Is ChatGPT a Good NLG Evaluator? A Preliminary Study

Jiaan Wang[1]*†, Yunlong Liang[2]*, Fandong Meng[3]
Zengkui Sun[2], Haoxiang Shi[4], Zhixu Li[5], Jinan Xu[2], Jianfeng Qu[1] and Jie Zhou[3]
[1]Soochow University, Suzhou, China [2]Beijing Jiaotong University, Beijing, China
[3]Pattern Recognition Center, WeChat AI, Tencent Inc, China [4]Waseda University, Tokyo, Japan
[5]Fudan Unversity, Shanghai, China

jawang.nlp@gmail.com {yunlongliang, zengksun}@bjtu.edu.cn
fandongmeng@tencent.com hollis.shi@toki.waseda.jp

arXiv:2303.04048v3 [cs.CL] 24 Oct 2023

Abstract

Recently, the emergence of ChatGPT has attracted wide attention from the computational linguistics community. Many prior studies have shown that ChatGPT achieves remarkable performance on various NLP tasks in terms of automatic evaluation metrics. However, the ability of ChatGPT to serve as an evaluation metric is still underexplored. Considering assessing the quality of natural language generation (NLG) models is an arduous task and NLG metrics notoriously show their poor correlation with human judgments, we wonder *whether ChatGPT is a good NLG evaluation metric.*

In this report, we provide a preliminary meta-evaluation on ChatGPT to show its reliability as an NLG metric. In detail, we regard ChatGPT as a human evaluator and give task-specific (*e.g.*, summarization) and aspect-specific (*e.g.*, relevance) instruction to prompt ChatGPT to evaluate the generated results of NLG models. We conduct experiments on five NLG meta-evaluation datasets (including summarization, story generation and data-to-text tasks). Experimental results show that compared with previous automatic metrics, ChatGPT achieves state-of-the-art or competitive correlation with human judgments in most cases. In addition, we find that the effectiveness of the ChatGPT evaluator might be influenced by the creation method of the meta-evaluation datasets. For the meta-evaluation datasets which are created greatly depending on the reference and thus are biased, the ChatGPT evaluator might lose its effectiveness. We hope our preliminary study could prompt the emergence of a general-purposed reliable NLG metric.[1]

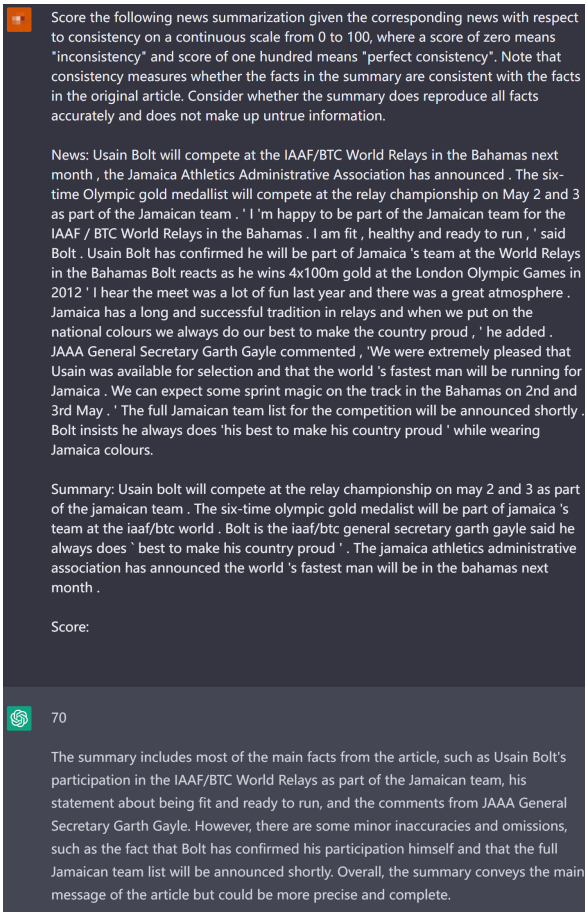

Figure 1: Prompting ChatGPT as an evaluator to score the generated results of NLG models (taking news summarization as an example).

1 Introduction

Pre-trained large language models (LLMs; *e.g.*, GPT-3.5, ChatGPT and GPT-4), which are performed through chatting (or asking) with it, have obtained promising results on various natural language understanding (NLU) and natural language generation (NLG) downstream tasks (Ouyang et al., 2022; Kocoń et al., 2023; Qin et al., 2023; Huang et al., 2023; Yang et al., 2023; Rao et al., 2023; Bang et al., 2023; Zuccon and Koopman, 2023). For example, Zhong et al. (2023) show that ChatGPT can attain the comparable understanding abil-

* Equal Contribution. Work was done when Wang and Liang was interning at Pattern Recognition Center, WeChat AI, Tencent Inc, China.

† Corresponding author.

[1]We have released the used data at https://github.com/krystalan/chatgpt_as_nlg_evaluator.

ity to some fine-tuned BERT-style models on NLU tasks while failing to surpass current task-specific NLU models. Wei et al. (2023) prove that ChatGPT can achieve good performance and even surpasses some full-shot models on several datasets through a multi-turn question-answering manner. For NLG tasks, Jiao et al. (2023) claim that ChatGPT performs competitively with commercial translation products (*e.g.*, Google Translator) on high-resource European languages. Wang et al. (2023a) demonstrate that ChatGPT can balance well between informativeness and conciseness, and generate great cross-lingual summaries. Although impressive performance on these tasks in terms of automatic evaluation metrics has been shown, it is still not clear *whether ChatGPT can evaluate the quality of textual generations as a human does*.

Recently, using pre-trained language models as NLG evaluation metric, *e.g.*, MoverScore (Zhao et al., 2019), BERTScore (Zhang et al., 2020), COMET (Rei et al., 2020), BLEURT (Sellam et al., 2020), BARTScore (Yuan et al., 2021) and MAUVE (Pillutla et al., 2022), receives increasing attention since it offers a decent human-related judgment from a deep semantic perspective. Given the powerful ability of ChatGPT as an intelligent conversational LLM, researchers also attempt to investigate whether it can evaluate the translation quality as a human evaluator (Kocmi and Federmann, 2023). However, the automated assessment of the general generation quality of NLG models still remains underexplored.

In this report, we aim to answer the following research question: *Is ChatGPT a good NLG evaluator?* To this end, we regard ChatGPT as a human evaluator and give task-specific (*e.g.*, summarization) and aspect-specific (*e.g.*, relevance) instruction to prompt ChatGPT to evaluate the generation of NLG models. As the example shows in Figure 1, we also attempt different scoring criteria and whether to provide golden references in the prompts to systematically test the reliability of the ChatGPT evaluator. We conduct experiments on five widely-used NLG meta-evaluation datasets (including summarization, story generation and data-to-text tasks). Experimental results show that ChatGPT exhibits a high correlation with human judgment in most cases especially for the story generation task, indicating its potential as an NLG metric. In addition, we find that the ChatGPT evaluator is sensitive to the prompts, and for different tasks or aspects, the prompts should be carefully designed. Moreover, the creation method of the meta-evaluation datasets has a significant influence on the effectiveness of different evaluation metrics. If a meta-evaluation dataset is created greatly depending on the reference, the similarity between model generation and references serves as a strong signal to reflect human judgments, where simple similarity-based metrics (*e.g.*, ROUGE) can achieve very strong performance. Therefore, the ChatGPT evaluator might lose its effectiveness in such situations.

Our main contributions are concluded as follows:

- To our knowledge, we are the first to utilize ChatGPT as a general NLG evaluation metric to study its correlations with human judgments.
- We use task-specific and aspect-specific prompts to guide ChatGPT to perform as a reference-free or reference-based NLG metric, and evaluate its effectiveness on five widely-used meta-evaluation datasets covering three NLG tasks.
- We find that the ChatGPT evaluator has a high correlation with humans in most cases, especially for creative NLG tasks (*e.g.*, story generation) where multiple generations can satisfy humans.
- We find that the ChatGPT evaluator is sensitive to the prompts. For different tasks and aspects, the prompt should be carefully designed.
- We find that the involved biases of the NLG meta-evaluation datasets also influence the effectiveness of NLG metrics, and might lead to the limited effectiveness of the ChatGPT evaluator.

2 Related Work

2.1 NLG Metrics

A good automatic NLG metric can effectively indicate the quality of the textual generations and thus can save lots of human labor from conducting human evaluation. Therefore, it is vital to design automatic evaluation metrics for NLG tasks, *e.g.*, text summarization, story generation, data-to-text generation, machine translation, and many others. Generally, the score that points out how well the systems perform on each task is computed by comparing the system texts with one or more reference texts for semantic matching. In the literature, the metrics can be roughly categorized into four types:

n-gram-based Metrics. Essentially, the n-gram-based metrics aim to measure the lexical overlap between a generated text and a reference text.

The standard n-gram overlap-based metrics generally include ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), Distinct-n (Li et al., 2016), and METEOR (Denkowski and Lavie, 2011). For example, ROUGE is the dominant metric in the summarization evaluation area. Its variants consider the overlap of unigrams (ROUGE-1) and bigrams (ROUGE-2), among others. BLEU metric is the common practice for the machine translation evaluation area. Although these metrics achieve good correlations (typically large overlaps) with golden references, they are not general enough because a system summary might convey the same meaning while using different surface forms.

Embedding-based Metrics. To further improve semantic similarity between a generated text and a reference text, embedding-based metrics are proposed based on the word embeddings (*e.g.*, WMD (Kusner et al., 2015)) or sentence embeddings (*e.g.*, BERTScore (Zhang et al., 2020) and MoverScore (Zhao et al., 2019)). These metrics further bridge the gap with human-related judgment while they largely depend on the quality of embeddings, which may limit their potential.

LLM-based Metrics. With the development of LLMs, some researchers show that LLMs could achieve great correlation with human judgment, *i.e.*, BARTScore (Yuan et al., 2021), and GPTScore (Fu et al., 2023). However, ChatGPT, as a more powerful conversational LLM, has not been investigated to evaluate the quality of the NLG model outputs.

Other Metrics. In different research fields, there are some paraphraser-based or task-specific metrics. For example, PRISM (Thompson and Post, 2020) is proposed to evaluate translation outputs based on the pre-trained paraphrase models. StoryER (Chen et al., 2022), a learning metric, mimics human preference when judging a story by three steps: Ranking, Rating, and Reasoning based on a specific story-generation dataset. Besides, a specifically developed metric named PARENT (Dhingra et al., 2019) is designed for the table2text generation. Other statistical indicators, such as omission errors, hallucination errors, addition errors, duplication errors, and extrinsic errors, are also applied in the table2text task. Although these metrics have obtained impressive results, human evaluation is still inevitable in table2text.

2.2 Research on ChatGPT

In recent years, from BERT (Devlin et al., 2019) to ChatGPT (OpenAI, 2022), a large number of pre-trained language models have been proposed one after another. Both their parameters and ability are gradually increased, facilitating much-advanced techniques. In particular, ChatGPT, which shows us a revolutionary change as an intelligent conversational large language model, sends shock waves through the research community and industries that have continued to reverberate to this day. With the emergence of ChatGPT, there are two growing research interests related to it: (1) leveraging ChatGPT to deal with various NLP tasks and evaluating its performance using traditional task-specific metrics (*i.e.*, evaluation), and (2) using as a metric to evaluate the outputs of other task-specific models (*i.e.*, evaluator) (Kocmi and Federmann, 2023).

Evaluation. Generally, the evaluation tasks on ChatGPT can be divided into two categories, *i.e.*, natural language understanding (NLU) and natural language generation (NLG). For NLU tasks, some researchers find that ChatGPT covers almost all NLU tasks (*e.g.*, sentiment analysis, textual similarity and textual entailment) and achieves competitive or even better performance (Qin et al., 2023; Bang et al., 2023; Zhong et al., 2023). For NLG tasks, machine translation (Jiao et al., 2023), summarization (Yang et al., 2023), query generation (Wang et al., 2023b), and radiology report simplification (Jeblick et al., 2022) are involved. Different from them, we regard ChatGPT as a human evaluator to automatically assess the quality of general textual generations rather than using it for solving tasks.

Evaluator. As an evaluator, there are two studies that evaluate the quality of translation (Kocmi and Federmann, 2023) and human personalities (Rao et al., 2023) by prompting ChatGPT. However, in this work, we aim to evaluate the more general textual outputs to further show the ability of ChatGPT as a general NLG metric.

3 ChatGPT for NLG Evaluation

In this section, we discuss how to prompt ChatGPT to serve as a reference-free NLG metric (§ 3.1) or a reference-based NLG metric (§ 3.2) to evaluate the generation quality of NLG models. We take the news summarization task as an example, and give the details of the prompt templates.

3.1 Reference-free Metric

To evaluate the generation quality of NLG models, we regard ChatGPT as a human evaluator and give it evaluation instruction via different prompts. Each prompt should specify (1) which NLG task (*e.g.*, summarization) needs to be evaluated and (2) which aspect (*e.g.*, fluency) of the generation result should be assessed currently.

Inspired by Kocmi and Federmann (2023), we utilize the following two prompts: direct assessment (**DA**) and one-to-five stars ranking (**star**).

(DA Prompt)
Score the following `[task-ins]` with respect to `[aspect]` on a continuous scale from 0 to 100, where a score of zero means "`[ant-aspect]`" and score of one hundred means "perfect `[aspect]`". Note that `[aspect]` measures `[aspect-ins]`.

`[Conditioned Text]`
`[Generated Text]`
Scores:

(Star Prompt)
Score the following `[task-ins]` with respect to `[aspect]` with one to five stars, where one star means "`[ant-aspect]`" and five stars means "perfect `[aspect]`". Note that `[aspect]` measures `[aspect-ins]`.

`[Conditioned Text]`
`[Generated Text]`
Stars:

where `[task-ins]` and `[aspect-ins]` are the instructions of the current task and aspect, respectively. `[aspect]` and `[ant-aspect]` denote the evaluated aspect and its antonym, respectively. `[Conditioned Text]` is the input of NLG models while `[Generated Text]` is the output. For example, when evaluating news summarization models in terms of fluency, the DA prompt may be like this:

Score the following news summarization given the corresponding news with respect to fluency on a continuous scale from 0 to 100, where a score of zero means "disfluency" and score of one hundred means "perfect fluency". Note that fluency measures the quality of individual sentences, are they well-written and grammatically correct. Consider the quality of individual sentences.

News: [a news article]
Summary: [one generated summary]
Scores:

In this manner, both the details of the task and the evaluation aspect are given to ChatGPT. Next, ChatGPT will give its judgment (*e.g.*, "score: 70") and the corresponding illustrative description (*e.g.*, "the summary covers the main points of the news, but ..."). A specific example is shown in Figure 1. Finally, the numerical scores could be extracted via several simple heuristic rules.

3.2 Reference-based Metric

In addition to reference-free metrics, we explicitly mention the golden references in the prompts to make ChatGPT a reference-based NLG metric:

(DA Prompt w/ Reference)
Score the following `[task-ins]` with respect to `[aspect]` on a continuous scale from 0 to 100, where a score of zero means "`[ant-aspect]`" and score of one hundred means "perfect `[aspect]`". Note that `[aspect]` measures `[aspect-ins]`.

`[Conditioned Text]`
Human reference: `[A Reference]`
`[Generated Text]`
Scores:

The star prompt with reference is also formed in a similar way:

(Star Prompt w/ Reference))
Score the following `[task-ins]` with respect to `[aspect]` with one to five stars, where one star means "`[ant-aspect]`" and five stars means "perfect `[aspect]`". Note that `[aspect]` measures `[aspect-ins]`.

`[Conditioned Text]`
Human reference: `[A Reference]`
`[Generated Text]`
Stars:

In this way, the ChatGPT evaluator will make its judgment and give the evaluation results under the consideration of the golden references.

Metrics	Coherence			Relevance			Consistency			Fluency			Avg.		
	Spear.	Pear.	Kend.	Spear.	Pear.	Kend.	Spear.	Pear.	Kend.	Spear.	Pear.	Kend.	Spear.	Pear.	Kend.
ROUGE-1	0.167	0.160	0.126	0.326	0.359	0.252	0.160	0.224	0.130	0.115	0.158	0.094	0.192	0.225	0.150
ROUGE-2	0.184	0.174	0.139	0.290	0.327	0.219	0.187	0.246	0.155	0.159	0.185	0.128	0.205	0.233	0.160
ROUGE-L	0.128	0.102	0.099	0.311	0.342	0.237	0.115	0.189	0.092	0.105	0.141	0.084	0.165	0.194	0.128
BERTScore	0.283	0.310	0.211	0.311	0.346	0.243	0.110	0.152	0.090	0.192	0.209	0.158	0.224	0.254	0.175
MoverScore	0.159	0.167	0.118	0.318	0.371	0.244	0.157	0.224	0.127	0.129	0.176	0.105	0.191	0.234	0.148
PRISM	0.249	0.258	0.196	0.212	0.232	0.163	0.345	0.352	0.285	0.254	0.264	0.205	0.265	0.276	0.212
BARTScore	0.322	0.345	0.250	0.264	0.290	0.197	0.311	0.321	0.256	0.248	0.260	0.203	0.286	0.304	0.227
BARTScore+CNN	0.448	0.458	0.342	0.356	0.369	0.273	0.382	0.422	0.315	0.356	0.407	0.292	0.385	0.414	0.305
BARTScore+CNN+Para	0.424	0.442	0.325	0.313	0.364	0.241	0.401	0.487	0.332	0.378	**0.448**	0.311	0.379	0.435	0.302
ChatGPT (DA w/o ref)	0.451	0.456	0.383	0.439	0.473	**0.379**	**0.432**	0.512	**0.399**	**0.380**	0.443	**0.351**	**0.425**	**0.471**	**0.378**
ChatGPT (Stars w/o ref)	0.470	0.484	0.403	0.428	0.454	0.374	0.419	0.517	0.389	0.353	0.415	0.329	0.417	0.468	0.374
ChatGPT (DA w/ ref)	0.420	0.435	0.346	**0.448**	**0.474**	0.378	0.424	**0.534**	0.384	0.316	0.381	0.285	0.402	0.456	0.348
ChatGPT (Stars w/ ref)	**0.474**	**0.491**	**0.407**	0.430	0.457	0.378	0.403	0.489	0.375	0.339	0.409	0.319	0.411	0.461	0.370

Table 1: Sample-level Spearman correlation (Spear.) correlation, Pearson (Pear.) correlation and Kendall's Tau (Kend.) of different aspects on SummEval (a text summarization meta-evaluation dataset). " Avg. " indicates the average performance. The **bold** indicates the best correlation.

4 Experiments

4.1 Experimental Setup

Metrics. To evaluate how well automatic metrics correlate with human judgment. Two widely-used correlation measures are adopted: (1) *Spearman correlation* (Zar, 2005) assesses the monotonic relationships between two variables; (2) *Pearson correlation* (Mukaka, 2012) measures the linear relationships between two sets of data; (3) *Kendall's Tau* (Kendall, 1938) evaluates the ordinal association between two measured quantities.

Evaluation Strategy. When calculating the correlation scores, there are different aggregation methods. Given a set of conditioned text $\{c_1, c_2, ..., c_n\}$ (*e.g.*, source documents in text summarization task) and M NLG models. The generated text of m-th model for the i-th condition text is denoted as $g_{i,m}$.
(1) *Sample-level* evaluation strategy calculates the correlation scores as follows:

$$\begin{aligned}\text{Corr}_{\text{sample}} =& \frac{1}{n}\sum_{1}^{n}(\rho([f_{\text{auto}}(g_{i,1}), ..., f_{\text{auto}}(g_{i,M})], \\ & [f_{\text{human}}(g_{i,1}), ..., f_{\text{human}}(g_{i,M})]))\end{aligned} \tag{1}$$

where ρ denotes the correlation metrics like Spearman correlation. f_{auto} and f_{human} indicate the automatic evaluation and human judgment functions, respectively.
(2) *Dataset-level* evaluation strategy calculates the correlation as follows:

$$\begin{aligned}\text{Corr}_{\text{dataset}} =& \rho([f_{\text{auto}}(g_{1,1}), ..., f_{\text{auto}}(g_{n,M})], \\ & [f_{\text{human}}(g_{1,1}), ..., f_{\text{human}}(g_{n,M})])\end{aligned} \tag{2}$$

4.2 Baselines

We compare the ChatGPT evaluator with the following widely-used automatic NLG metrics to provide deeper analyses:

- ROUGE-1, ROUGE-2 and ROUGE-L (Lin, 2004) measure the lexical overlap between the generated text and corresponding references based on unigram, bigram and longest common subsequence, respectively.
- BERTScore (Zhang et al., 2020) and MoverScore (Zhao et al., 2019) evaluate the semantic similarity via pre-trained BERT model (Devlin et al., 2019).
- PRISM (Thompson and Post, 2020) is used to evaluate NLG models via pre-trained paraphrase models.
- BARTScore (Yuan et al., 2021) is a state-of-the-art NLG metrics based on vanilla pre-trained BART model (Lewis et al., 2020).
- BARTScore+CNN (Yuan et al., 2021) could be regarded as an enhanced version of BARTScore. This metric is based on the BART fine-tuned on the CNN/DM dataset (Hermann et al., 2015).
- BARTScore+CNN+Para (Yuan et al., 2021) is another enhanced version of BARTScore. The metric is based on the BART fine-tuned on both CNN/DM and Paraphrase2.0 (Hu et al., 2019).
- Perplexity (PPL) is a commonly-used NLG metric to evaluate whether the generation result is grammatical and fluent.

4.3 Text Summarization

We conduct meta-evaluation on SummEval (Fabbri et al., 2021), NewsRoom (Grusky et al., 2018) and

Metrics	Coherence			Relevance			Informativeness			Fluency			Avg.		
	Spear.	Pear.	Kend.	Spear.	Pear.	Kend.	Spear.	Pear.	Kend.	Spear.	Pear.	Kend.	Spear.	Pear.	Kend.
ROUGE-1	0.095	-0.071	0.076	0.147	-0.001	0.112	0.130	-0.039	0.099	0.104	-0.074	0.082	0.119	-0.046	0.092
ROUGE-2	0.026	-0.108	0.009	0.091	-0.056	0.065	0.079	-0.087	0.052	0.048	-0.101	0.032	0.061	-0.088	0.092
ROUGE-L	0.064	-0.091	0.051	0.106	-0.034	0.083	0.089	-0.068	0.064	0.072	-0.090	0.061	0.083	-0.071	0.065
BERTScore	0.147	0.043	0.116	0.162	0.080	0.126	0.130	0.044	0.105	0.171	0.046	0.128	0.152	0.053	0.119
MoverScore	0.161	0.008	0.127	0.195	0.077	0.157	0.188	0.045	0.151	0.120	-0.008	0.086	0.166	0.030	0.130
PRISM	0.573	0.605	0.478	0.553	0.636	0.460	0.561	0.629	0.472	0.532	0.547	0.443	0.555	0.604	0.463
BARTScore	**0.679**	**0.709**	**0.568**	**0.604**	**0.744**	**0.507**	**0.646**	**0.749**	**0.543**	**0.670**	**0.662**	**0.564**	**0.650**	**0.716**	**0.545**
BARTScore+CNN	0.653	0.690	0.547	0.567	0.718	0.478	0.616	0.712	0.510	0.640	0.653	0.540	0.619	0.693	0.519
BARTScore+CNN+Para	0.657	0.675	0.544	0.562	0.739	0.465	0.614	0.727	0.507	0.652	0.630	0.545	0.621	0.693	0.515
ChatGPT (DA w/o ref)	0.469	0.487	0.405	0.461	0.587	0.392	0.578	0.645	0.498	0.507	0.524	0.427	0.504	0.561	0.430
ChatGPT (Stars w/o ref)	0.428	0.424	0.375	0.402	0.488	0.348	0.557	0.602	0.487	0.451	0.448	0.385	0.460	0.490	0.399
ChatGPT (DA w/ ref)	0.431	0.494	0.369	0.436	0.535	0.372	0.429	0.484	0.368	0.459	0.490	0.387	0.439	0.501	0.374
ChatGPT (Stars w/ ref)	0.423	0.424	0.369	0.443	0.506	0.395	0.404	0.463	0.352	0.503	0.504	0.430	0.443	0.474	0.387

Table 2: Sample-level Spearman correlation (Spear.) correlation, Pearson (Pear.) correlation and Kendall's Tau (Kend.) of different aspects on NewsRoom (a text summarization meta-evaluation dataset). " Avg. " indicates the average performance. The **bold** indicates the best correlation.

Metrics	Sample-level			Dataset-level		
	Spear.	Pear.	Kend.	Spear.	Pear.	Kend.
ROUGE-1	**0.498**	**0.526**	**0.408**	**0.533**	**0.555**	**0.383**
ROUGE-2	0.423	0.449	0.353	0.514	0.513	0.369
ROUGE-L	0.488	0.515	0.403	**0.533**	0.549	**0.383**
BERTScore	0.441	0.477	0.347	0.503	0.517	0.358
MoverScore	0.372	0.400	0.290	0.427	0.451	0.303
PRISM	0.411	0.458	0.324	0.478	0.494	0.339
BARTScore	0.441	0.467	0.342	0.467	0.458	0.327
BARTScore+CNN	0.475	0.500	0.374	0.436	0.455	0.306
BARTScore+CNN+Para	0.471	0.512	0.374	0.499	0.515	0.357
ChatGPT (DA w/o ref)	0.173	0.179	0.152	0.185	0.193	0.145
ChatGPT (Stars w/o ref)	0.145	0.162	0.129	0.170	0.179	0.136
ChatGPT (DA w/ ref)	0.184	0.208	0.154	0.276	0.288	0.206
ChatGPT (Stars w/ ref)	0.195	0.207	0.174	0.224	0.261	0.181

Table 3: Sample-level and Dataset-level correlation on RealSumm (a text summarization meta-evaluation dataset) (Spear.: Spearman correlation; Pear.: Pearson correlation; Kend.: Kendall's Tau). The **bold** indicates the best correlation.

RealSumm (Bhandari et al., 2020) to evaluate the performance of ChatGPT as an NLG metric for text summarization. SummEval collects 16 model-generated summaries on the CNN/DM dataset and annotates human judgments upon these summaries covering aspects of coherence, relevance, consistency and fluency. Newsroom, as a text summarization dataset, also provides human judgments on 7 model-generated summaries, including coherence, relevance, informativeness and fluency. RealSumm evaluates the pyramid (Nenkova and Passonneau, 2004) recall of 25 model-generated summaries.

The Potentiality of ChatGPT. Table 1 and Table 2 show the sample-level evaluation results on SummEval and NewsRoom, respectively (dataset-level evaluation results on SummEval and NewsRoom also shown in Table 4 and Table 5 with the similar trends). Experimental results show that ChatGPT achieves a new state-of-the-art correlation in most aspects of SummEval, demonstrating its potential of serving as an NLG metric. For results on Newsroom, ChatGPT also outperforms dominant summarization metrics (*i.e.*, ROUGE and BERTScore) by a large margin. Note that our experiments only estimate the lower bound of ChatGPT's performance, and better performances would like to be achieved once using better prompts or updated versions of ChatGPT.

The Impact of Dataset Biases. As shown in Table 3, we find that the experimental results on RealSumm show different trends from those on SummEval, *i.e.*, ChatGPT significantly underperforms other baseline metrics. For example, ChatGPT (Stars w/ ref) achieves 0.195 sample-level Spearman correlation, which is far behind the counterpart of ROUGE-1 (*i.e.*, 0.498). We conjecture this is because the human judgments in RealSumm are collected via pyramid method (Nenkova and Passonneau, 2004). In detail, this method first requires human evaluators to extract semantic content units from golden references, and then score each system summary based on how many extracted semantic content units are mentioned in the system summary.

In this manner, the more similarity between one generated summary and the corresponding golden reference, the more human evaluation scores will be achieved. Therefore, this reference-oriented annotation method makes the traditional n-gram-based metric (such as ROUGE) already achieve well correlations with human judgments, which we named as *lexical biases*. As for SummEval and NewsRoom, human evaluators are required to directly score different summaries without comparing them with the golden references, and thus do

Metrics	Coherence			Relevance			Consistency			Fluency			Avg.		
	Spear.	Pear.	Kend.	Spear.	Pear.	Kend.	Spear.	Pear.	Kend.	Spear.	Pear.	Kend.	Spear.	Pear.	Kend.
ROUGE-1	0.184	0.193	0.129	0.302	0.341	0.217	0.137	0.175	0.108	0.080	0.143	0.062	0.176	0.213	0.129
ROUGE-2	0.145	0.140	0.102	0.245	0.254	0.175	0.129	0.152	0.102	0.062	0.092	0.048	0.145	0.160	0.107
ROUGE-L	0.141	0.148	0.100	0.284	0.318	0.204	0.109	0.152	0.086	0.079	0.132	0.061	0.153	0.188	0.113
BERTScore	0.317	0.326	0.224	0.362	0.381	0.262	0.117	0.146	0.092	0.150	0.196	0.117	0.237	0.262	0.174
MoverScore	0.178	0.177	0.125	0.294	0.328	0.211	0.150	0.171	0.118	0.119	0.160	0.092	0.185	0.209	0.136
PRISM	0.286	0.281	0.204	0.280	0.297	0.202	0.323	0.297	0.256	0.236	0.248	0.184	0.281	0.281	0.212
BARTScore	0.335	0.357	0.241	0.363	0.386	0.263	0.269	0.276	0.212	0.187	0.206	0.146	0.288	0.306	0.215
BARTScore+CNN	0.408	0.434	0.292	0.394	0.423	0.286	0.334	0.377	0.264	0.285	0.354	0.223	0.355	0.397	0.266
BARTScore+CNN+Para	0.424	0.430	0.304	0.398	0.431	0.289	**0.379**	0.452	0.301	**0.346**	**0.410**	**0.271**	**0.387**	0.431	0.291
ChatGPT (DA w/o ref)	0.394	0.399	0.310	0.455	0.435	0.365	0.339	0.500	0.300	0.286	0.380	0.250	0.368	0.428	0.306
ChatGPT (Stars w/o ref)	0.435	0.438	0.353	0.448	0.459	0.366	0.356	**0.515**	**0.320**	0.300	0.385	0.268	0.385	**0.449**	0.327
ChatGPT (DA w/ ref)	0.418	0.426	0.327	**0.494**	**0.506**	**0.389**	0.363	0.507	0.315	0.237	0.329	0.203	0.378	0.442	0.308
ChatGPT (Stars w/ ref)	**0.465**	**0.472**	**0.385**	0.458	0.476	0.381	0.333	0.475	0.299	0.285	0.385	0.258	0.385	0.452	**0.331**

Table 4: Dataset-level Spearman correlation (Spear.) correlation, Pearson (Pear.) correlation and Kendall's Tau (Kend.) of different aspects on SummEval (a text summarization meta-evaluation dataset). " Avg. " indicates the average performance. The **bold** indicates the best correlation.

Metrics	Coherence			Relevance			Informativeness			Fluency			Avg.		
	Spear.	Pear.	Kend.	Spear.	Pear.	Kend.	Spear.	Pear.	Kend.	Spear.	Pear.	Kend.	Spear.	Pear.	Kend.
ROUGE-1	0.100	0.015	0.070	0.122	0.061	0.084	0.149	0.043	0.106	0.064	-0.009	0.043	0.109	0.028	0.076
ROUGE-2	0.080	0.033	0.060	0.124	0.071	0.092	0.158	0.060	0.119	0.045	0.018	0.032	0.102	0.045	0.076
ROUGE-L	0.079	-0.010	0.055	0.101	0.031	0.069	0.136	0.018	0.098	0.045	-0.030	0.029	0.090	0.002	0.063
BERTScore	0.169	0.138	0.122	0.176	0.158	0.127	0.196	0.153	0.141	0.154	0.129	0.109	0.174	0.145	0.125
MoverScore	0.173	0.119	0.122	0.192	0.156	0.132	0.232	0.148	0.166	0.112	0.091	0.076	0.177	0.129	0.124
PRISM	0.483	0.485	0.350	0.540	0.550	0.398	0.567	0.569	0.414	0.420	0.421	0.303	0.503	0.506	0.366
BARTScore	**0.656**	**0.666**	**0.495**	**0.588**	**0.700**	**0.439**	**0.645**	**0.710**	**0.485**	**0.615**	**0.610**	**0.464**	**0.626**	**0.671**	**0.471**
BARTScore+CNN	0.623	0.640	0.466	0.557	0.665	0.411	0.592	0.665	0.440	0.596	0.592	0.448	0.592	0.641	0.441
BARTScore+CNN+Para	0.621	0.639	0.465	0.575	0.692	0.427	0.615	0.694	0.459	0.593	0.577	0.444	0.601	0.650	0.449
ChatGPT (DA w/o ref)	0.383	0.418	0.297	0.491	0.541	0.392	0.527	0.576	0.413	0.401	0.398	0.309	0.451	0.483	0.353
ChatGPT (Stars w/o ref)	0.370	0.374	0.294	0.422	0.444	0.343	0.518	0.527	0.423	0.381	0.362	0.302	0.423	0.427	0.341
ChatGPT (DA w/ ref)	0.381	0.407	0.292	0.434	0.458	0.339	0.377	0.412	0.291	0.386	0.403	0.298	0.394	0.420	0.305
ChatGPT (Stars w/ ref)	0.370	0.355	0.295	0.425	0.426	0.342	0.373	0.400	0.301	0.439	0.425	0.353	0.402	0.402	0.323

Table 5: Dataset-level Spearman correlation (Spear.) correlation, Pearson (Pear.) correlation and Kendall's Tau (Kend.) of different aspects on NewsRoom (a text summarization meta-evaluation dataset). " Avg. " indicates the average performance. The **bold** indicates the best correlation.

Metrics	Sample-level			Dataset-level		
	Spear.	Pear.	Kend.	Spear.	Pear.	Kend.
ROUGE-1	0.014	0.020	0.013	-0.023	-0.010	-0.016
ROUGE-2	0.035	0.041	0.035	0.009	0.012	0.007
ROUGE-L	0.013	0.021	0.014	-0.016	-0.004	-0.011
BERTScore	0.140	0.120	0.105	0.081	0.084	0.056
BARTScore	-0.065	-0.082	-0.061	-0.065	-0.092	-0.045
BARTScore+CNN	0.049	0.026	0.041	0.047	0.053	0.033
BARTScore+CNN+Para	0.064	0.050	0.062	0.062	0.074	0.043
PPL	0.324	0.330	0.265	0.306	0.255	0.213
ChatGPT (DA w/o ref)	**0.507**	**0.533**	**0.439**	**0.471**	**0.494**	**0.366**
ChatGPT (Stars w/o ref)	0.472	0.490	0.427	0.415	0.439	0.342
ChatGPT (DA w/ ref)	0.411	0.434	0.357	0.363	0.375	0.281
ChatGPT (Stars w/ ref)	0.478	0.493	0.435	0.346	0.372	0.291

Table 6: Sample-level and Dataset-level correlation on OpenMEVA (a story generation meta-evaluation dataset) (Spear.: Spearman correlation; Pear.: Pearson correlation; Kend.: Kendall's Tau).

not involve such lexical biases.

The Impact of Different Prompt. In this work, we attempt four prompts to guide ChatGPT to evaluate the generation of NLG models. As we can see, the performances of ChatGPT are sensitive to the prompt design. For different aspects, the prompt should be carefully designed, just like formulating instructions for human evaluators.

4.4 Story Generation

Story generation is another NLG task with more emphasis on open-ended generation compared with text summarization, which also means for a given beginning of a story, various generated storylines and different plots could satisfy people. Therefore, story generation models are extremely challenging to evaluate. The automatic similarity-based metrics between the generated storylines and so-called references cannot fully evaluate the quality of the storylines since they do not consider creativity.

To show the effectiveness of ChatGPT as an NLG metric for the story generation task, we conduct experiments on OpenMEVA-ROC (Guan et al., 2021). The OpenMEVA-ROC dataset manually annotates five model-generated storylines under the consideration of their overall quality.

Metrics	Informativeness			Naturalness			Quality			Avg.		
	Spear.	Pear.	Kend.	Spear.	Pear.	Kend.	Spear.	Pear.	Kend.	Spear.	Pear.	Kend.
ROUGE-1	0.092	0.093	0.073	0.265	0.274	0.206	0.235	0.234	0.184	0.197	0.200	0.154
ROUGE-2	0.133	0.137	0.103	0.233	0.241	0.177	0.192	0.192	0.145	0.186	0.190	0.142
ROUGE-L	0.079	0.084	0.063	0.237	0.255	0.183	0.210	0.216	0.163	0.175	0.185	0.136
BERTScore	0.231	0.261	0.174	0.288	0.327	0.216	0.264	0.304	0.197	0.261	0.297	0.196
MoverScore	**0.284**	**0.276**	0.209	0.189	0.189	0.140	0.152	0.161	0.114	0.208	0.209	0.154
PRISM	0.255	0.268	0.189	0.301	0.350	0.223	0.308	0.337	0.226	0.288	0.318	0.213
BARTScore	0.234	0.270	0.174	0.221	0.280	0.164	0.186	0.245	0.139	0.214	0.265	0.159
BARTScore+CNN	0.237	0.253	0.177	0.312	0.382	0.233	0.294	0.357	0.219	0.281	0.331	0.210
BARTScore+CNN+Para	0.240	0.266	0.177	**0.335**	**0.416**	0.248	**0.343**	**0.383**	0.255	**0.306**	**0.355**	0.227
ChatGPT (DA w/o ref)	-	-	-	0.243	0.293	0.202	0.310	0.319	0.260	-	-	-
ChatGPT (Stars w/o ref)	-	-	-	0.316	0.389	**0.269**	0.307	0.367	**0.266**	-	-	-
ChatGPT (DA w/ ref)	0.247	0.255	0.198	0.305	0.344	0.248	0.269	0.343	0.215	0.274	0.314	0.220
ChatGPT (Stars w/ ref)	0.266	0.262	**0.224**	0.293	0.374	0.235	0.276	0.328	0.239	0.278	0.321	**0.233**

Table 7: Dataset-level Spearman correlation (Spear.) correlation, Pearson (Pear.) correlation and Kendall's Tau (Kend.) of different aspects on BAGEL (a data-to-text generation meta-evaluation dataset). " Avg. " indicates the average performance. The **bold** indicates the best correlation.

The Potentiality of ChatGPT. As shown in Table 6, ChatGPT achieves the best performance in terms of all correlations, and significantly outperforms the second-best metric (*i.e.*, PPL). For example, ChatGPT (DA w/o ref) achieves 0.507 sample-level Spearman correlation, while PPL only achieves 0.324 sample-level Spearman correlation. In addition, we also find that all similarity-based metrics (*i.e.*, ROUGE-1, ROUGE-2, ROUGE-L, BERTScore and BARTScore) show their weak correlations with the human judgments. This finding indicates that the ChatGPT evaluator has more powerful and reliable judgments on the open-ended and creative text generation tasks, where many diversified generated results could also be regarded as high-quality.

The Impact of Different Prompt. The results in Table 6 also show the sensitivity of the correlation results led by the different prompts. For example, there are large performance gaps between ChatGPT (DA w/o ref) and ChatGPT (DA w/ ref). This finding is also consistent with that in text summarization (Section 4.3). More recently, some researchers also discuss the robustness of LLMs on different (adversarial) prompts (Zhu et al., 2023), and we think this under-explored LLM research direction deserves more research attention.

4.5 Data-to-Text Generation

Data-to-text generation aims at generating a fluent free-text description for a given structured table. We conduct experiments on BAGEL (Mairesse et al., 2010) to show the effectiveness of the ChatGPT evaluator on data-to-text generation.

Table 7 shows the experimental results, where ChatGPT achieves competitive correlations compared with the previous state-of-the-art baselines, indicating its strong potentiality serving as a metric for data-to-text generation. It is worth noting that we do not provide reference-free ChatGPT performance in terms of informativeness because informativeness in BAGEL is defined as "whether the system generation contains all the information in the gold reference", which also means that when evaluating informativeness the golden references must be given.

5 Conclusion

In this technical report, we explore a research question: "*Is ChatGPT a good NLG evaluator?*". To this end, we design task-specific as well as aspect-specific prompts to guide ChatGPT to perform as an NLG metric. Experimental results on five widely-used meta-evaluation datasets, covering text summarization, story generation and data-to-text tasks, show the potentiality of ChatGPT as an NLG metric. ChatGPT achieves the new state-of-the-art correlations (with human judgments) on SummEval and OpenMEVA meta-evaluation datasets, and obtains competitive results on NewsRoom and BAGEL datasets.

In addition, we also find that the lexical biases involved in the meta-evaluation datasets would influence the effectiveness of NLG metrics, and might lead to the limited performance of the ChatGPT evaluator. Besides, the performances of ChatGPT as an NLG evaluator are sensitive to the format

of the prompt, for different tasks and aspects, the prompt should be carefully designed.

We believe that ChatGPT will exceed its current performance and provide a reliable NLG metric for the research community in the near future.

Limitations

While we show that ChatGPT achieves state-of-the-art or competitive correlation with human judgments on various NLG tasks, there are limitations that provide avenues for future work: (1) ChatGPT's performance as an NLG metric is related to prompts, and future work could explore more powerful prompts to achieve better performance; (2) This preliminary report misses experiments on some mainstream NLG tasks, *e.g.*, dialogue generation and report generation; (3) When we did the experiments, the OpenAI ChatGPT did not release the official API. Thus, we conducted the experiments on the ChatGPT website with default temperature, making the results difficult to reproduce. All experiments related to ChatGPT are conducted between February 24 to February 27, 2023; and March 17 to March 22. (4) The experiments are only conducted on the English NLG meta-evaluation datasets, and future work could extend this method into other languages or cross-lingual scenes. (5) The correlation between the ChatGPT evaluator and humans is also related to the quality and challenge of the corresponding meta-evaluation datasets. Our experiments are conducted on the traditional NLG meta-evaluation datasets (that appear before the LLM era). Recently, Zeng et al. (2023) propose LLM-BAR, a challenging meta-evaluation benchmark to test the ability of an LLM evaluator. Future work could adapt our method to other challenging datasets and study the performance of the ChatGPT evaluator.

Acknowledgement

We thank anonymous reviewers for their constructive suggestions and comments.

References

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359, Online. Association for Computational Linguistics.

Hong Chen, Duc Vo, Hiroya Takamura, Yusuke Miyao, and Hideki Nakayama. 2022. StoryER: Automatic story evaluation via ranking, rating and reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1739–1753, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 85–91, Edinburgh, Scotland. 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.

Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4884–4895, Florence, Italy. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. 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.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. 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.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1693–1701.

J. Edward Hu, Abhinav Singh, Nils Holzenberger, Matt Post, and Benjamin Van Durme. 2019. Large-scale, diverse, paraphrastic bitexts via sampling and clustering. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 44–54, Hong Kong, China. Association for Computational Linguistics.

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

Katharina Jeblick, Balthasar Schachtner, Jakob Dexl, Andreas Mittermeier, Anna Theresa Stüber, Johanna Topalis, Tobias Weber, Philipp Wesp, Bastian Sabel, Jens Ricke, and Michael Ingrisch. 2022. Chatgpt makes medicine easy to swallow: An exploratory case study on simplified radiology reports.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520*.

Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, Anna Kocoń, Bartłomiej Koptyra, Wiktoria Mieleszczenko-Kowszewicz, Piotr Miłkowski, Marcin Oleksy, Maciej Piasecki, Łukasz Radliński, Konrad Wojtasik, Stanisław Woźniak, and Przemysław Kazienko. 2023. Chatgpt: Jack of all trades, master of none.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 957–966, Lille, France. PMLR.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

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

François Mairesse, Milica Gašić, Filip Jurčíček, Simon Keizer, Blaise Thomson, Kai Yu, and Steve Young. 2010. Phrase-based statistical language generation using graphical models and active learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1552–1561, Uppsala, Sweden. 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.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 145–152, Boston, Massachusetts, USA. Association for Computational Linguistics.

OpenAI. 2022. Introducing chatgpt. `https://openai.com/blog/chatgpt`.

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*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Krishna Pillutla, Lang Liu, John Thickstun, Sean Welleck, Swabha Swayamdipta, Rowan Zellers, Sewoong Oh, Yejin Choi, and Zaid Harchaoui. 2022. Mauve scores for generative models: Theory and practice.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Haocong Rao, Cyril Leung, and Chunyan Miao. 2023. Can chatgpt assess human personalities? a general evaluation framework.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

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.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2023a. Cross-lingual summarization via chatgpt. *arXiv preprint*.

Shuai Wang, Harrisen Scells, Bevan Koopman, and Guido Zuccon. 2023b. Can chatgpt write a good boolean query for systematic review literature search?

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, Yong Jiang, and Wenjuan Han. 2023. Zero-shot information extraction via chatting with chatgpt.

Xianjun Yang, Yan Li, Xinlu Zhang, Haifeng Chen, and Wei Cheng. 2023. Exploring the limits of chatgpt for query or aspect-based text summarization. *arXiv preprint arXiv:2302.08081*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation.

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

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2023. Evaluating large language models at evaluating instruction following. *arXiv preprint arXiv:2310.07641*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Neil Zhenqiang Gong, Yue Zhang, et al. 2023. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv preprint arXiv:2306.04528*.

Guido Zuccon and Bevan Koopman. 2023. Dr chatgpt, tell me what i want to hear: How prompt knowledge impacts health answer correctness.