AN EVALUATION OF LLM- BASED EVALUATOR (STORY WRITING) *

ABSTRACT

LLMs have ushered in a new era of large scale production and public dissemination of natural language text. This necessitates the rise of LLM-based evaluators on the consumer side for tasks such as filtering/ quality checks etc. In this paper, we evaluate the LLM-based evaluator scores for alignment with human expert ratings. The task considered here is creative writing and LLM-based evaluation of LLM responses to mathematical reasoning datasets.

***K*eywords** LLM-based Evaluator · Alignment · Creative Task ·

1 Introduction

Content glut is an inevitable consequence of the ongoing Gen-AI revolution. This pushes for LLM-based evaluators on the consumer-side for filtering/ quality checks etc. Many recent works [RLAIF] explore the usage of LLMs for evaluation tasks. Self-evaluation is a widely explored component of many of the state-of-the-art prompting techniques.

Our contributions:

1. **Experiments to analyse the alignment between LLMs and human experts on the perception of creativity**
   - **Ablation studies concerning independent/ dependant evaluation and provision of related study reference guide.**
   - Compare **Thought/ Reason –> Answer** Versus **Answer –> Reason/ Thought**
   - **Experiments to study if there is a hidden implicit bias for an LLM towards content generated by itself/ same model**
2. **We propose an agentic framework tasked with auto-evaluation of LLM reponses on mathematical reasoning datasets.**

2 Experimental Setup

2.1 Primer

In [1] 12 LLMs were provided a creative writing task with the following input prompt in a zero-shot setting:

Write an epic narration of a single combat between Ignatius J. Reilly and a pterodactyl, in the style of John Kennedy Toole.

5 stories were sampled from each LLM and 5 human written stories were added resulting in a total of 65 stories. In this paper, we evaluate this collection of stories using LLM-Evaluators. Further, our objective to evaluate the alignment of the LLM-evaluations with human expert ratings. In [1], 10 expert human raters were involved in the evaluation process. We use these ratings to set the human evaluation standards. For more details regarding the dataset creation process, please refer to the paper.

* *Citation*: **Authors. Title. Pages.... DOI:000000/11111.**

2.2 Replication of the Evaluation Environment

We need to replicate the evaluation setup provided to the human evaluators for the LLM-based evaluators. Here are few key setup details that we considered to ensure this:

2.2.1 Calibration: Dependant Evaluation

Human Evaluator Setting: Raters were sent all stories at once and they were free to go back and change the ratings of previously-rated stories.

Analogous LLM-Evaluator Setting: We provide all the stories within a group in a single prompt for evaluation. This also reduces calibration issues.

2.2.2 Context Information

Human Evaluator Setting: We provided raters with detailed information about the plot, setting, imagery, tone, characters, main protagonist, and derivative/imitative style of the author, taken from a generic and popular study guide $Link$.

2.3 Evaluation

There are 5 groups with 13 stories each. We share all the 13 stories within the same prompt to get the evaluation scores. For each of the 13 stories in a group, we sample scores 5 times.

3 Questions worth Exploring

Why do we need to run these experiments?

Some other topics on which we hope to get idea on after running these experiments?

- With and without guide
- Reasoning first versus reasoning last
- Is the overall ordering of LLMs by performance made by LLM same as that made by humans.
- Ablation studies concerning independent/ dependant evaluation and provision of related study reference guide.
- Do LLMs have a preference for content generated by itself?
- Is human rating of creativity aligned with the perception of creativity for LLMs? Comparing the rubrics score like does llm percieved humour match with those of human ratings for creativity etc?
- Can random reordering of the items before evaluation affect its scores?

Analogous LLM-Evaluator Setting: We provide the same guide referenced in the paper to the LLM-evaluators via the prompt.

4 Experimental Results

For Anthropic Claude-3.1 model. For each group, we sample 5 independant evaluations and report the average value. We also report the variance amongst the 5 sample evaluations.

Comparison on overall rating by humans vs Comparison on overall rating by Cluade-3.1 Opus

References

[1] Carlos Gómez-Rodríguez and Paul Williams. A confederacy of models: a comprehensive evaluation of LLMs on creative writing. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14504–14528, Singapore, December 2023. Association for Computational Linguistics.