPLICIT SCORE				
ChatGPT (sampling)	**50.3**	8.6	31.7	44.3
ChatGPT (greedy)	43.7	16.8	**32.8**	**52.5**
COMPARISON				
ChatGPT (sampling)	22.6	7.8	24.2	30.5
ChatGPT (greedy)	34.5	17.4	22.0	34.0

Table 5: Estimated Kendall's tau-b (Kend.) correlation of different aspects on SummEval.

5 Detailed Analysis

5.1 Why does the pairwise comparison paradigm perform worse?

In the main experiments, it is noteworthy that direct pairwise comparison using ChatGPT did not yield satisfactory results. To investigate whether this was caused by poorly designed prompts, alternative prompts were also evaluated. These prompts are briefly described in Table 9, with detailed information provided in Appendix B. Surprisingly, changing the prompt did not improve performance, but rather worsened it, as illustrated in Figure 1.

To gain further insights, we examined the confusion matrices of results based on different prompts, as shown in Figure 2. Our analysis revealed that, although we have provided the option of "both storylines equally good" in the default prompt (Prompt V1), ChatGPT still tended to choose one storyline that it deemed "better", as observed from Fig-

Metrics	Kend.									
	COH	CON	DIV	DEP	LIK	UND	FLE	INF	INQ	Overall
IMPLICIT SCORE										
text-davinci-001	33.3	32.0	29.6	25.1	25.6	49.9	32.8	44.8	49.5	33.6
text-davinci-003	28.8	30.5	18.8	36.9	41.9	43.2	34.0	45.8	43.0	36.7
EXPLICIT SCORE										
ChatGPT (sampling)	48.4	**44.1**	32.4	47.5	46.7	48.0	36.2	45.6	**45.9**	**44.2**
ChatGPT (greedy)	**50.2**	39.6	**45.5**	**53.5**	**50.8**	**53.7**	50.5	**47.7**	38.1	41.7
COMPARISON										
ChatGPT (sampling)	28.3	16.1	28.5	31.5	43.0	27.5	**55.5**	35.2	24.5	38.6
ChatGPT (greedy)	24.3	13.7	28.5	33.8	41.9	27.5	55.5	34.1	25.6	37.5

Table 6: Estimated Kendall's tau-b (Kend.) correlation of different aspects on dialogue-level FED.

Metrics	Kend.			
	Hard	Medium	Easy	All
IMPLICIT SCORE				
text-davinci-001	6.3	29.8	44.4	16.6
text-davinci-003	**27.9**	36.8	66.7	33.2
EXPLICIT SCORE				
ChatGPT (sampling)	18.5	47.3	74.3	31.2
ChatGPT (greedy)	16.8	**62.6**	**82.5**	**36.2**
COMPARISON				
ChatGPT (sampling)	8.1	22.8	33.3	14.5
ChatGPT (greedy)	9.9	29.8	55.6	19.7

Table 7: Estimated Kendall's tau-b (Kend.) correlation on OpenMEVA.

Metrics	Kend.			
	Hard	Medium	Easy	All
IMPLICIT SCORE				
text-davinci-001	21.6	34.6	13.6	20.4
text-davinci-003	25.5	19.2	59.1	28.6
EXPLICIT SCORE				
ChatGPT (sampling)	**27.8**	**40.0**	53.8	**34.9**
ChatGPT (greedy)	15.3	38.5	57.0	31.2
COMPARISON				
ChatGPT (sampling)	14.6	31.0	**68.3**	31.3
ChatGPT (greedy)	10.0	22.2	65.1	26.3

Table 8: Estimated Kendall's tau-b (Kend.) correlation on Twitter (Extend).

ure 2(a). This could be attributed to the bias introduced by adding "Answer: I will choose Option" at the end of the prompt, which may have induced the model to make a biased choice at the beginning of the answer. To address this issue, we modified the prompt to require ChatGPT to present its reasoning process before making the final decision

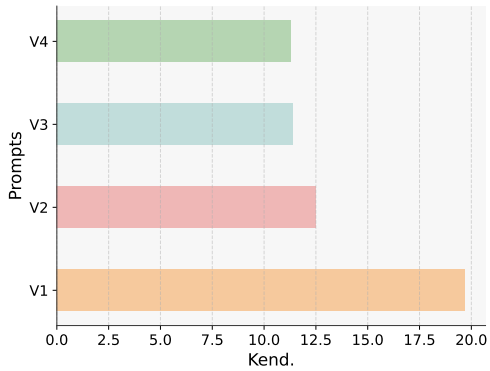


Figure 1: Estimated Kendall's tau-b (Kend.) correlation of Pairwise Comparison using ChatGPT with different prompts on OpenMEVA. We use greedy decoding for Prompt V1∼V3. Whereas, for Prompt V4 we use Top-P sampling five times to obtain multiple results and vote for the final decision.

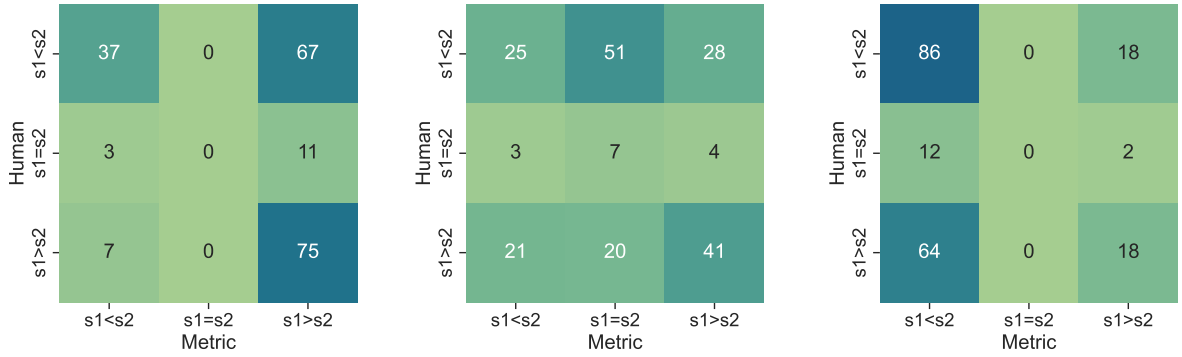
(Prompt V4). With this prompt, the model was more likely to choose the "tie" option, as indicated by the "s1=s2" column in Figure 2(b).

After analyzing ChatGPT's reasoning process, we discovered that ChatGPT frequently concludes that "the quality of the two given storylines is equally poor." As a result, we prompted ChatGPT to choose the "worse" storyline instead of the "better" one (Prompt V3). However, this questioning approach did not yield a better outcome. In addition, Figure 2(c) shows that although Prompt V3 is a mirrored version of Prompt V1, which changes the prompt from selecting the better option to choosing the worse one, ChatGPT's results based on these two prompts are not always consistent. For example, in one case, ChatGPT selected Storyline-1 as better based on Prompt V1, but under the guidance of Prompt V3, it may not necessarily choose Storyline-2 as worse.

Overall, we speculate that the poor quality of the

Prompts for Pairwise Comparison on Story Generation	
Prompt V1	The default prompt where we first provide the beginning of the story and the corresponding two storylines for comparison before presenting the question.
Prompt V2	A revised version of Prompt V1 where we first propose the question, then provide the beginning of the story and present the two storylines to be compared in the form of options.
Prompt V3	A mirrored version of Prompt V1 where we instruct the model to choose "which one is worse" instead of "which one is better" from the two given storylines.
Prompt V4	A "chain-of-thought" version of Prompt V1 where we require the model to illustrate the reasoning process before presenting the final answer.
Prompts for Explicit Score on Story Generation	
Prompt V1	The default prompt where we only specify the rating criteria for zero and full marks.
Prompt V2	A rephrased version of Prompt V1.
Prompt V3	A simplified version of Prompt V1 where we only describe the dimensions that need to be evaluated.
Prompt V4	A detailed prompt where we divide the scores into 5 scales and list the corresponding evaluation criteria for each score scale.
Prompt V5	A "chain-of-thought" version of Prompt V1 where we require the model to first present the reasons for the evaluation, and then provide the final score.

Table 9: Prompts designed for Pairwise Comparison and Explicit Score for assessing the quality of storylines in story generation. Note that Prompt V4 of Explicit Score is cited from (Wang et al., 2023a).



(a) Prompt V1 (Default) (b) Prompt V4 ("Chain-of-Thought") (c) Prompt V3 (Mirrored)

Figure 2: Confusion matrices of pairwise comparisons on OpenMEVA based on different prompts using ChatGPT. Prompt V1 is the default prompt used in the main experiments. Prompt V4 and V3 are the "chain-of-thought" and "mirrored" versions of Prompt V1 respectively. Details of these prompts are presented in Table 9 and Appendix B.

candidate texts used in our experiments is the main reason why comparing pairs directly with ChatGPT did not yield good results. ChatGPT perceives the candidate texts as generally low quality, making it to select a "better" or "worse" one from them. This might lead to ChatGPT's unstable decisions.

5.2 Why does Explicit Score generally perform better than Implicit Score?

In order to obtain the Explicit Score, we utilize ChatGPT to generate scores in a natural language format. However, as we do not have access to ChatGPT's token probabilities, we instead rely on the confidence of text-davinci series models to determine the Implicit Score, which reflects how well a text meets a particular evaluation criterion. As stated in the Main Experiments (§4), the Explicit

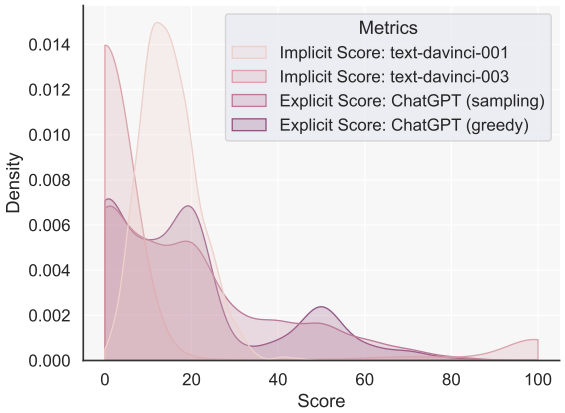


Figure 3: Distribution of different types of Individual Scores on OpenMEVA. The Implicit Score is rescaled into [0,100].

Explicit Score	Spear.	Pear.
ChatGPT		
w/ Prompt V1 (greedy)	49.9	51.7
w/ Prompt V2 (greedy)	**50.8**	**53.6**
w/ Prompt V3 (greedy)	49.4	52.0
w/ Prompt V4 (greedy)	46.1	48.4
w/ Prompt V5 (sampling)	47.2	50.8

Table 10: Sample-level Spearman (Spear.) and Pearson (Pear.) correlation for Explicit Score based on ChatGPT with different prompts on OpenMEVA. We use greedy decoding for Prompt V1∼V4. Whereas, for Prompt V5, we employ Top-P sampling five times to generate multiple reasons and average the resulting scores.

Score is generally more effective than the Implicit Score. This difference in effectiveness could be attributed not only to the variation in the models used but also to the distribution of the two scores. Figure 3 illustrates that the Implicit Score distribution has a peaked structure and is concentrated within a small range. In contrast, the Explicit Score distribution is smoother, allowing for greater discrimination between scores for different texts.

5.3 How does the prompt design affect Explicit Score?

We also investigate the impact of prompt design on the performance of rating Explicit Scores generated by ChatGPT. The detailed prompts are provided in Appendix A, and their main features and differences are summarized in Table 9. Our results, presented in Table 10, indicate that paraphrasing (V2) or simplifying (V3) the default prompt (V1) does not significantly affect the performance of Explicit Score based on ChatGPT. In contrast, refining scoring criteria (V4) or providing reasons before scoring (V5) results in a slight decrease in performance. The former may be due to the fact that the refined scoring rules in Prompt V4 do not fully match the standards used for actual manual annotation, and dividing scores into five scales reduces the distinction between scores for different samples. The latter may be due to the overall low quality of the dataset. Our observation indicates that ChatGPT's evaluations for each text are similar and mostly negative. After giving reasons before scoring, ChatGPT's scoring focuses more on the reasons rather than the text itself, resulting in lower scores for each text based on Prompt V5 and reducing the distinction between scores. The detailed distribution of scores derived from different prompts is demonstrated using a violin plot in Figure 4.

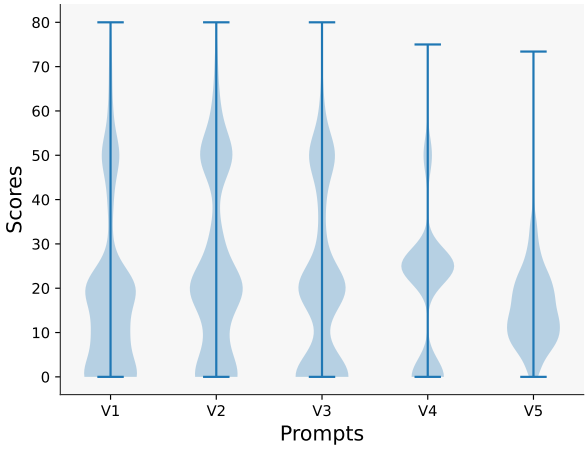


Figure 4: Distribution of Explicit Scores based on ChatGPT with different prompts on OpenMEVA. For Prompt V4, the scores are normalized into [0, 100].

6 Related Work

In the field of text quality evaluation, researchers have devised two main lines of approaches: reference-based and reference-free methods. The reference-based text evaluation aims to assess the quality by comparing outputs with ground truth, e.g. ROUGE (Lin, 2004), BERTScore (Zhang* et al., 2020) and BARTScore (Yuan et al., 2021). However, due to the inherent complexity and diversity of text, it is impossible to obtain references covering the entire spectrum of potential outputs. This limitation has prompted researchers to explore reference-free evaluation methods without relying on predefined references e.g. iBLEU (Sun and Zhou, 2012) and ParaScore (Shen et al., 2022). In this line, a reliable sentence representation model is required (Gao et al., 2021; Shen et al., 2023a,b). Recent studies have indicated that LLM-based evaluation methods can exhibit good consistency with human evaluation in assessing text quality (Fu et al., 2023; Wang et al., 2023a; Kocmi and Federmann, 2023; Ji et al., 2023). However, most of these works are preliminary explorations or require gold references. On the contrary, we are the first to conduct extensive experiments to investigate the optimal evaluation approaches using LLMs without references, and moreover propose some clues for customized text evaluation.

7 Conclusion

This paper explores the feasibility of LLMs, specifically ChatGPT and text-davinci series models, for evaluating text quality in a reference-free mode. Through an empirical study, we compare different

methods for the evaluation of text quality and recommend the use of an Explicit Score generated by ChatGPT as the most effective and stable approach. This paper also highlights the potential problem of directly comparing the quality of two texts using ChatGPT and the limitations of Implicit Scores obtained through the confidence of text-davinci series models. The prompt design is another crucial factor impacting the performance of the Explicit Score generated by ChatGPT. Overall, this paper demonstrates the potential of LLMs in evaluating text quality without reference and we hope it will provide useful insights for future research.

Limitations

- **Meta Evaluation Strategy**

We primarily assess the reliability of metrics based on their correlation with human scores. However, it should be noted that the consistency between scores annotated by different raters may not always be high in certain datasets. Hence, the correlation with human ratings may not always reflect the performance of metrics appropriately.

- **Coverage of Texts**

We only conducted experiments on four text-generation tasks. Additionally, the quality distribution of the evaluated texts may be non-uniform, potentially lacking in extremely high-quality texts. Even if a metric performs well in evaluating a set of low-quality texts, it does not necessarily imply the same level of discrimination for high-quality texts, and vice versa. Furthermore, our evaluation has been limited to short texts, omitting the consideration of long-text generation.

- **Coverage of Models**

We utilize OpenAI's API to access their language models, including ChatGPT (gpt3.5-turbo-0301), text-davinci-003, and text-davinci-001. However, these models may be updated over time, which can result in inconsistencies in experimental outcomes. Moreover, we have not considered a wider range of LLMs, such as text-babbage-001, text-curie-001, and the FLAN-T5 series. Regrettably, due to API limitations, we were unable to obtain results from the more powerful GPT4 model.

- **Prompt Design**

Our exploration of prompts was limited to a few basic variations. Future research may benefit from more sophisticated prompt designs, such as incorporating few-shot demonstrations, providing more precise annotation guidelines, or guiding the model through multi-turn conversations to facilitate a more accurate assessment of text quality.

Acknowledgements

This research was supported in part by the National Natural Science Foundation of China(62006062, 62176076), the Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies(2022B121201000 5), Natural Science Foundation of Guangdong(2023A1515012922), and Key Technologies Research and Development Program of Shenzhen JSGG20210802154400001.

References

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.

Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Ziyang Chen, and Jia Li. 2022a. What would harry say? building dialogue agents for characters in a story. *arXiv preprint arXiv:2211.06869*.

Yi Chen, Haiyun Jiang, Lemao Liu, Rui Wang, Shuming Shi, and Ruifeng Xu. 2022b. Mcpg: A flexible multi-level controllable framework for unsupervised paraphrase generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5948–5958.

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

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, Dublin, Ireland. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021a. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021b. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Sarik Ghazarian, Johnny Wei, Aram Galstyan, and Nanyun Peng. 2019. Better automatic evaluation of open-domain dialogue systems with contextualized embeddings. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 82–89, Minneapolis, Minnesota. Association for Computational Linguistics.

Jian Guan, Zhexin Zhang, Zhuoer Feng, Zitao Liu, Wenbiao Ding, Xiaoxi Mao, Changjie Fan, and Minlie Huang. 2021. OpenMEVA: A benchmark for evaluating open-ended story generation metrics. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6394–6407, Online. Association for Computational Linguistics.

Yunjie Ji, Yan Gong, Yiping Peng, Chao Ni, Peiyan Sun, Dongyu Pan, Baochang Ma, and Xiangang Li. 2023. Exploring chatgpt's ability to rank content: A preliminary study on consistency with human preferences.

M. G. Kendall. 1945. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Longxuan Ma, Ziyu Zhuang, Weinan Zhang, Mingda Li, and Ting Liu. 2022. SelF-eval: Self-supervised fine-grained dialogue evaluation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 485–495, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Shikib Mehri and Maxine Eskenazi. 2020a. Unsupervised evaluation of interactive dialog with DialoGPT. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235, 1st virtual meeting. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskenazi. 2020b. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online. Association for Computational Linguistics.

Mohsen Mesgar, Sebastian Bücker, and Iryna Gurevych. 2020. Dialogue coherence assessment without explicit dialogue act labels. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1439–1450, Online. Association for Computational Linguistics.

Mavuto M Mukaka. 2012. A guide to appropriate use of correlation coefficient in medical research. *Malawi medical journal*, 24(3):69–71.

Tong Niu, Semih Yavuz, Yingbo Zhou, Nitish Shirish Keskar, Huan Wang, and Caiming Xiong. 2021. Unsupervised paraphrasing with pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5136–5150, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

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.

Lingfeng Shen, Haiyun Jiang, Lemao Liu, and Shuming Shi. 2023a. Sen2pro: A probabilistic perspective to sentence embedding from pre-trained language model. *arXiv preprint arXiv:2306.02247*.

Lingfeng Shen, Haiyun Jiang, Lemao Liu, and Shuming Shi. 2023b. A simple and plug-and-play method for unsupervised sentence representation enhancement. *arXiv preprint arXiv:2305.07824*.

Lingfeng Shen, Lemao Liu, Haiyun Jiang, and Shuming Shi. 2022. On the evaluation metrics for paraphrase generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3178–3190, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mingyang Song, Haiyun Jiang, Shuming Shi, Songfang Yao, Shilong Lu, Yi Feng, Huafeng Liu, and Liping Jing. 2023. Is chatgpt a good keyphrase generator? a preliminary study. *arXiv preprint arXiv:2303.13001*.

Hong Sun and Ming Zhou. 2012. Joint learning of a dual SMT system for paraphrase generation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 38–42, Jeju Island, Korea. Association for Computational Linguistics.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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

Rui Wang, Jianzhu Bao, Fei Mi, Yi Chen, Hongru Wang, Yasheng Wang, Yitong Li, Lifeng Shang, Kam-Fai Wong, and Ruifeng Xu. 2023b. Retrieval-free knowledge injection through multi-document traversal for dialogue models. In *Annual Meeting of the Association for Computational Linguistics*.

Wei Xu, Chris Callison-Burch, and Bill Dolan. 2015. SemEval-2015 task 1: Paraphrase and semantic similarity in Twitter (PIT). In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 1–11, Denver, Colorado. Association for Computational Linguistics.

Wei Xu, Alan Ritter, Chris Callison-Burch, William B. Dolan, and Yangfeng Ji. 2014. Extracting lexically divergent paraphrases from Twitter. *Transactions of the Association for Computational Linguistics*, 2:435–448.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Jerrold H Zar. 2005. Spearman rank correlation. *Encyclopedia of biostatistics*, 7.

Chen Zhang, Yiming Chen, Luis Fernando D'Haro, Yan Zhang, Thomas Friedrichs, Grandee Lee, and Haizhou Li. 2021. DynaEval: Unifying turn and dialogue level evaluation. 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 5676–5689, Online. Association for Computational Linguistics.

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

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

A Different Prompts for Explicit Score on Story Generation

======= PROMPT FOR EXPLICIT SCORE V1 =======

Score the following storyline given the beginning of the story on a continual scale from 0 (worst) to 100 (best), where a score of 0 means "The storyline makes no sense and is totally not understandable" and a score of 100 means "The storyline is perfect-written and highly consistent with the given beginning of the story".

The beginning of the story:
`[Conditioned Text]`

Storyline:
`[Generated Text]`

Score:

======= PROMPT FOR EXPLICIT SCORE V2 =======

On a scale of 0 to 100, evaluate the storyline based on the given beginning. A score of 0 indicates that the storyline is incomprehensible, while a score of 100 means that the storyline is flawlessly written and logically follows from the beginning of the story.

The beginning of the story:
`[Conditioned Text]`

Storyline:
`[Generated Text]`

Score:

======= PROMPT FOR EXPLICIT SCORE V3 =======

Score the overall quality of the following storyline given the beginning of the story on a continual scale from 0 (worst) to 100 (best). Consider whether the storyline is well-written and consistent with the given beginning of the story.

The beginning of the story:
`[Conditioned Text]`

Storyline:
`[Generated Text]`

Score:

B Different Prompts for Pairwise Comparison on Story Generation

======= PROMPT FOR EXPLICIT SCORE V4 =======

Score the following storyline given the beginning of the story with one to five stars. Where

- one star means "Nonsense",
- two stars mean "The storyline has some connections with the beginning, but is not understandable",
- three stars mean "The storyline has some connections with the beginning and is understandable",
- four stars mean "The storyline is consistent with the beginning and possibly involves a few grammar mistakes",
- and five stars mean "Perfect storyline and grammar".

The beginning of the story:
`[Conditioned Text]`

Storyline:
`[Generated Text]`

Stars (1-5):

======= PROMPT FOR EXPLICIT SCORE V5 =======

Score the following storyline given the beginning of the story on a continual scale from 0 (worst) to 100 (best), where a score of 0 means "The storyline makes no sense and is totally not understandable" and a score of 100 means "The storyline is perfect-written and highly consistent with the given beginning of the story". Please first give your reason carefully (indicated by "Reason:") and then decide your final score (indicated by "Score: 1-100").

The beginning of the story:
`[Conditioned Text]`

Storyline:
`[Generated Text]`

===== PROMPT FOR PAIRWISE COMPARISON V1 =====

Consider the following two storylines written according to the given beginning of the story:

The beginning of the story:
`[Conditioned Text]`

Storyline-1:
`[Generated Text-1]`

Storyline-2:
`[Generated Text-2]`

Question: Which storyline is better-written and more consistent with the beginning of the story? Please answer with one of the following options.

Options:
(A) Storyline-1
(B) Storyline-2
(C) Both storylines are equally well-written and consistent with the beginning of the story.

Answer: I will choose Option

===== PROMPT FOR PAIRWISE COMPARISON V2 =====

Question: Which storyline is better-written and more consistent with the beginning of the story? Please answer with one of the following options.

The beginning of the story:
`[Conditioned Text]`

Options:
(A) `[Generated Text-1]`
(B) `[Generated Text-2]`
(C) Both storylines are equally well-written and consistent with the beginning of the story.

Answer: I will choose Option

===== PROMPT FOR PAIRWISE COMPARISON V3 =====

Consider the following two storylines written according to the given beginning of the story:
The beginning of the story:
`[Conditioned Text]`

Storyline-1:
`[Generated Text-1]`

Storyline-2:
`[Generated Text-2]`

Question: Which storyline has poorer writing and is less consistent with the beginning of the story? Please answer with one of the following options.

Options:
(A) Storyline-1
(B) Storyline-2
(C) Both storylines are equally poor-written and inconsistent with the beginning of the story.

Answer: I will choose Option

===== PROMPT FOR PAIRWISE COMPARISON V4 =====

Consider the following two storylines written according to the given beginning of the story:
The beginning of the story:
`[Conditioned Text]`

Storyline-1:
`[Generated Text-1]`

Storyline-2:
`[Generated Text-2]`

Question: Which storyline is better-written and more consistent with the beginning of the story? Please first give your reason carefully (indicated by "Reason:") and then choose one of the following options (indicated by "Answer: A/B/C").

Options:
(A) Storyline-1
(B) Storyline-2
(C) Both storylines are equally well-written (poor-written) and consistent (inconsistent) with the beginning of the story.

C An Explanation of Kendall's Tall-b

Kendall's Tau-b is a measure of the correlation between two variables, specifically designed to handle ties and ranks. The formula to calculate Kendall's Tau-b is as follows:

$$\tau_B = \frac{P - Q}{\sqrt{(P + Q + T)(P + Q + U)}}. \quad (2)$$

where P is the number of concordant pairs, Q is the number of discordant pairs, T is the number of ties only in human judgments, and U is the number of ties only in the given metric. To better understand the calculation of P, Q, T, and U, we can refer to the following table:

		Metric		
		$s_1 < s_2$	$s_1 = s_2$	$s_1 > s_2$
Human	$s_1 < s_2$	P	U	Q
	$s_1 = s_2$	T	-	T
	$s_1 > s_2$	Q	U	P