G-EVAL: NLG Evaluation using GPT-4 with Better Human Alignment

Yang Liu Dan Iter Yichong Xu
Shuohang Wang Ruochen Xu Chenguang Zhu

Microsoft Cognitive Services Research
{***yaliu10***, *iterdan, yicxu, shuowa, ruox, chezhu*}*@microsoft.com*

Abstract

The quality of texts generated by natural language generation (NLG) systems is hard to measure automatically. Conventional reference-based metrics, such as BLEU and ROUGE, have been shown to have relatively low correlation with human judgments, especially for tasks that require creativity and diversity. Recent studies suggest using large language models (LLMs) as reference-free metrics for NLG evaluation, which have the benefit of being applicable to new tasks that lack human references. However, these LLM-based evaluators still have lower human correspondence than medium-size neural evaluators. In this work, we present G-EVAL, a framework of using large language models with chain-of-thoughts (CoT) and a form-filling paradigm, to assess the quality of NLG outputs. We experiment with two generation tasks, text summarization and dialogue generation. We show that G-EVAL with GPT-4 as the backbone model achieves a Spearman correlation of 0.514 with human on summarization task, outperforming all previous methods by a large margin. We also propose analysis on the behavior of LLM-based evaluators, and highlight the potential concern of LLM-based evaluators having a bias towards the LLM-generated texts. [1]

1 Introduction

Evaluating the quality of natural language generation systems is a challenging problem even when large language models can generate high-quality and diverse texts that are often indistinguishable from human-written texts (Ouyang et al., 2022). Traditional automatic metrics, such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005), are widely used for NLG evaluation, but they have been shown to have relatively low correlation with human judgments, especially for open-ended generation tasks. Moreover, these metrics require associated reference output, which is costly to collect for new tasks.

Recent studies propose directly using LLMs as reference-free NLG evaluators (Fu et al., 2023; Wang et al., 2023). The idea is to use the LLMs to score the candidate output based on its generation probability without any reference target, under the assumption that the LLMs have learned to assign higher probabilities to high-quality and fluent texts. However, the validity and reliability of using LLMs as NLG evaluators have not been systematically investigated. In addition, meta-evaluations show that these LLM-based evaluators still have lower human correspondence than medium-size neural evaluators (Zhong et al., 2022). Thus, there is a need for a more effective and reliable framework for using LLMs for NLG evaluation.

In this paper, we propose G-EVAL, a framework of using LLMs with chain-of-thoughts (CoT) (Wei et al., 2022) to evaluate the quality of generated texts in a form-filling paradigm. By only feeding the Task Introduction and the Evaluation Criteria as a prompt, we ask LLMs to generate a CoT of detailed Evaluation Steps. Then we use the prompt along with the generated CoT to evaluate the NLG outputs. The evaluator output is formatted as a form. Moreover, the probabilities of the output rating tokens can be used to refine the final metric. We conduct extensive experiments on three meta-evaluation benchmarks of two NLG tasks: text summarization and dialogue generation. The results show that G-EVAL can outperform existing NLG evaluators by a large margin in terms of correlation with human evaluations. Finally, we conduct analysis on the behavior of LLM-based evaluators, and highlight the potential issue of LLM-based evaluator having a bias towards the LLM-generated texts.

[1] https://github.com/nlpyang/geval

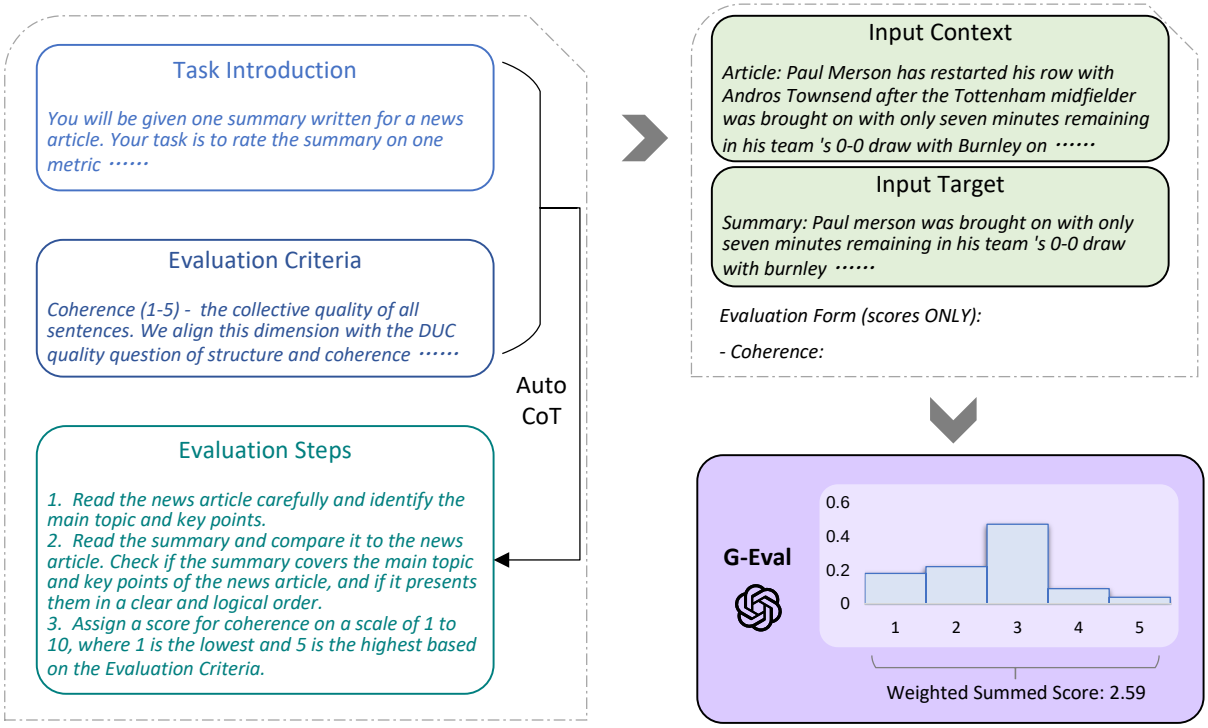

Figure 1: The overall framework of G-EVAL. We first input Task Introduction and Evaluation Criteria to the LLM, and ask it to generate a CoT of detailed Evaluation Steps. Then we use the prompt along with the generated CoT to evaluate the NLG outputs in a form-filling paradigm. Finally, we use the probability-weighted summation of the output scores as the final score.

To summarize, our main contributions in this paper are:

1. LLM-based metrics generally outperform reference-based and reference-free baseline metrics in terms of correlation with human quality judgments, especially for open-ended and creative NLG tasks, such as dialogue response generation.

2. LLM-based metrics are sensitive to the instructions and prompts, and chain-of-thought can improve the performance of LLM-based evaluators by providing more context and guidance.

3. LLM-based metrics can provide a more fine-grained continuous score by re-weighting the discrete scores by their respective token probabilities.

4. LLM-based metrics have a potential issue of preferring LLM-generated texts over human-written texts, which may lead to the self-reinforcement of LLMs if LLM-based metrics are used as the reward signal for improving themselves.

2 Method

G-EVAL is a prompt-based evaluator with three main components: 1) a prompt that contains the definition of the evaluation task and the desired evaluation criteria, 2) a chain-of-thoughts (CoT) that is a set of intermediate instructions generated by the LLM describing the detailed evaluation steps, and 3) a scoring function that calls LLM and calculates the score based on the probabilities of the return tokens.

Prompt for NLG Evaluation The prompt is a natural language instruction that defines the evaluation task and the desired evaluation criteria. For example, for text summarization, the prompt can be:

> *You will be given one summary written for a news article. Your task is to rate the summary on one metric.*
>
> *Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.*

The prompt should also contain customized eval-

uation criteria for different NLG tasks and, such as coherence, conciseness, or grammar. For example, for evaluating coherence in text summarization, we add the following content to the prompt:

> *Evaluation Criteria:*
>
> *Coherence (1-5) - the collective quality of all sentences. We align this dimension with the DUC quality question of structure and coherence whereby "the summary should be well-structured and well-organized. The summary should not just be a heap of related information, but should build from sentence to sentence to a coherent body of information about a topic."*

Auto Chain-of-Thoughts for NLG Evaluation The chain-of-thoughts (CoT) is a sequence of intermediate representations that are generated by the LLM during the text generation process. For evaluation tasks, some criteria need a more detailed evaluation instruction beyond the simple definition, and it is time-consuming to manually design such evaluation steps for each task. We find that LLM can generate such evaluation steps by itself. The CoT can provide more context and guidance for the LLM to evaluate the generated text, and can also help to explain the evaluation process and results. For example, for evaluating coherence in text summarization, we add a line of *"Evaluation Steps:"* to the prompt and let LLM to generate the following CoT automatically:

> *1. Read the news article carefully and identify the main topic and key points.*
>
> *2. Read the summary and compare it to the news article. Check if the summary covers the main topic and key points of the news article, and if it presents them in a clear and logical order.*
>
> *3. Assign a score for coherence on a scale of 1 to 5, where 1 is the lowest and 5 is the highest based on the Evaluation Criteria.*

Scoring Function The scoring function calls the LLM with the designed prompt, auto CoT, the input context and the target text that needs to be evaluated. Unlike GPTScore (Fu et al., 2023) which uses the conditional probability of generating the target text as an evaluation metric, G-EVAL directly performs the evaluation task with a form-filling paradigm. For example, for evaluating coherence in text summarization, we concatenate the prompt, the CoT, the news article, and the summary, and then call the LLM to output a score from 1 to 5 for each evaluation aspect, based on the defined criteria.

However, we notice this direct scoring function has two issues:

1. For some evaluation tasks, one digit usually dominates the distribution of the scores, such as 3 for a 1 - 5 scale. This may lead to the low variance of the scores and the low correlation with human judgments.

2. LLMs usually only output integer scores, even when the prompt explicitly requests decimal values. This leads to many ties in evaluation scores which do not capture the subtle difference between generated texts.

To address these issues, we propose using the probabilities of output tokens from LLMs to normalize the scores and take their weighted summation as the final results. Formally, given a set of scores (like from 1 to 5) predefined in the prompt $S = \{s_1, s_2, ..., s_n\}$, the probability of each score $p(s_i)$ is calculated by the LLM, and the final score is:

$$score = \sum_{i=1}^{n} p(s_i) \times s_i \tag{1}$$

This method obtains more fine-grained, continuous scores that better reflect the quality and diversity of the generated texts.

3 Experiments

Following Zhong et al. (2022), we meta-evaluate our evaluator on three benchmarks, SummEval, Topical-Chat and QAGS, of two NLG tasks, summarization and dialogue response generation.

3.1 Implementation Details

We use OpenAI's GPT family as our LLMs, including GPT-3.5 (text-davinci-003) and GPT-4. For GPT-3.5, we set decoding temperature to 0 to increase the model's determinism. For GPT-4, as it does not support the output of token probabilities, we set '$n = 20, temperature = 1, top\_p = 1$' to sample 20 times to estimate the token probabilities. We use G-EVAL-4 to indicate G-EVAL with GPT-4

Metrics	Coherence		Consistency		Fluency		Relevance		AVG	
	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
ROUGE-1	0.167	0.126	0.160	0.130	0.115	0.094	0.326	0.252	0.192	0.150
ROUGE-2	0.184	0.139	0.187	0.155	0.159	0.128	0.290	0.219	0.205	0.161
ROUGE-L	0.128	0.099	0.115	0.092	0.105	0.084	0.311	0.237	0.165	0.128
BERTScore	0.284	0.211	0.110	0.090	0.193	0.158	0.312	0.243	0.225	0.175
MOVERSscore	0.159	0.118	0.157	0.127	0.129	0.105	0.318	0.244	0.191	0.148
BARTScore	0.448	0.342	0.382	0.315	0.356	0.292	0.356	0.273	0.385	0.305
UniEval	0.575	0.442	0.446	0.371	0.449	0.371	0.426	0.325	0.474	0.377
GPTScore	0.434	–	0.449	–	0.403	–	0.381	–	0.417	–
G-EVAL-3.5	0.440	0.335	0.386	0.318	0.424	0.347	0.385	0.293	0.401	0.320
- Probs	0.359	*0.313*	0.361	*0.344*	0.339	*0.323*	0.327	*0.288*	0.346	*0.317*
G-EVAL-4	**0.582**	**0.457**	**0.507**	**0.425**	**0.455**	**0.378**	**0.547**	**0.433**	**0.514**	**0.418**
- Probs	0.560	*0.472*	0.501	*0.459*	0.438	*0.408*	0.511	*0.444*	0.502	*0.446*
- CoT	0.564	0.454	0.493	0.413	0.403	0.334	0.538	0.427	0.500	0.407

Table 1: Summary-level Spearman (ρ) and Kendall-Tau (τ) correlations of different metrics on SummEval benchmark. G-EVAL without probabilities (*italicized*) should not be considered as a fair comparison to other metrics on τ, as it leads to many ties in the scores. This results in a higher Kendall-Tau correlation, but it does not fairly reflect the true evaluation ability. More details are in Section 4.

as the backbone model, and G-EVAL-3.5 to indicate G-EVAL with GPT-3.5 as the backbone model. Example prompts for each task are provided in the Appendix.

3.2 Benchmarks

We adopt three meta-evaluation benchmarks to measure the correlation between G-EVAL and human judgments.

SummEval (Fabbri et al., 2021) is a benchmark that compares different evaluation methods for summarization. It gives human ratings for four aspects of each summary: `fluency`, `coherence`, `consistency` and `relevance`. It is built on the CNN/DailyMail dataset (Hermann et al., 2015)

Topical-Chat (Mehri and Eskenazi, 2020) is a testbed for meta-evaluating different evaluators on dialogue response generation systems that use knowledge. We follow (Zhong et al., 2022) to use its human ratings on four aspects: `naturalness`, `coherence`, `engagingness` and `groundedness`.

QAGS (Wang et al., 2020) is a benchmark for evaluating hallucinations in the summarization task. It aims to measure the `consistency` dimension of summaries on two different summarization datasets.

3.3 Baselines

We evaluate G-EVAL against various evaluators that achieved state-of-the-art performance.

BERTScore (Zhang et al., 2019) measures the similarity between two texts based on the contextualized embedding from BERT (Devlin et al., 2019).

MoverScore (Zhao et al., 2019) improves BERTScore by adding soft alignments and new aggregation methods to obtain a more robust similarity measure.

BARTScore (Yuan et al., 2021) is a unified evaluator which evaluate with the average likelihood of the pretrained encoder-decoder model, BART (Lewis et al., 2020). It can predict different scores depending on the formats of source and target.

FactCC and **QAGS** (Kryściński et al., 2020; Wang et al., 2020) are two evaluators that measure the factual consistency of generated summaries. FactCC is a BERT-based classifier that predicts whether a summary is consistent with the source document. QAGS is a question-answering based evaluator that generates questions from the summary and checks if the answers can be found in the source document.

USR (Mehri and Eskenazi, 2020) is evaluator that assess dialogue response generation from different perspectives. It has several versions that assign different scores to each target response.

UniEval (Zhong et al., 2022) is a unified evaluator that can evaluate different aspects of text gen-

eration as QA tasks. It uses a pretrained T5 model (Raffel et al., 2020) to encode the evaluation task, source and target texts as questions and answers, and then computes the QA score as the evaluation score. It can also handle different evaluation tasks by changing the question format.

GPTScore (Fu et al., 2023) is a new framework that evaluates texts with generative pre-training models like GPT-3. It assumes that a generative pre-training model will assign a higher probability of high-quality generated text following a given instruction and context. Unlike G-EVAL, GPTScore formulates the evaluation task as a conditional generation problem instead of a form-filling problem.

3.4 Results for Summarization

We adopt the same approach as Zhong et al. (2022) to evaluate different summarization metrics using summary-level Spearman and Kendall-Tau correlation. The first part of Table 1 shows the results of metrics that compare the semantic similarity between the model output and the reference text. These metrics perform poorly on most dimensions. The second part shows the results of metrics that use neural networks to learn from human ratings of summary quality. These metrics have much higher correlations than the similarity-based metrics, suggesting that they are more reliable for summarization evaluation.

In the last part of Table 1 which corresponds to GPT-based evaluators, GPTScore also uses GPTs for evaluating summarization texts, but relies on GPT's conditional probabilities of the given target. G-EVAL substantially surpasses all previous state-of-the-art evaluators on the SummEval benchmark. G-EVAL-4 achieved much higher human correspondence compared with G-EVAL-3.5 on both Spearman and Kendall-Tau correlation, which indicates that the larger model size of GPT-4 is beneficial for summarization evaluation. G-EVAL also outperforms GPTScore on several dimension, demonstrating the effectiveness of the simple form-filling paradigm.

3.5 Results for Dialogue Generation

We use the Topical-chat benchmark from Mehri and Eskenazi (2020) to measure how well different evaluators agree with human ratings on the quality of dialogue responses. We calculate the Pearson and Spearman correlation for each turn of the dialogue. Table 2 shows that similarity-based metrics have good agreement with humans on how `engaging` and `grounded` the responses are, but not on the other aspects. With respect to the learning-based evaluators, before G-EVAL, UniEval predicts scores that are most consistent with human judgments across all aspects.

As shown in the last part, G-EVAL also substantially surpasses all previous state-of-the-art evaluator on the Topical-Chat benchmark. Notably, the G-EVAL-3.5 can achieve similar results with G-EVAL-4. This indicates that this benchmark is relatively easy for the G-EVAL model.

3.6 Results on Hallucinations

Advanced NLG models often produce text that does not match the context input (Cao et al., 2018), and recent studies find even powerful LLMs also suffer from the problem of hallucination. This motivates recent research to design evaluators for measuring the `consistency` aspect in summarization (Kryściński et al., 2020; Wang et al., 2020; Cao et al., 2020; Durmus et al., 2020). We test the QAGS meta-evaluation benchmark, which includes two different summarization datasets: CNN/DailyMail and XSum (Narayan et al., 2018) Table 3 shows that BARTScore performs well on the more extractive subset (QAGS-CNN), but has low correlation on the more abstractive subset (QAGS-Xsum). UniEval has good correlation on both subsets of the data.

On average, G-EVAL-4 outperforms all state-of-the-art evaluators on QAGS, with a large margin on QAGS-Xsum. G-EVAL-3.5, on the other hand, failed to perform well on this benchmark, which indicates that the consistency aspect is sensitive to the LLM's capacity. This result is consistent with Table 1.

4 Analysis

Will G-EVAL prefer LLM-based outputs? One concern about using LLM as an evaluator is that it may prefer the outputs generated by the LLM itself, rather than the high-quality human-written texts. To investigate this issue, we conduct an experiment on the summarization task, where we compare the evaluation scores of the LLM-generated and the human-written summaries. We use the dataset collected in Zhang et al. (2023), where they first ask freelance writers to write high-quality summaries for news articles, and then ask annotators to compare human-written summaries and LLM-generated summaries (using GPT-3.5, text-davinci-