Are Large Language Models Good Evaluators for Abstractive Summarization?

Chenhui Shen [*] [1,2] **Liying Cheng** [1] **Yang You**[2] **Lidong Bing**[†1]
[1]DAMO Academy, Alibaba Group [2] National University of Singapore
{chenhui.shen, liying.cheng, l.bing}@alibaba-inc.com
youy@comp.nus.edu.sg

Abstract

Human evaluations are often required for abstractive summary evaluations to give fairer judgments. However, they are often time-consuming, costly, inconsistent, and non-reproducible. To overcome these challenges, we explore the potential of using an out-of-the-box LLM (i.e. "gpt-3.5-turbo") for summarization evaluation without manually selecting demonstrations or complex prompt tuning. We compare different evaluation methods, including 2 methods for Likert-scale scoring and 1 method for head-to-head comparisons, to investigate the performance of the LLM as a zero-shot evaluator. We further propose a meta-correlation metric to measure the stability of the LLM's evaluation capability. With extensive experiments, we show that certain prompt formats can produce better results than others. We also bring attention to the LLM's deteriorating evaluation capability with the rising qualities of summaries. In addition, we find that the LLM's evaluation capability also depends on the evaluated dimensions. We discuss the pros and cons of each method, make recommendations, and suggest some future directions for improvement. [1]

1 Introduction

To evaluate the performance of abstractive summarization systems, researchers often use both automatic and human evaluation metrics. Automatic evaluation metrics like ROUGE (Lin, 2004) are easy to use, yet are only compared against the "gold" output summaries. However, it is time-consuming and costly to manually prepare summaries for each input article, and many datasets (e.g., CNN/DM, XSUM) use automatically compiled "gold" summaries, which are often unnatural for containing sentence fragments or click-baits (Tanya Goyal, 2022; Fabbri et al., 2020). Moreover, summarization is an open task that could have various good summaries for the same input article, so simply comparing against 1 or 2 "gold" outputs may not be sufficient.

In order to evaluate the generated summaries more comprehensively, human evaluation is usually required for judgments in various dimensions, such as fluency and consistency. However, such fine-grained evaluation requires domain experts, and the crowd-sourced non-professional annotators often produce inconsistent and unreliable evaluations (Chiang and Lee, 2023; Fabbri et al., 2020).

To save time and cost, researchers usually conduct human evaluations on a small subset (e.g., 50-100 samples) of the test set. The evaluated samples are often not standardized across datasets, making the evaluation difficult to reproduce. In addition, the small sample size may result in sample biases, further inhibiting the reliability of human evaluation. Therefore, how to define and design a better evaluation method is an inevitable research topic.

Recently, large language models (LLMs) have demonstrated their strong abilities in understanding contexts and generating reasonable replies. Some concurrent works (Chiang and Lee, 2023; Wang et al., 2023a; Wu et al., 2023; Luo et al., 2023) have demonstrated the potential of using LLM as a zero-shot evaluator for natural language generation tasks. Despite the preliminary success suggested by these works, many questions remain. For instance, to what degree can we trust the evaluation results by LLMs for different evaluation dimensions? Is there a need for extensive prompt design or can we simply use a LLM out-of-the-box? Are there evaluation methods that are more reliable than others? These are some of the questions we must address before we can readily use LLM evaluations to replace human evaluations.

In this paper, we seek to address some of the

[*] Chenhui is under the Joint PhD Program between Alibaba and National University of Singapore.

[†] Corresponding author.

[1]Our code and data will be fully released at https://github.com/DAMO-NLP-SG/LLM_summeval.

above questions. Specifically, we explore various evaluation methods on an out-of-the-box LLM ("gpt-3.5-turbo"[2]) with minimal effort in prompt design. We use ChatGPT in the following sections to refer to this model. In short, we investigate the zero-shot capability of ChatGPT for evaluating abstractive summarization systems without the need for manually selecting demonstrations or complex prompt tuning. We follow two common human evaluation approaches for ChatGPT-conducted evaluations, namely Likert-scale scoring (He et al., 2022; Shen et al., 2022b; Zhang et al., 2020a) and head-to-head (H2H) comparisons (Shen et al., 2022a; Li et al., 2020; Liu and Lapata, 2019).

We formulate the Likert-scale scoring as either a direct reason-then-score (RTS) generation method or a multiple-choice question (MCQ) method. We further define a meta-correlation metric to measure the stability of ChatGPT's performance across different summarization systems. With extensive analysis, we find that the MCQ method performs better, and is more stable than the RTS method. However, both methods suffer from negative meta-correlations, which indicate the ChatGPT's deteriorating evaluation capability with increasing qualities of summarization systems. In terms of the H2H method, although it achieves impressive performance on the consistency dimension, it is much less effective on other dimensions. Finally, we discuss the implications of our findings and possible future directions.

Our contributions are as follows:

- We investigate the effectiveness of ChatGPT as a zero-shot evaluator for abstractive summarization systems with various evaluation methods.
- We propose a meta-correlation metric to measure the stability of ChatGPT's evaluation performance.
- We highlight the strengths and weaknesses of different evaluation methods and make recommendations on how to better use ChatGPT for out-of-the-box evaluations.

2 Related Work

Summarization The summarization task involves generating a summary that contains succinct and important contents of the original input article (Nenkova and McKeown, 2012). Common summarization systems can be mostly divided into 2 categories, either extractive or abstractive. Unlike extractive summarization systems (Ernst et al., 2022; Chen et al., 2021; Zhou et al., 2018; Dong et al., 2018) that directly extract important sentences or phrases from the input article, abstractive summarization systems (Shen et al., 2023; Liu et al., 2022; Xiao et al., 2022; Lewis et al., 2020; Zhang et al., 2020a; Ziegler et al., 2019) often generate the summary using their own words, applying sentence fusion or paraphrasing techniques (Bing et al., 2015; Xu and Durrett, 2021). We focus on evaluating the abstractive summarization systems, as they pose more significant challenges for human evaluation (Saha et al., 2022; Pagnoni et al., 2021).

Human Evaluation Methods Many previous works (He et al., 2022; Shen et al., 2022b; Zhang et al., 2020a; Cheng et al., 2020; Gao et al., 2019; Liu et al., 2018; Li et al., 2017; Kryściński et al., 2018) employ a Likert scale to evaluate the summaries on discrete ranges, such as from 1 to 5. On the other hand, Shen et al. (2022a), Li et al. (2020), Liu and Lapata (2019), Fan et al. (2018), and Fabbri et al. (2019) adopt a comparison approach by asking the human annotators to select the best summary out of 2 or more generated summaries by different systems on the same input article. We use both the Likert-scale scoring approach as well as the head-to-head comparison approach (the simplest comparison approach) for ChatGPT-conducted evaluations.

Automatic Evaluation Metrics ROUGE (Lin, 2004) is a common lexical overlap metric frequently used for automatic evaluation for summarization systems. However, it may experience limited effectiveness, especially for abstractive summarization systems that use various fusion and re-writing techniques. To mitigate the above issue, Zhang et al. (2020b) proposes BERTScore, which leverages the word embeddings in the pre-trained BERT model (Devlin et al., 2019) to evaluate the semantic similarity of individual tokens. Yuan et al. (2021) takes a step further and introduces BARTScore, which makes use of the language modeling capability of the pre-trained BART (Lewis et al., 2020) to estimate the probability of a summary given its input article. In addition, the BARTScore can be adapted to specific domains by additional fine-tuning on the corresponding summarization datasets, giving rise to other metrics

[2] The "gpt-3.5-turbo" API is introduced as the ChatGPT API in https://openai.com/blog/introducing-chatgpt-and-whisper-apis.

such as BARTScore-CNN and BARTScore-CNN-Para (Yuan et al., 2021). We compare the results of ChatGPT-conducted evaluations against all these metrics in terms of correlation with human scores.

LLM Evaluations There are many concurrent works that demonstrate the potential of LLMs as zero-shot evaluators. Chiang and Lee (2023) use LLMs for open-ended story evaluations, Luo et al. (2023) apply ChatGPT specifically for evaluating the consistency of summaries, and Wu et al. (2023) formulate LLMs as diverse role-players to evaluate summaries from the perspectives of different personas. Wang et al. (2023a) also explores the LLM's evaluation potential in various dimensions for the natural language generation task. Our work differs from the above, as we focus on investigating the ChatGPT's evaluation capability with various evaluation methods across different dimensions for abstractive summarization.

3 ChatGPT as a Zero-Shot Evaluator

We investigate ChatGPT's evaluation capabilities in the dimensions of coherence, consistency, fluency, and relevance respectively, as defined by Fabbri et al. (2020). Following common human evaluation approaches, we propose two methods for Likert-scale scoring, namely the Reason-Then-Score method and the Multiple-Choice Question method, as well as one method for head-to-head comparisons. We introduce each method in Sec.3.1.

Besides investigating the performance of ChatGPT with different methods, we believe the stability of ChatGPT's evaluations across different summarization systems is equally important. Ideally, a good evaluator should perform equally well regardless of the evaluated systems. In Sec.3.2, we discuss how to investigate the evaluation stability of ChatGPT. In addition, we introduce a meta-correlation metric, to gauge the extent to which ChatGPT's evaluation performance depends on the evaluated systems.

3.1 Summary Evaluation Methods

Reason-then-Score (RTS) Given the success of chain-of-thought prompting (Kojima et al., 2022; Wei et al., 2022), the most intuitive method would be to ask ChatGPT to evaluate a specific dimension by outputting the reasons followed by a score. Since the SummEval dataset (Fabbri et al., 2020) contains human scores on a Likert scale of 1 to 5, we also ask ChatGPT to score the summaries in this range. We show an example of the prompt used for this method on the relevance dimension in Table 1. The definition of relevance and all other dimensions used are copied from Fabbri et al. (2020).

Score the following Summary given the corresponding Article with respect to relevance from 1 to 5. Note that relevance measures the Summary's selection of important content from the Article. 5 points indicate all information included in the Summary are important and non-trivial to the Article. Provide your reason in 1 sentence, then give a final score.

Article: {article}

Summary: {summary}

Reason: ?

Score: ?

Table 1: Example prompt for the RTS method on the relevance dimension. Text in {cyan}: the specific article and the corresponding summary to be evaluated.

Choose an option from A to E that gives a suitable score for the following Summary given the corresponding Article with respect to relevance. Note that relevance measures the Summary's selection of important content from the Article.

Article: {article}

Summary: {summary}

A: No information included in the Summary is relevant to the Article. The Summary is totally irrelevant to the Article. 1 point.
B: Only a small portion of the Summary is relevant to the Article. The majority of the Summary is irrelevant to the Article. 2 points.
C: Some information in the Summary is relevant to the Article whereas some are not. 3 points.
D: The majority of the Summary is relevant to the Article. Only a small portion of the Summary is irrelevant to the Article. 4 points.
E: All information included in the Summary are relevant to the Article. 5 points.

Your Answer (enter 1 letter from A to E): ?

Table 2: Example prompt for the MCQ method on the relevance dimension. Text in {cyan}: the specific article and the corresponding summary to be evaluated.

MCQ Scoring (MCQ) Nevertheless, some previous works find that the reasons generated by LLMs do not always make sense (Lyu et al., 2023; Wang et al., 2023b; Gao et al., 2022). To avoid

Evaluate the relevance of Summary 1 and Summary 2 with the corresponding Article, then choose an option from A, B, and C below. Note that relevance measures the summary's selection of important content from the Article. A perfectly relevant summary should include information that is important to the Article and should not include other trivial information.

Article: {article}

Summary 1: {summary from model A}

Summary 2: {summary from model B}

A: Summary 1 is more relevant than Summary 2.
B: Summary 2 is more relevant than Summary 1.
C: Both Summary 1 and Summary 2 are equally relevant.

Your choice (enter A, B, or C): ?

Table 3: Example prompt for the H2H method on the relevance dimension. Text in {cyan}: the specific article, and the corresponding summaries generated by a pair of compared models.

the misguidance of wrongly generated reasons, we explore a more constrained multiple-choice question (MCQ) method for the Likert-scale scoring. As shown in Table 2, instead of allowing ChatGPT to freely generate its reasons, we fix the specific reasons for each score value, then ask the model to output its choice.

Head-to-Head Comparison (H2H) Lastly, some concurrent works (Ma et al., 2023a,b) also observe that ChatGPT can act as an effective ranking model. We thus explore the head-to-head comparison approach. As shown in Table 3, we present 2 summaries (Summary 1 and Summary 2) generated by different summarization systems on the same input article, then prompt ChatGPT to select a better summary (Option A or B). We further include an option for ChatGPT to indicate that both summaries are equally good (Option C). We ask ChatGPT to choose from three options rather than directly generating natural text responses, because we observe from our preliminary studies that ChatGPT can rarely indicate that both summaries are equally good with direct generations.

Moreover, to avoid potential biases due to the 2 summaries' order of appearance, we separately conduct each evaluation twice, presenting the same summary as Summary 1 for the first run and as Summary 2 for the second run. We conclude that ChatGPT prefers a specific summary only if it has chosen the same summary for both runs. Otherwise, we assume that ChatGPT has evaluated both summaries to be equally good.

3.2 Stability of ChatGPT's Evaluations

To ensure fairness across different evaluated systems, we believe that it is crucial for ChatGPT to produce stable evaluations. That is, regardless of evaluated systems, ChatGPT should still be able to evaluate the summaries in a consistent manner. We investigate the stability of ChatGPT in two ways as below.

First, we group the evaluated summaries according to the corresponding summarization system that they are generated from. For each system, we investigate the correlation between ChatGPT's evaluation and human evaluations. The summaries used are all generated from the same set of input articles. Ideally, if ChatGPT is stable across systems, it should produce evaluations that are similarly correlated to human evaluations. If the correlations differ significantly across various summarization systems, then we may conclude that ChatGPT's evaluation is system-dependent.

Second, we define a **meta-correlation** metric to measure the extent to which the evaluation capability of ChatGPT is affected by the quality of the evaluated systems. Specifically, we use the average human score for each system as an indicator of its summarization quality (Q_i), as shown in Eqn.1:

$$Q_i = \frac{1}{N} \sum_{j=1}^{N} f_{human}(g_{i,j}) \tag{1}$$

where $f_{human}(\cdot)$ indicates the human evaluation, $g_{i,j}$ represents the j^{th} summary generated by the i^{th} summarization system. Each system's quality is calculated as an average of N generated summaries. Next, we use the correlation between ChatGPT and human scores an indicator of the evaluation performance (P_i) of ChatGPT for a specific system, as shown in Eqn.2:

$$\begin{aligned} P_i = \rho([f_{LLM}(g_{i,1}), ..., f_{LLM}(g_{i,N})], \\ [f_{human}(g_{i,1}), ..., f_{human}(g_{i,N})]) \end{aligned} \tag{2}$$

where ρ denotes the correlation metrics such as Spearman correlation, Pearson correlation, or Kendall's Tau[3], and $f_{LLM}(\cdot)$ indicates the ChatGPT evaluation for each summary $g_{i,j}$. Finally, we

[3] For consistency, we use the same correlation metric for both Eqn.2 and 3.

ID	Model Name	Avg	Coh	Con	Flu	Rel
M22	BART	4.57	4.18	4.94	4.90	4.25
M23	Pegasus (C4)	4.55	4.16	4.91	4.88	4.26
M17	T5	4.52	4.00	4.93	4.93	4.23
M12	Unified-ext-abs	4.32	3.60	4.96	4.85	3.85
M13	ROUGESal	4.24	3.44	4.82	4.86	3.83
M15	Closed book decoder	4.19	3.35	4.95	4.80	3.67
M14	Multi-task (Ent + QG)	4.12	3.20	4.90	4.74	3.63
M8	Pointer Generator	4.07	3.29	4.65	4.79	3.55
M9	Fast-abs-rl	3.77	2.38	4.67	4.50	3.52
M10	Bottom-Up	3.70	2.73	4.25	4.42	3.38
M20	GPT-2 (zero-shot)	3.58	3.63	3.40	3.97	3.30
M11	Improve-abs	3.09	2.28	3.27	3.65	3.15

Table 4: The average human evaluation scores of various abstractive summarization models reported by Fabbri et al. (2020). We calculate the average (Avg) score of the reported coherence (Coh), consistency (Con), fluency (Flu), and relevance (Rel) scores. Rows are sorted according to the Avg column values in descending order.

calculate the meta-correlation (M) on a total of k evaluated systems as follows:

$$M = \rho([Q_1, ..., Q_k], [P_1, ..., P_k]) \tag{3}$$

Ideally, ChatGPT should work equally well regardless of the quality of the evaluated systems, which means that M should be close to zero. On the other hand, a significant M would indicate a linear relationship between the evaluation capability of ChatGPT and the quality of the evaluated systems, suggesting that ChatGPT's evaluation is not stable, but dependent on the quality of the evaluated systems.

4 Experiments

We use the ChatGPT API ("gpt-3.5-turbo") for evaluating the summaries. Similar to Luo et al. (2023) and Wu et al. (2023), we set the temperature to 0 for more reproducible performance. We reset the dialogue history by sending in each request separately.

Dataset We use the SummEval benchmark dataset (Fabbri et al., 2020). This dataset contains expert human annotations for coherence, consistency, fluency, and relevance on the generation results from 12 abstractive systems on the CNN/Daily Mail dataset (Hermann et al., 2015). Each system generates summaries for the same 100 news articles, and each summary is evaluated by 3 expert annotators[4].

For the RTS and MCQ methods, we use ChatGPT to generate a score from 1 to 5 for all 1200 summaries. For H2H, due to a much larger number of possible combinations and a limited budget, we only evaluate the model pairs of consecutive performances. For instance, as shown in Tab.4, M22 has the best average human score of 4.57, followed by M23 of 4.55, then M17 of 4.52. We thus compare model pairs of "M22-M23" and "M23-M17" (See all evaluated pairs in Tab.10). This setting is not only the most challenging but also useful due to the common need to compare 2 systems with close performances.

Evaluations We conduct zero-shot evaluation following our prompt format given in Tab.1, 2, and 3. We re-use the definitions of the evaluation dimensions from Fabbri et al. (2020) as below: (1) Coherence (Coh): "the collective quality of all sentences"; (2) Consistency (Con): "the factual alignment between the summary and the summarized source "; (3) Fluency (Flu): " the quality of individual sentences"; (4) Relevance (Rel): "the selection of important content from the source".

For the RTS and MCQ methods, we calculate the correlation between the expert human scores reported in Fabbri et al. (2020) and the ChatGPT scores for a total of 1200 summaries. Specifically, we report the Spearman correlation (Zar, 2005), Pearson correlation (Cohen et al., 2009), and Kendall's Tau (Kendall, 1938). In addition, as discussed in Sec.3.2, we also investigate the evaluation stability of ChatGPT by calculating its various correlations with human scores for each individual summarization system, and further report the meta-correlations using the 3 correlation metrics mentioned above.

For the H2H method, we calculate the percentage of times that ChatGPT and humans (Fabbri et al., 2020) select the same summarization system as the better system out of a given pair. We coin this value as the **success rate** of ChatGPT. Specifically, for each dimension, we count the number of times that ChatGPT prefers one system over the other, and regard the system with more counts as the better system. Similarly, we count the number of times one system obtains a higher average

[4] Although each summary is also evaluated by 5 crowd-sourced workers, we don't use these scores as Fabbri et al. (2020) find the crowd-sourced annotations unreliable and inaccurate.

	Coh			Con			Flu			Rel		
ID	RTS	MCQ	human	RTS	MCQ	human	RTS	MCQ	human	RTS	MCQ	human
M8	1.92	3.85	3.29	4.34	4.80	4.65	2.35	4.20	4.79	3.37	4.40	3.55
M9	1.86	3.71	2.38	4.23	3.74	4.67	1.97	3.88	4.50	3.23	4.41	3.52
M10	1.90	3.73	2.73	4.09	4.52	4.25	2.06	3.85	4.42	3.20	4.33	3.38
M11	1.61	3.01	2.28	3.36	3.96	3.27	1.71	3.03	3.65	3.18	4.18	3.15
M12	2.00	4.01	3.60	4.49	4.86	4.96	2.47	4.35	4.85	3.54	4.50	3.85
M13	1.98	4.12	3.44	4.32	4.84	4.82	2.54	4.34	4.86	3.64	4.47	3.83
M14	1.90	4.02	3.20	4.21	4.79	4.90	2.25	4.15	4.74	3.45	4.42	3.63
M15	2.00	3.98	3.35	4.42	4.84	4.95	2.47	4.29	4.80	3.49	4.42	3.67
M17	2.11	4.31	4.00	4.64	4.90	4.93	2.56	4.31	4.93	3.75	4.53	4.23
M20	1.94	3.32	3.63	3.41	3.46	3.40	2.53	3.38	3.97	3.08	4.09	3.30
M22	2.23	4.15	4.18	4.60	4.83	4.94	2.96	4.36	4.90	3.85	4.54	4.25
M23	2.04	4.27	4.16	4.53	4.81	4.91	2.54	4.33	4.88	3.77	4.57	4.26
Avg	1.96	3.87	3.35	4.22	4.53	4.55	2.37	4.04	4.61	3.46	4.41	3.72

Table 5: Comparison between the average ChatGPT scores and human scores.

	Coh			Con			Flu			Rel		
Metrics	Spear.	Pear.	Kend.	Spear.	Pear.	Kend.	Spear.	Pear.	Kend.	Spear.	Pear.	Kend.
Rouge-1*	0.193	0.202	0.136	0.155	0.186	0.121	0.075	0.153	0.058	0.323	0.361	0.231
Rouge-2*	0.145	0.146	0.101	0.137	0.156	0.107	0.053	0.095	0.041	0.255	0.262	0.181
Rouge-L*	0.148	0.158	0.105	0.133	0.167	0.103	0.078	0.146	0.060	0.306	0.340	0.219
BERTScore*	0.375	0.383	0.265	0.163	0.182	0.127	0.167	0.229	0.130	0.396	0.414	0.285
BARTScore*	0.381	0.391	0.275	0.271	0.265	0.212	0.168	0.187	0.131	0.381	0.391	0.276
BARTScore-CNN*	**0.461**	**0.480**	0.332	0.389	0.413	0.305	0.310	0.378	0.241	**0.425**	**0.450**	0.309
BARTScore-CNN-Para*	0.455	0.455	0.328	**0.413**	0.459	0.324	**0.368**	**0.417**	**0.286**	0.414	0.440	0.299
ChatGPT-RTS	0.305	0.306	0.257	0.313	0.396	0.282	0.271	0.259	0.240	0.348	0.378	0.281
ChatGPT-MCQ	0.417	0.396	**0.346**	0.361	**0.516**	**0.336**	0.238	0.367	0.212	0.372	0.375	**0.319**

Table 6: Spearman's (Spear.) correlations, Pearson's (Pear.) correlations, and Kendal's Tau (Kend.) between various metrics and human scores for a total of 1200 summaries. *: results derived from Wang et al. (2023a). **Bolded**: best results. Values in light gray color are insignificant (p-value $\geq$ 0.05).

human score and determine the better system according to human evaluation.

Baselines Metrics We use Rouge (Lin, 2004) F1 scores for Rouge-1, Rouge-2, and Rouge-L, BERTScore (Zhang et al., 2020b), BARTScore (Yuan et al., 2021), BARTScore-CNN, and BARTScore-CNN-PARA as baseline metrics. The last three metrics make use of the language modeling capabilities of pre-trained language models, and are especially strong.

5 Results and Analysis

5.1 RTS Results

In Table 5, we compare the average ChatGPT scores using the RTS method against human-evaluated scores (the "RTS" and "human" columns) on all 4 dimensions. It can be seen that with the RTS method, ChatGPT scores different systems much more conservatively than humans across all dimensions.

As shown in Table 6, when considering a total of 1200 summaries combined from all systems, the ChatGPT-RTS have a low but significant correlation with human evaluations. The correlation of RTS scores is stronger than that of the automatic Rouge-1/2/L metrics, comparable to that of the BARTScore metric, but lags behind those of BARTScore-CNN and BARTScore-CNN-Para that make use of pre-trained language models specifically fine-tuned with the in-domain news data.

Given that many works (Lyu et al., 2023; Wang et al., 2023b; Gao et al., 2022) find the reasons generated by LLMs to be unfaithful, we further investigate the soundness of the reasons generated by the RTS method. Similar to Luo et al. (2023), we find ChatGPT to conduct false inferences. One major issue with the RTS method is that ChatGPT penalizes the summary for reasons not related to the evaluated dimension (such as penalizing the consistency dimension for repetitiveness, or penalizing the fluency dimension for missing important details)[5]. In

[5] Fabbri et al. (2020) have observed the same issue with the crowd-sourced (non-expert) annotators, but have seen an

	Coh			Con			Flu			Rel		
ID	Spear.	Pear.	Kend.	Spear.	Pear.	Kend.	Spear.	Pear.	Kend.	Spear.	Pear.	Kend.
M8	0.170	0.204	0.140	0.224	0.324	0.208	0.222	0.173	0.203	0.350	0.383	0.280
M9	0.263	0.253	0.227	0.234	0.224	0.210	0.099	0.177	0.089	0.266	0.298	0.218
M10	0.316	0.275	0.262	0.062	0.175	0.053	0.126	0.113	0.113	0.240	0.284	0.192
M11	0.345	0.408	0.302	0.534	0.548	0.429	0.336	0.344	0.289	0.308	0.349	0.244
M12	0.153	0.159	0.128	-0.010	-0.042	-0.009	0.289	0.242	0.258	0.321	0.300	0.262
M13	0.204	0.184	0.171	0.157	-0.024	0.144	0.178	0.140	0.163	0.425	0.417	0.346
M14	0.231	0.249	0.197	0.093	0.039	0.086	0.086	0.042	0.077	0.372	0.415	0.303
M15	0.143	0.143	0.120	0.029	0.207	0.027	-0.014	-0.077	-0.013	0.259	0.287	0.206
M17	0.102	0.143	0.087	-0.024	-0.052	-0.022	-0.098	0.022	-0.092	0.122	0.152	0.100
M20	0.316	0.344	0.271	0.416	0.460	0.329	0.423	0.421	0.347	0.336	0.349	0.264
M22	0.245	0.256	0.209	-0.111	-0.075	-0.107	0.088	0.105	0.081	0.177	0.232	0.151
M23	0.046	0.117	0.042	0.106	0.267	0.101	-0.008	-0.060	-0.008	0.236	0.268	0.194

Table 7: Spearman's (Spear.) correlations, Pearson's (Pear.) correlations, and Kendal's Tau (Kend.) between **RTS** and human scores on the 100 summaries for each model. Values in light gray color are insignificant (p-value $\geq$ 0.05).

	Coh			Con			Flu			Rel		
ID	Spear.	Pear.	Kend.	Spear.	Pear.	Kend.	Spear.	Pear.	Kend.	Spear.	Pear.	Kend.
M8	0.385	0.397	0.318	0.127	0.351	0.121	0.158	0.349	0.146	0.414	0.474	0.359
M9	-0.009	0.049	-0.008	0.132	0.465	0.121	0.222	0.299	0.191	0.257	0.244	0.223
M10	0.459	0.443	0.385	0.070	0.264	0.060	0.024	0.114	0.018	0.323	0.324	0.278
M11	0.321	0.346	0.274	0.308	0.317	0.258	0.241	0.234	0.193	0.357	0.391	0.306
M12	0.285	0.304	0.241	0.051	-0.023	0.050	0.142	0.159	0.132	0.264	0.270	0.231
M13	0.310	0.279	0.256	0.267	0.221	0.257	0.062	0.008	0.057	0.492	0.431	0.424
M14	0.375	0.333	0.317	0.211	0.039	0.202	-0.076	-0.075	-0.068	0.250	0.246	0.218
M15	0.277	0.227	0.226	0.025	0.593	0.023	0.157	0.121	0.144	0.454	0.414	0.394
M17	0.319	0.281	0.276	0.033	-0.038	0.032	-0.025	0.035	-0.024	0.286	0.262	0.254
M20	0.497	0.474	0.396	0.585	0.531	0.493	0.387	0.370	0.323	0.258	0.295	0.215
M22	0.215	0.141	0.189	-0.082	-0.061	-0.081	-0.032	-0.024	-0.032	0.277	0.319	0.246
M23	0.274	0.267	0.241	0.185	0.385	0.181	-0.115	-0.171	-0.109	0.255	0.283	0.227

Table 8: Spearman's (Spear.) correlations, Pearson's (Pear.) correlations, and Kendal's Tau (Kend.) between **MCQ** and human scores on the 100 summaries for each model. Values in light gray color are insignificant (p-value $\geq$ 0.05).

addition, we catch ChatGPT at times to overlook the mistakes or inconsistencies in the summaries, or even generate false statements. Therefore, we believe that ChatGPT with the RTS method shall not be fully trusted.

Regarding the stability of the RTS method, as shown in Table 7, if we only look at the correlation on 100 summaries generated by the same summarization systems, the correlation is only significant for a small subset of the summarization systems. The correlation is only significant for most systems across the dimension of relevance. We also find the correlation of the RTS scores to be system dependent. For instance, correlations on the summaries generated by M11 and M20 are significant across all dimensions, whereas correlations on those generated by M17 are insignificant across all dimensions.

To quantify the stability of the RTS method, we further present the meta-correlation in Table 9. As shown, there exists a significant and high negative meta-correlation for the RTS method, suggesting that as the quality of the summarization system increases, it is increasingly difficult for the RTS method to produce consistent scores with the human expert scores. Therefore, we conclude that the RTS method is still not stable across different summarization systems. With as little as 100 summaries, ChatGPT with the RTS method may not be suitable to be used as a zero-shot evaluator.

5.2 MCQ Results

Given that the ChatGPT tends to penalize the summaries for the wrong reasons with the RTS method, we experiment with the MCQ method that con-

effective reduction of this issues with the expert annotators.

	Coh			Con			Flu			Rel		
ID	Spear.	Pear.	Kend.	Spear.	Pear.	Kend.	Spear.	Pear.	Kend.	Spear.	Pear.	Kend.
RTS	-0.531	-0.476	-0.531	-0.846	-0.762	-0.846	-0.629	-0.622	-0.629	-0.413	-0.476	-0.350
MCQ	-0.259	-0.252	-0.161	-0.734	-0.441	-0.734	-0.671	-0.615	-0.671	-0.140	-0.175	0.028

Table 9: Meta-correlation for RTS and MCQ methods. The correlation is calculated between a) the average human scores for each summarization system, and b) the corresponding correlation between scores produced by each method and the human scores. Values in light gray color are insignificant (p-value $\geq$ 0.05). The smaller the absolute value of the meta-correlation, the more stable the evaluation across different summarization systems.

		Coh			Con			Flu			Rel		
Model A	Model B	A	B		A	B		A	B		A	B	
M22 (4.57)	M23 (4.55)	26 / 42	5 / 35	✓	25 / 8	5 / 3	✓	34 / 11	7 / 12	×	22 / 39	5 / 40	×
M23 (4.55)	M17 (4.52)	10 / 40	17 / 35	×	12 / 7	18 / 9	✓	14 / 7	24 / 16	✓	14 / 40	18 / 36	×
M17 (4.52)	M12 (4.32)	15 / 59	12 / 26	✓	15 / 4	18 / 7	✓	24 / 18	19 / 9	✓	20 / 67	23 / 22	×
M12 (4.32)	M13 (4.24)	12 / 44	10 / 42	✓	20 / 14	13 / 5	✓	16 / 16	16 / 16	✓	12 / 30	13 / 40	✓
M13 (4.24)	M15 (4.19)	13 / 41	11 / 42	×	20 / 6	14 / 14	×	19 / 22	20 / 18	×	16 / 51	13 / 30	✓
M15 (4.19)	M14 (4.12)	16 / 48	13 / 40	✓	21 / 12	15 / 5	✓	24 / 21	16 / 17	✓	24 / 46	17 / 32	✓
M14 (4.12)	M8 (4.07)	16 / 36	13 / 48	×	13 / 18	12 / 9	✓	22 / 19	22 / 26	×	17 / 46	18 / 39	×
M8 (4.07)	M9 (3.77)	41 / 78	5 / 14	✓	49 / 24	15 / 18	✓	54 / 43	6 / 16	✓	41 / 46	9 / 38	✓
M9 (3.77)	M10 (3.70)	18 / 29	24 / 57	✓	23 / 31	35 / 15	×	33 / 25	32 / 36	×	21 / 49	21 / 37	×
M10 (3.70)	M20 (3.58)	27 / 23	11 / 75	×	40 / 53	18 / 25	✓	39 / 45	23 / 22	✓	43 / 49	15 / 40	✓
M20 (3.58)	M11 (3.09)	27 / 80	26 / 16	✓	33 / 47	22 / 41	✓	35 / 53	28 / 36	✓	35 / 47	34 / 41	✓
Success Rate		63.6%			81.8%			63.6%			54.5%		

Table 10: LLM's preference over human preference for H2H evaluation. We use "✓" to indicate both LLM and human prefer the same model, and "×" otherwise. The success rate indicates the percentage of pairs for which both LLM and humans prefer the same model. The model pairs are sorted in descending order according to the average human scores (value in brackets after the model IDs), which we believe to be a challenging but common setting.

strains ChatGPT to only select from a set of predefined reasons for scores from 1 to 5.

As shown in Table 5, now the MCQ evaluation scores are in a much closer range to the human scores except for relevance where the ChatGPT becomes overly optimistic. Nevertheless, as shown in Table 6, for the overall evaluation of 1200 summaries, the MCQ method outperforms the RTS method in terms of correlation with human scores across almost all dimensions. The improvement is especially large in the dimensions of coherence and consistency. It is very likely that the MCQ method improves from the RTS method by preventing the generation of wrong reasons, forcing ChatGPT to only consider the relevant criteria amongst the set of reasons provided for a specific dimension.

Next, Table 8 shows the correlation between the MCQ scores and human scores on 100 summaries generated by each summarization system. It is obvious that with the MCQ method, more systems now obtain significantly correlated ChatGPT evaluations with human scores, with slightly improved correlation values than those of the RTS method. Moreover, the MCQ method's correlation with human scores is generally significant along the dimensions of coherence and relevance. We believe that the above demonstrates that the MCQ method can make ChatGPT evaluations more stable, which is also supported by the smaller meta-correlation values in Table 9.

5.3 H2H Results

We show the results of H2H comparisons in Table 10. We can see that the H2H method is especially effective for ChatGPT in identifying the better model on the dimension of consistency. Even with the challenging setting of comparing models with very close performance, ChatGPT can still identify the more consistent model at a high success rate of 81.8% with only 100 pieces of compared summaries. However, ChatGPT is weak in identifying the more relevant models given the model pairs of close performances. The success rate for relevance is only 54.5%, barely above the chance level. Therefore, we conclude that the H2H evaluation capability of ChatGPT is also dependent on the evaluated dimensions.

6 Discussion and Future Work

Given the above results, we believe that for evaluations on a Likert scale, some evaluations can be better than others. For instance, rather than allowing ChatGPT to freely generate its reasons and scores, researchers may consider defining a specific set of reasons for different score values, then request ChatGPT to output its evaluation in an MCQ format. It may be interesting to investigate if there exist alternative methods that can make ChatGPT perform better, and whether our findings apply to other LLMs.

Moreover, we wish to highlight the potential pitfall of the negative meta-correlation phenomenon observed in this paper. Such an observation suggests that, as the summarization systems quality improves, ChatGPT may have limited evaluation capability. This can limit the usefulness of the ChatGPT as a potential evaluator, especially because of the recent leaps of improvements in the summarization quality of LLMs. It remains questionable whether ChatGPT could be used as a fair and effective out-of-the-box evaluator for summaries generated by LLMs.

On the other hand, ChatGPT seems very effective at selecting the more consistent summary out of the two competitive systems, although its success may not transfer well to other dimensions such as relevance. However, we note that our findings may be limited to the comparison settings with only a small number (i.e., 100) of compared summaries. As also noted by Chiang and Lee (2023), while evaluating a large number of summaries may be very laborious and expensive for humans, it may not be the case for LLMs. One possible future direction of research may be to investigate the minimal number of evaluations required by an LLM to reliably determine if one system is better than the other.

Given our current experiments, we believe that, although ChatGPT may have great potential to act as a zero-shot evaluator for abstractive summarization systems, it is still not ready to replace human evaluations for only a small number of evaluated summaries. If ChatGPT alone is not sufficient, another possible research direction may be to find ways to effectively combine ChatGPT with human evaluation, such that we could not only save time and cost, but also obtains much more reproducible evaluation results.

7 Conclusion

In this work, we explore the potential of using an out-of-the-box LLM as a zero-shot evaluator for abstractive summarization systems, in order to overcome the common disadvantages of traditional evaluation methods. Specifically, we explore the potential of using ChatGPT for summary evaluations on 4 dimensions with various methods, including 2 methods for Likert-scale scoring and 1 method for head-to-head comparisons. We further examine the evaluation stability of ChatGPT across different systems. Extensive experimental results and analysis show that the MCQ-formulated method is much more effective and stable than the direct reason-to-score generation method. For head-to-head evaluations, ChatGPT is much more effective on the consistency dimension than others. We conclude that with only a limited number of evaluations, ChatGPT is still not ready to replace human evaluations. Possible future directions may include investigating new evaluation methods with LLMs, exploring the minimal number of evaluations required for LLMs to behave in a more stable manner, as well as incorporating LLMs with human evaluation to make the process more efficient and reproducible.

Acknowledgements

Yang You is supported by the NUS startup grant and the Singapore MOE Tier-1 grant.

References

Lidong Bing, Piji Li, Yi Liao, Wai Lam, Weiwei Guo, and Rebecca Passonneau. 2015. Abstractive multi-document summarization via phrase selection and merging. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1587–1597.

Moye Chen, Wei Li, Jiachen Liu, Xinyan Xiao, Hua Wu, and Haifeng Wang. 2021. Sgsum: Transforming multi-document summarization into sub-graph selection. In *Proceedings of EMNLP*.

Liying Cheng, Dekun Wu, Lidong Bing, Yan Zhang, Zhanming Jie, Wei Lu, and Luo Si. 2020. Ent-desc: Entity description generation by exploring knowledge graph. In *Proceedings of EMNLP*.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of ACL*.

Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. *Noise reduction in speech processing*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.

Yue Dong, Yikang Shen, Eric Crawford, Herke van Hoof, and Jackie Chi Kit Cheung. 2018. Banditsum: Extractive summarization as a contextual bandit. In *Proceedings of EMNLP*.

Ori Ernst, Avi Caciularu, Ori Shapira, Ramakanth Pasunuru, Mohit Bansal, Jacob Goldberger, and Ido Dagan. 2022. Proposition-level clustering for multi-document summarization. In *Proceedings of NAACL*.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of ACL*.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2020. Summeval: Re-evaluating summarization evaluation. *arXiv preprint arXiv:2007.12626*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of ACL*.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2022. Pal: Program-aided language models. *arXiv preprint arXiv:2211.10435*.

Shen Gao, Xiuying Chen, Piji Li, Zhaochun Ren, Lidong Bing, Dongyan Zhao, and Rui Yan. 2019. Abstractive text summarization by incorporating reader comments. In *Proceedings of AAAI*.

Junxian He, Wojciech Kryscinski, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2022. CTRLsum: Towards generic controllable text summarization. In *Proceedings of EMNLP*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *NeuRIPS*.

MG Kendall. 1938. A new measure of rank correlation. *Biometrika*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *NeurIPS*.

Wojciech Kryściński, Romain Paulus, Caiming Xiong, and Richard Socher. 2018. Improving abstraction in text summarization. In *Proceedings of EMNLP*.

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 ACL*.

Piji Li, Wai Lam, Lidong Bing, and Zihao Wang. 2017. Deep recurrent generative decoder for abstractive text summarization. In *Proceedings of EMNLP*.

Wei Li, Xinyan Xiao, Jiachen Liu, Hua Wu, Haifeng Wang, and Junping Du. 2020. Leveraging graph to improve abstractive multi-document summarization. In *Proceedings of ACL*.

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

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

Yang Liu and Mirella Lapata. 2019. Hierarchical transformers for multi-document summarization. In *Proceedings of ACL*.

Yixin Liu, Ansong Ni, Linyong Nan, Budhaditya Deb, Chenguang Zhu, Ahmed H Awadallah, and Dragomir Radev. 2022. Leveraging locality in abstractive text summarization. In *Proceedings of EMNLP*.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for abstractive text summarization. *arXiv preprint arXiv:2303.15621*.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. *arXiv preprint arXiv:2301.13379*.

Xueguang Ma, Xinyu Zhang, Ronak Pradeep, and Jimmy Lin. 2023a. Zero-shot listwise document reranking with a large language model. *arXiv preprint arXiv:2305.02156*.

Yubo Ma, Yixin Cao, YongChing Hong, and Aixin Sun. 2023b. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! *arXiv preprint arXiv:2303.08559*.

Ani Nenkova and Kathleen McKeown. 2012. A survey of text summarization techniques. *Mining text data*.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with frank: A benchmark for factuality metrics. In *Proceedings of NAACL-HLT*.

Swarnadeep Saha, Shiyue Zhang, Peter Hase, and Mohit Bansal. 2022. Summarization programs: Interpretable abstractive summarization with neural modular trees. In *ICLR*.

Chenhui Shen, Liying Cheng, Lidong Bing, Yang You, and Luo Si. 2022a. Sentbs: Sentence-level beam search for controllable summarization. *EMNLP 2022*.

Chenhui Shen, Liying Cheng, Yang You, and Lidong Bing. 2023. A hierarchical encoding-decoding scheme for abstractive multi-document summarization. *arXiv preprint arXiv:2305.08503*.

Chenhui Shen, Liying Cheng, Ran Zhou, Lidong Bing, Yang You, and Luo Si. 2022b. Mred: A meta-review dataset for structure-controllable text generation. *Findings of ACL*.

Greg Durrett Tanya Goyal, Junyi Jessy Li. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint*.

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. *arXiv preprint arXiv:2303.04048*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. In *ICLR*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed H Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Ning Wu, Ming Gong, Linjun Shou, Shining Liang, and Daxin Jiang. 2023. Large language models are diverse role-players for summarization evaluation. *arXiv preprint arXiv:2303.15078*.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. Primera: Pyramid-based masked sentence pre-training for multi-document summarization. In *Proceedings of ACL*.

Jiacheng Xu and Greg Durrett. 2021. Dissecting generation modes for abstractive summarization models via ablation and attribution. In *Proceedings of ACL-IJCNLP*.

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

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

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020a. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of ICML*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert. In *ICLR*.

Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. 2018. Neural document summarization by jointly learning to score and select sentences. In *Proceedings of ACL*.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.