Exploring Prompting Large Language Models as Explainable Metrics

Ghazaleh Mahmoudi
School of Computer Engineering
Iran University of Science and Technology
Gh_Mahmoodi@comp.iust.ac.ir

Abstract

This paper describes the IUST NLP Lab submission to the Prompting Large Language Models as Explainable Metrics Shared Task at the Eval4NLP 2023 Workshop on Evaluation & Comparison of NLP Systems. We have proposed a zero-shot prompt-based strategy for explainable evaluation of the summarization task using Large Language Models (LLMs). The conducted experiments demonstrate the promising potential of LLMs as evaluation metrics in Natural Language Processing (NLP), particularly in the field of summarization. Both few-shot and zero-shot approaches are employed in these experiments. The performance of our best provided prompts achieved a Kendall correlation of 0.477 with human evaluations in the text summarization task on the test data. Code and results are publicly available on GitHub [1].

1 Introduction

Summarization is crucial for quickly understanding large textual documents. The goal of text summarization is to condense lengthy documents into a concise, coherent, and easily understandable format while retaining the essential information from the original source. However, assessing the quality and performance of summarization systems has proven to be a challenging task. Commonly used evaluation methods for summarization, such as ROUGE scores (Lin, 2004), have certain limitations. They fail to capture the overall quality, coherence, and interpretability of summaries. Additionally, they rely on human-generated reference summaries, which are time-consuming and subjective. Given the limitations of traditional evaluation approaches, it is important to explore alternative evaluation methods that leverage the capabilities of LLMs and offer explainable metrics.

LLMs, such as GPT-3 (Brown et al., 2020) and LLaMA (Touvron et al., 2023), have shown remarkable summarization capabilities. They can generate coherent and contextually grounded summaries. This makes them ideal for evaluation purposes. LLMs provide both interpretability and inherent summarization abilities. They can generate explanations and reasoning for their predictions, giving evaluators a deeper understanding of system strengths and weaknesses.

Moreover, LLMs reduce the dependency on gold-standard reference summaries. By using LLMs as evaluators, we can generate comparative summaries and objectively assess system-generated summaries. This eliminates potential biases from human references.

In summary, using LLMs as explainable metrics in summarization evaluation offers several benefits. It overcomes the limitations of traditional methods, provides interpretability, and reduces reliance on human-generated references. This emerging field of research holds promise for a more comprehensive and objective assessment of summarization systems.

The main contribution of this paper is the investigation of various prompt-based methods for explainable evaluation of summarization tasks. We explore both few-shot and zero-shot approaches in our experiments. The best performance prompt follows the zero-shot strategy and is introduced in the paper under the name P1. In this prompt, the criteria for assessing the quality of summaries are described (e.g., how well the main idea of the main document is captured in the summary). This prompt achieves a Kendall correlation score of 0.477, outperforming other methods in comparison. Our conducted experiments highlight the promising potential of LLMs as evaluation metrics in the field of NLP, with a specific focus on summarization.

[1] https://github.com/ghazaleh-mahmoodi/Prompting_LLMs_AS_Explainable_Metrics

Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems, pages 219–227
November 1, 2023 ©2023 Association for Computational Linguistics

2 Related Work

Several recent studies have focused on utilizing LLMs for the evaluation of several different tasks in NLP (e.g., text generation, machine translation, and summarization).

GPTScore (Fu et al., 2023) is a novel framework for evaluating generated texts using large pre-trained language models, particularly GPT-3. The framework leverages the emergent abilities of generative pre-trained models, such as zero-shot instruction, to score generated texts. GPTScore operates under the assumption that a large pre-trained language model is more likely to assign higher probabilities to high-quality generated text when provided with adequate instruction and context. By leveraging the power of GPT-3, GPTScore aims to assess the quality of generated text based on the model's generative capabilities.

In a similar vein, Wang et al. (2023) conducted a preliminary survey on using ChatGPT, a variant of the GPT model, as a natural language generation (NLG) evaluator. The study explores the potential of ChatGPT in evaluating the quality of generated text in various NLG tasks.

In the context of translation quality assessment, GEMBA (Kocmi and Federmann, 2023) is introduced as a GPT-based metric that can effectively evaluate translations with or without a reference translation. The evaluation focuses on zero-shot prompting and involves comparing four prompt variants in two modes, depending on the availability of a reference. Results from the evaluation demonstrate that GEMBA achieves state-of-the-art accuracy when compared to MQM-based human labels, as evidenced by the WMT22 Metrics shared task.

Instructscore (Xu et al., 2023) is an open-source and explainable evaluation metric for text generation. This model fine-tunes the LLaMA model to predict a fine-grained error diagnosis of machine translated content. This work presents a novel framework for explainable text generation evaluation, addressing the limitations associated with black-box metrics and showcasing the potential of LLMs to provide meaningful and interpretable evaluations.

G-Eval (Liu et al., 2023), is a framework that utilizes LLMs with chain-of-thoughts (CoT) and a form-filling paradigm to assess the quality of NLG outputs, specifically in text summarization and dialogue generation tasks. The experiments demonstrate that G-Eval, utilizing GPT-4 as the backbone model, achieves a high Spearman correlation of 0.514 with human evaluations in the text summarization task, outperforming previous methods significantly.

3 Task Description

The topic of Eval4NLP shared task (Leiter et al., 2023) is to provide explainable metrics for summarization and machine translation evaluation by prompting LLMs. The main goal is to investigate different prompting methods (e.g., zero-shot, few-shot, Chain of Thought, Fine-Grained, Majority Vote, Self-Refinement), therefore, **fine-tuning the LLMs is not allowed**. Also, a number of LLMs are allowed to be used. The shared task has two tracks based on the model sizes (One for models bigger than 25B parameters, and one for smaller models).

This work has been done on the summarization task and using small models. In the following, the dataset and the evaluation metric are explained.

3.1 Data

The Shared Task organizers opted for SummEval during the training and development phase for summarization. (Fabbri et al., 2021) introduced SummEval as an evaluation benchmark, aiming to compare various methods for assessing summarization. This benchmark entails the assignment of human ratings on four key dimensions of every summary, including `fluency`, `coherence`, `consistency`, and `relevance`. SummEval draws upon the renowned CNN/DailyMail dataset proposed by (Hermann et al., 2015) for its construction.

Furthermore, in the testing phase, a new reference-free dataset with summary-level quality scores is collected for summarization. As source data, sentences and paragraphs were collected from English Wikipedia pages created after 15.07.2023. Test-phase scores are constructed from fine-grained scores.

3.2 Evaluation Metrics

To determine how well LLMs explainable metrics correlate with humans, We follow the evaluation protocol of the WMT22 metrics shared task. we use Kendall's Tau correlation (Freitag et al., 2022). In addition to Kendall correlation, Pearson (Mukaka, 2012) and Spearman (Zar, 2005) are also used in the test phase.

4 Methodology

In this research, We have used pre-trained orca_mini_v3_7b (tuned Llama2-7b model) (Mathur, 2023) on the HuggingFace Transformers[2]. We employed two strategies, zero-shot and few-shot, for constructing prompts.

The zero-shot strategy included the combination of evaluation criteria for the quality of summarization in the form of questions or detailed explanations provided to the model.

There are examples of summarization evaluation written in **the few-shot strategy**. In this way, the main document, the summarized document, and the scores received are mentioned precisely.

To assess summarization quality via prompting an LLM, the following parameters are needed:

- Prompt variant (from a pre-defined set)
- Main document `Source Text`
- Summary `Summary`

4.1 Prompt variants

For **P1** (Table 2), we formulated the main criteria for assessing summary quality, which were originally expressed by humans. In this prompt, the following items are mentioned to be checked:

- Comparing the key points.
- Capturing the main ideas.
- Score on a continuous scale from 0 to 100.
- Meaning of zero score: irrelevant, factually incorrect, and not readable summary.
- Meaning of a hundred score: relevant, factually correct, good readability summary.
- Explain the result.

To create prompt **P2**(Table 3), we consulted the ChatGPT4 Bot and asked what questions would be relevant for evaluating summarization. We then modified and adapted the generated questions accordingly. In this prompt, in addition to the items mentioned in P1, the following items have been added in the form of questions.

- The overall length of the summary. Concise representation of the original document.
- Grammatical accuracy and fluency of the summary.
- Evaluate The ranking of information in the summary.
- Analyze the level of abstraction in the summary.
- Contextual understanding is exhibited by the summary.

Prompts P1 and P2 also include an Explanation for the model's output score, thus containing questions that aid in better understanding the received score's reasoning.

Prompt **P3**(Table 4) and **P4**(Table 5) are similar to the P1 prompts, and only the wording and the way of expression have changed.

In **P5** (Table 6), we guide the model to calculate the desired score by calculating the similarity of the main and summarized documents. P5 includes examples of how one can calculate the similarity of two documents.

Prompt **P6** (Table 7) follows the few-shot strategy, where two examples consisting of the main document, and the written summary, along with their respective score, are included in the input prompt.

5 Results and Analysis

We experiment with six different distinct prompt types. One of them is few-shot (P6) and the rest are zero-shot. Table 1 shows results for all prompt variants we have experimented with.

	Kendall	Pearson	Spearman
P1	**0.477**	0.495	**0.619**
P2	0.470	0.468	0.607
P3	0.472	0.498	0.612
P4	0.467	0.504	0.610
P5	0.454	**0.543**	0.589
P6	0.283	0.513	0.376

Table 1: Test phase Segment-level Kendall (τ) and Pearson (ρ) and Spearman correlation.

The execution of each prompt takes approximately 1 hour. If we also include the explanation of the results in the output, each execution of the test data takes close to 4 hours.

Based on the Kendall measure (which serves as the primary evaluation metric), the best result is associated with P1. This prompt follows the zero-shot strategy. In this prompt, some of the SummEval criteria are also mentioned. Additionally, P1 achieves the highest value in the Spearman measure and serves as the final strategy for the test phase.

The results of P2, P3, and P4 are very close to each other. The reason for the difference observed

[2] https://huggingface.co/pankajmathur/orca_mini_v3_7b

Prompt P1:
Score the effectiveness of the summarization by **comparing the key points and overall coherence** of the summarized with the main document.

Checked whether the summary **captures the main ideas, maintains the intended tone and style, and provides a concise yet comprehensive overview** of the main document.

Score the summarization **with respect to the summarized document** on a **continuous scale from 0 to 100**, where a score of zero means `irrelevant, factually incorrect and good readable` and score of 100 means `relevant, factually correct, no readability` summarized.

Also explain your process to get this score to summary.

Also please perform error Analysis of given summary.

What should we change to have a better result?",

main document: {main document},

Summary: {Summary}",

Score: *gen 'score' pattern='(100|[1-9]?[0-9])'*,

Explanation: *gen 'explanation'*

Table 2: The best-performing prompt based on zero-shot prompting strategy expecting a score between 0–100.

is the variation in the way the evaluation method is expressed. In this regard, it can be said that LLMs are sensitive to manner of expression. However, considering the proximity of the Kendall output value, it can be concluded that they have a low sensitivity to the mentioned changes.

Furthermore, considering the results of P5, it can be stated that introducing scientific methods for examining the similarity between summaries and the main document did not effectively guide the model. Instead, criteria such as "captures the main ideas" yielded better results.

Contrary to our expectations, P6 (few-shot approach) obtains the lowest score in the Kendall measure. We expected that the few-shot strategy would outperform zero-shot since it allows the model to observe multiple instances of scoring, thereby enhancing its capabilities. However, our experiments yielded results contrary to this assumption. There may be several reasons for this result. It is possible that a larger number of input samples would have been beneficial. Furthermore, the quality of the input samples might not have been sufficient for the model to comprehend the problem-solving process fully.

In conclusion, based on the obtained results, it can be inferred that by explicitly defining evaluation metrics, language models can be utilized as an interpretable method for evaluating the summarization task.

6 Conclusion

In this paper, we have investigated different prompts to define explainable evaluation metrics for summarization Using LLMs.

The experiments conducted indicate that LLMs have great potential as evaluation metrics in NLP tasks, especially summarization. In these experiments, both the few-shot and zero-shot approaches were used. Our best prompt achieved a Kendall correlation of 0.477 compared to the human score.

In future work, other prompt strategies, such as the Chain of Thought, can also be explored. Experiments can also be repeated with existing prompts and other Language Models and compare the results obtained to determine the effect of the Language Model on changing the quality of the output.

7 Limitations

Due to hardware limitations, we were unable to investigate other eligible models in this series of experiments. In future research, it would be beneficial to examine the impact of other models on the introduced prompts more extensively. Furthermore, the lack of fine-tuning the model on the defined tasks may have an effect on its performance. In future research, by fine-tuning the model, we can explore its impact on improving the output quality.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

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.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Christoph Leiter, Juri Opitz, Daniel Deutsch, Yang Gao, Rotem Dror, and Steffen Eger. 2023. The eval4nlp 2023 shared task on prompting large language models as explainable metrics. In *Proceedings of the 4th Workshop on Evaluation and Comparison for NLP systems*.

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.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment.

Pankaj Mathur. 2023. orca_mini_v3_7b: An explain tuned llama2-7b model. `https://https://huggingface.co/psmathur/orca_mini_v3_7b`,.

MM Mukaka. 2012. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi medical journal : the journal of Medical Association of Malawi*, 24(3):69—71.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

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

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Yang Wang, and Lei Li. 2023. Instructscore: Towards explainable text generation evaluation with automatic feedback.

Jerrold H. Zar. 2005. *Spearman Rank Correlation*. John Wiley Sons, Ltd.

A Appendix: Prompt Templates

Below, we present our prompt templates utilized in the described experiments in this paper.

Prompt P2:
Score the effectiveness of the summarization by **comparing the key points and overall coherence** of the summarized with the main document.

Checked whether the summary **captures the main ideas, maintains the intended tone and style, and provides a concise yet comprehensive overview** of the main document.

Score the summarization **with respect to the summarized document** on a **continuous scale from 0 to 100**, where a score of zero means `irrelevant, factually incorrect and no readability` and score of 100 means `relevant, factually correct, good readable` summarized.

To calculate Score, first answer the following questions.
Then, according to the answers to the questions, scored the quality between 0 and 100.
1. Assess the overall length of the summary. Does it provide a concise representation of the original document without omitting important information?
2. Examine the grammatical accuracy and fluency of the summary. Are the sentences well-structured, free of errors, and coherent?
3. Evaluate the ranking of information in the summary. Are the most salient and crucial details given appropriate emphasis and positioned prominently?
4. Analyze the level of abstraction in the summary. Does it effectively distill complex ideas and concepts into more accessible and simplified language?
5. Consider the contextual understanding exhibited by the summary. Does it demonstrate an understanding of the original text beyond simple keyword extraction?

Also explain your process to get this score to summary.

Also please perform error Analysis of given summary.

What should we change to have a better result?",

main document: {main document},

Summary: {Summary}",

Score: *gen ’score’ pattern=’(100|[1-9]?[0-9])’*,

Explanation: *gen ’explanation’*

Table 3: Prompt P2

Prompt P3:
Your Task is to score the Samaritan quality. The original document is collected from English Wikipedia pages created after 15.07.2023.

Score the effectiveness of the summarization by comparing the key points and overall coherence of the summarized with the main document.

"Checked whether the summary captures the main ideas, maintains the intended tone and style, and provides a concise yet comprehensive overview of the main document.

Score the summarization **with respect to the summarized document** on a **continuous scale from 0 to 100**, where a score of zero means `irrelevant, factually incorrect and not readable` and score of 100 means `relevant, factually correct, good readability, grammatical correctness, covers the main topic and key points of the main document article`

Source text: {main document},

Summary: {Summary}",

Score: *gen ’score’ pattern=’(100|[1-9]?[0-9])’*,

Table 4: Prompt P3

Prompt P4:
Score the effectiveness of the summarization by comparing the key points and overall coherence of the summarized with the main document.

Checked whether the summary captures the main ideas, maintains the intended tone and style, and provides a concise yet comprehensive overview of the main document.

Score the summarization with respect to the summarized document, on a continuous scale from 0 to 100.

Source text: {main document},

Summary: {Summary}",

Score: *gen ’score’ pattern=’(100|[1-9]?[0-9])’*,

Table 5: Prompt P4

Prompt P5:
Score the summarization with respect to the summarized document on a **continuous scale from 0 to 100**, where a score of zero means `irrelevant, factually incorrect and not readable` and score of 100 means `relevant, factually correct, good readability`.
let’s think step by step.
In other words, this Score should show the similarity between the main document and the summarized document.
For similarity measurement, It’s possible to compare the main and summarized document with a similarity measure such as Cosine Similarity.
word2vec, Bert embedding or n-gram are some of the approaches to calculate similarity.
Source text: {main document},

Summary: {Summary}",

Score: *gen ’score’ pattern=’(100|[1-9]?[0-9])’*,

Table 6: Prompt P5

Prompt P6:
Consider these example that summarization is graded in scale 0 - 100.

1. **Source text:** Usain Bolt will compete at the IAAF/BTC World Relays in the Bahamas next month, the Jamaica Athletics Administrative Association has announced. The six-time Olympic gold medallist will compete at the relay championship on May 2 and 3 as part of the Jamaican team. 'I'm happy to be part of the Jamaican team for the IAAF / BTC World Relays in the Bahamas. I am fit, healthy and ready to run,' said Bolt. Usain Bolt has confirmed he will be part of Jamaica's team at the World Relays in the Bahamas Bolt reacts as he wins 4x100m gold at the London Olympic Games in 2012 'I hear the meet was a lot of fun last year and there was a great atmosphere. Jamaica has a long and successful tradition in relays and when we put on the national colours we always do our best to make the country proud,' he added. JAAA General Secretary Garth Gayle commented, 'We were extremely pleased that Usain was available for selection and that the world's fastest man will be running for Jamaica. We can expect some sprint magic on the track in the Bahamas on 2nd and 3rd May.' The full Jamaican team list for the competition will be announced shortly. Bolt insists he always does 'his best to make his country proud' while wearing Jamaica colours.

1. **Summary:** Jamaican sprinter Usain Bolt has confirmed he will be part of the Jamaican team at the IAAF/BTC World Relays in the Bahamas.
1. **Score : 95,**

2. **Source text:** Referee Mark Clattenburg has been named to take charge of the Manchester derby on Sunday, despite having sent off three players from United and City this season. City captain Vincent Kompany was dismissed for two bookable offences during Belgium's narrow 1-0 defeat of Israel in their Euro 2016 qualifier on March 31, meaning he is now suspended for the match against Wales in June. And, although Clattenburg has been accused of favouring Louis van Gaal's side in the past, it's worth noting that the 40-year-old has only sent off two players season in the Premier League this season and both have been from United; Tyler Blackcett in the 5-3 defeat by Leicester and Luke Shaw in the 1-1 draw with West Ham. Mark Clattenburg will officiate the Manchester derby between United and City at Old Trafford The English referee sent off City and Belgium captain Vincent Kompany during the international break Leicester 5-3 Manchester United West Ham 1-1 Manchester United Manchester United 3-0 Tottenham Manchester City 3-1 West Ham Liverpool 2-1 Manchester City Chelsea 1-1 Manchester City Clattenburg has courted controversy during his career but is generally regarded as one of the Premier League's leading referees alongside Michael Oliver. The champion's shock 2-1 loss to Crystal Palace on Monday saw United move a point above their local rivals to add extra incentive for both sides ahead of the derby at Old Trafford, which could ultimately decide who finishes second behind expected winners Chelsea. While Manuel Pellegrini's side have struggled since the turn of the year, turning from title challengers to fourth place chases, United are coasting on confidence having won their last five consecutive league games. Clattenburg will be joined on Sunday by assistants Simon Beck and Jake Collin, while Jonathan Moss will serve as the fourth official. Clattenburg has shown only two red cards this season, both to United players including Luke Shaw (centre).

2. **Summary:** United's win over Liverpool was their first league win since the 3-0 win over Leicester on March 31 City's win over West Ham was their first league win since the 3-0 win over Chelsea on March 31 Manchester City's win over West Ham was their first league win since the 3-0 win over Chelsea on March 31 Manuel Pellegrini's side are top of the Premier League table, four points clear of Chelsea, who have a game.
2. **Score : 26.666666666**

following these examples, please score the following input.

Source text: {main document},

Summary: {Summary}",

Score: *gen 'score' pattern='(100|[1-9]?[0-9])'*,

Table 7: Prompt P6