Limitations of Large Language Models as Automatic Evaluators

Rickard Stureborg[1,2] **Dimitris Alikaniotis**[1] **Yoshi Suhara**[3,*]
[1]Grammarly [2]Duke University [3]NVIDIA
rickard.stureborg@duke.edu
dimitrios.alikaniotis@grammarly.com
ysuhara@nvidia.com

Abstract

The zero-shot capability of Large Language Models (LLMs) has enabled us to use LLMs as reference-free automatic evaluation metrics. Existing studies have shown that LLMs can be high-quality evaluators for various NLP tasks such as summarization. However, little is known about the robustness of LLM evaluators, as the existing work has focused on pursuing the best performance of LLM evaluators with respect to correlations between LLM scores and human expert scores. In this paper, we conduct a series of analysis using the SummEval dataset and report that LLM evaluators for text summarization are sensitive to prompt differences that are trivial to human understanding of text quality. This includes the rating scale itself, scores assigned to previous dimensions of analysis biasing future scores in the same generation, bias towards lower-perplexity summaries, and reliance on features that are uncorrelated the true summary quality (such as worsened performance on rating Fluency of a summary when the source document is not included). We share recipes for how we should configure LLM evaluators while clarifying the limitations, resulting in significantly better performance than G-Eval-style evaluation on the SAMSum partition of the RoSE dataset.

1 Introduction

A core limitation of automatic evaluation is in developing new metrics and scaling them beyond limited benchmark datasets, primarily due to their common reliance on reference outputs. While there is a line of work in reference-free automatic evaluation metrics, it is known that it is less reliable than the current reference-based metrics (Fabbri et al., 2021; Deutsch et al., 2022). To address the limitation, Large Language Models (LLMs) have proven useful in this domain due to their demonstrated high natural language understanding abilities and performance at adhering to instructions. For example, LLMs can be used for aspect-based evaluation in summarization, which is considered to require human annotators for scoring (Fabbri et al., 2021; Liu et al., 2023c). LLM evaluators have become part of automatic evaluation for commonly used benchmarks for LLMs (Zheng et al., 2023).

However, little is known about the abilities of these LLM evaluators. A few studies have looked deeper into this point (Wang et al., 2023; Zheng et al., 2023; Liu et al., 2023b); there is a need for further analysis into potential risks and failure points when using them, especially if used in sensitive applications. Therefore, in this paper, we aim to study two important characteristics of the LLM evaluator, namely *bias* and *consistency*, in order to understand and share the limitations of LLM evaluators. To this end, we conduct extensive experiments using GPT-3.5 and GPT-4, which are commonly used as LLM evaluators, with various prompts and generation configurations on the summarization evaluation benchmarks SummEval and RoSE datasets.

Throughout analyzing these issues, we compiled findings into a set of recipes for LLM evaluators. Experiment results on the RoSE dataset show that our new LLM evaluator shows better performance on the RoSE dataset (Liu et al., 2023c).

2 Methodology

2.1 Datasets

To investigate the performance of LLM-based evaluators, we test predictions on two main datasets. We use SummEval (Fabbri et al., 2021) as our development set, perform extensive analyses of LLM-based evaluators on this set, and then use RoSE (Liu et al., 2023c) as a test set after selecting hyper-parameters of our LLM evaluator system through the first analyses.

[*] Work done while at Grammarly.

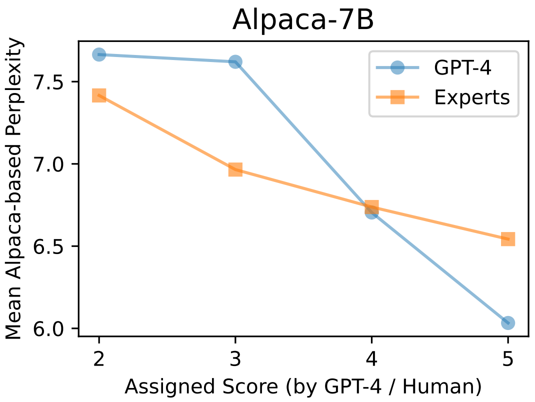

Figure 1: **Average Perplexity by Assigned Score**

2.2 Evaluation Metrics

The goal of automatic evaluation is to provide scores highly correlated with human judgments on the task at hand. In our work, we primarily measure this through Kendall's τ correlation on scores produced for each label in SummEval (Coherence, Consistency, Fluency, Relevance) and RoSE (ACU).

3 Results and Analysis

3.1 Perplexity Bias

Summaries are first grouped by evaluation scores (as assigned either by Experts or an LLM evaluator). Perplexities are then computed with Alpaca-7B, Llama-2, and GPT-2 on the summary text, and a mean score is calculated for each group of summaries.

GPT-4 is disproportionately biased towards low perplexity summaries as compared with expert annotators. The mean perplexities of high assigned scores (5s) are lower than for expert raters, and mean perplexities of low assigned scores (2s and 3s) are lower. An implication of this is that LLM evaluators may show a biased preference for text which is generated by itself.

3.2 Scoring Granularity

A common scale for scoring is 1-5. However, when producing scores for automatic evaluation, ties between candidate examples are often undesirable. To reduce ties, we aim to increase scoring granularity: the distinct number of possible scores for candidate responses. One solution is to increase the range of scores to 1-100.

However, we find that models sparsely predict scores within the range. Frequencies of some scores, such as 90 and 95, are far higher than 'odd' scores such as 92 or 19, and much of the range is almost entirely ignored (1-60).

4 Case Study

Using the lessons learned from our analysis (§3), we find that deterministic, non-CoT prompting with a 1-10 score granularity and using GPT-4 performed the best. We compare our best method for solving the issues raised in Section 3, as determined by performance on the SummEval dataset. Our method outperforms modified G-Eval on both of the out-of-domain test sets, SAMSum and XSum. We note that for XSum, this difference is not statistically significant.

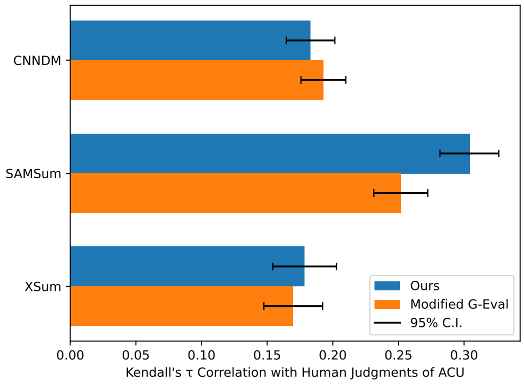

Figure 2: **Performance Comparison of Modified G-Eval versus 1-10 scoring on the RoSE benchmark**. Our approach performs statistically significantly better on the SAMSum dataset partition, while we only perform worse (not significant) on the CNNDM partition.

5 Conclusion

Our findings show that LLM evaluators are disproportionately biased towards low perplexity summaries than is helpful, they fail to respect scoring scales given to them when attempting to increase the granularity of scores, and they are inconsistent their own judgements depending on settings such as inclusion of source documents or generating multiple judgements at the same time. In attempts to solve some of these issues, we find that we are able to significantly outperform a modified version of G-Eval (Liu et al., 2023a) on the SAMSum partition of RoSE annotations with with 95% confidence. Our work suggests that more effort should be allocated towards understanding and remedying the issues exhibited by LLM evaluators.

Limitations

Reliance on GPT-based models. We experiment primarily on GPT-based, proprietary models from OpenAI due to their SOTA performance on automatic evaluation of text summarization. However, this means it is unclear how well our results generalize to other LLMs such as Llama-2, Vicuna, Alpaca, etc. Do to constraints in time and budget, extending the analysis to investigate other LLMs was not possible during the time this work was carried out. This project involved generating more than 560,000 outputs from OpenAI models; repeating the experiments on several models amounts to substantial effort and resources. Future work could aim to replicate and extend our analysis to further models.

Reliance on SummEval for analysis. Our analysis section primarily investigates issues by measuring performance of various model and prompt configurations against SummEval. There is a risk that our results to do generalize well beyond For this reason, we also sought to measure performance on the RoSE benchmark, which is comprised of three datasets in different domains. We find that addressing the issuess seen in SummEval significantly improves performance on one of the domains, and has insignificant but positive results on the other domains.

Limited solutions. Although we investigate solutions to some of the identified issues in this paper, many remain to be studied and may provide the research community with directions for future research efforts. LLM's inconsistencies and biases as automatic evaluators is tough to build solutions around. There is ample opportunity for creative solutions, and while our work offers some, its main focus is in identifying the existing issues in the first place.

Ethics Statement

All datasets utilized in this study are well-established and widely accessible. There are no ethical considerations pertaining to issues such as privacy, consent, or the use of human or animal subjects in our research.

Acknowledgments

We thank the Grammarly research team for their comments and feedback.

References

Daniel Deutsch, Rotem Dror, and Dan Roth. 2022. On the limitations of reference-free evaluations of generated text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10960–10977, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

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

Yiqi Liu, Nafise Sadat Moosavi, and Chenghua Lin. 2023b. LLMs as narcissistic evaluators: When ego inflates evaluation scores.

Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023c. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4140–4170, Toronto, Canada. Association for Computational Linguistics.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.