LEVERAGING LARGE LANGUAGE MODELS AS LABELERS FOR OBJECTIVE EVALUATIONS *

Archie Chaudhury, Garrett Allen
LayerLens
{ac, ga}@layerlens.ai

ABSTRACT

Consistent benchmarking and evaluation are becoming more and more crucial to the processes of fine-tuning, post-training, and monitoring foundational models. Traditionally, objective evaluation datasets consist of a set of prompts, a corresponding set of ground truths, and discrete evaluation logic to compare the response generated by the model to the ground truth. Contemporary evaluations can also be subjective: here, an external reviewer manually rates the response for correctness. In the latter, utilizing an external, third-party LLM rather than a human judge to rate the responses has become increasingly common. However, this often results in some common issues, where the LLM-based judges often fail to rate the responses in an unbiased and objective manner. In this work, we propose utilizing LLMs not as judges for subjective evaluations, but rather as labelers for objective datasets, scoring model-generated text for metrics such as instruction following, readability, and toxicity. We find that utilizing LLMs as scorers for datasets with objective criteria allows for greater efficiency during parsing, while also allowing for subjective analysis of individual responses to prompts.

***K*eywords** LLMs · Evaluations · Judges

1 Introduction

Even before the large scale commercial and enterprise adoption of foundational models that coincided with the release of ChatGPT in 2022, evaluations and benchmarking had become a critical aspect of the development of text models. Datasets such as Massive Multitask Language Understanding (MMLU), which is a test to measure the ability for a text model to answer generalized questions in different subject areas [1], and HumanEval, which measures the ability for a model to generate workable code [2], have allowed for the generalized comparison of models across objective metrics. However, getting an accurate and nuanced measurement of a model's capabilities has become increasingly difficult with traditional datasets: most of them are thoroughly delineated from the majority of contemporary use-cases, and their parsing guidelines often mean that correct answers are marked incorrect: for example, an answer of "4" for a mathematics question may be marked incorrect because the expected answer was "Answer: 4". [3]

Subjective evaluations are evaluations that require manual labeling and review; perhaps the most popular implementation of this is ChatBot Arena, which is a public application that asks users to directly compare the responses of two different text models by ranking them against one another. [4] However, subjective evaluations often suffer from significant bias, and require collecting feedback from a disproproptiately large set of humans. An alternative approach that has gained popularity is utilizing an independent language model to evaluate the correctness of a response, according to some predefined criteria. Popular evaluations that leverage this methodology, popularly termed "LLM as a Judge" include WildBench, MtBench, and more. [5], [6] However, these methodologies often also suffer from the same problem, with models showcasing similar bias as their human counterparts, such as preferring responses that were generated by the same family of models or overrating the first response when scoring two responses head to head.

* *Citation*: **Chaudhury, Archie. Allen, Garrett Leveraging Large Language Models as Labelers for Objective Evaluators.**

In this initial work, we explore using Large Language as objective evaluators. Rather than utilizing model judges to rate responses in a subjective fashion, we prompt them with specific instructions centered around an existing static evaluation dataset, specifically centered around determining the agreement of a response with certain metrics. We test this approach by utilizing Google's Gemini model as a judge [7], asking it to determine the correctness of a response based on the evaluation instructions, and rating it on external metrics such as readability and safety.

We find that this method correlates with the actual results of these evaluations, thus having the potential to save time and upfront cost, as the need to implement a custom evaluator and executing the corresponding evaluation logic locally post-inference is no longer needed. In contrast to traditional LLM as a Judge mechanisms, our methodology is objective, and does not rely upon a linear scoring scheme: responses for a particular prompt are either correct or incorrect.

To validate this methodology, we ran a single pass against 5 different models through IFEval, an evaluation dataset measuring the capability of a model to follow generalized instructions. [8] We ask the judge to rate the model's ability to follow instructions, and to rate the readability and toxicity of the generated text. We then use a Spearman's rank correlation to determine the correlation between the heuristic scores, calculated directly against the semantics of the scores, and the model graded responses.

2 Results

We tested the following models against IFEval, using Gemini as the judge: Llama-3.18-8-B, Llama-3.23-3B, qwen-2.5-70B, and Phi-2.5-mini-128k. We note that the objective of this evaluation is to not measure the capabilities of these models, but rather the ability of the judge to effectively rate responses subjectively on various metrics. We utilize Spearman's rank correlation to directly measure the directional strength, independent of the scale, between the model-graded scores and the heuristic scores. For the heuristic scores, human readability was calculated using Flesch Readability, while toxicity was calculated using the "detoxify" package. [9] We note that the toxicity score is technically from a trained machine learning model, rather than pure heuristics. However, it is still a valid proxy due to its popularity and standardization for analyzing bodies of text. Raw IFEval scores were taken from pre-determined leaderboards such as LLM Extractum, or the model results themselves. [10] We avoid using the LayerLens platform for this evaluation in order to curtail potential bias in the experiment. The difference between the model-graded ratings and the actual evaluation scores listed in the leaderboards or papers is noted, along with the Spearman strength correlation between the

The following table showcases the correlation strength for all the models that were tested.

Table 1: Metrics for instruction following, toxicity and human readability on IFEval for different models

Model	Instruction Following Difference	Human Readability Correlation	Toxicity Correlation
meta llama 3.1-8b instruct	13 percent	52.4 percent	51.7 percent
meta llama 3.1-70 instruct	2.6 percent	50.2 percent	52.2 percent
qwen 2.5 72b instruct	13.1 percent	52.2 percent	50.9 percent
microsoft phi 3.5 mini 128k-instruct	4.4 percent	49.5 percent	58.6 percent

3 Discussion

Gemini, despite not being an advanced reasoning model, showed a high degree of agreeablity with actual evaluation scores when acting as a judge. It also shows a relatively decent correlation with heuristic scores: in all cases, the Spearman correlation was above 0, indicating that there was some degree of correlation between the heuristic scores and the raw judge ratings. It is worth noting that there are not enough samples or variability for the results of this experiment to have any formal degree of statistical relevance. We mark this report as merely preliminary, with more experimentation needed to validate the proposed ideas. Specifically, more models should be tested, along with more metrics. A stronger, state-of-the-art judge should also be used to validate the central claim. We propose utilizing GPT-4, the standard in most LLM as a Judge evaluations, as the judge, and including a more comprehensive set of models, spanning multiple sizes and architectures, to reach a more formalized conclusion.

4 Potential Impacts

4.1 Automated Toxcity Detection

As most LLM as a Judge scenarios are, by their very nature, long form datasets that cannot be verified through heuristics or static evaluation methods, there is some potential for generated text to include some degree of toxicity. The same also applies for COT or multi-turn datasets that may require human annotation or scoring: because of the nature of the dataset, responses to individual prompts may have some presence of toxicity. If a model-based judge is found to be able to predict the underlying toxicity of base-level responses such as this with a relatively high degree of accuracy, it might be able to potentially detect the presence of toxic phrases or words in responses to common scenarios that may be included in larger datasets. This allows for organizations and users to be able to use human language to directly prompt a judge to scan for incorrect or toxic language instead of relying on post-data processing or human annotation, thus potentially saving time and resources.

4.2 Readability Improvements

Using an automated evaluator for scoring readability can also enable model creators to be able to directly improve the capability of their models to generate human-readable text. Rather than relying solely on subjective, recursive feedback from actual humans, long-form evaluations that use chain of thought or are conversation based in nature could also be graded for average human readability, given model creators additional data to improve their baseline models. While this is not a full replacement for human feedback, it could be an additional set of data that could help offset the impact of sybil ratings or other incorrect data from traditional human ratings.

5 Conclusion

In conclusion, we explored the potential for automated evaluators, or judges, to serve as labelers for metrics in long-form datasets. Specifically, we ran an experiment that evaluated the capability of Google's flagship Gemini model to measure a generated text's readability, toxicity, and ability to follow instructions based on the input. These scores were then compared against the raw evaluation score, and heuristic-based scores. We note that while the results do show some promise, they are not significant, and more analysis is needed in order to draw any meaningful conclusions.

References

[1] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[2] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[3] Sean Welleck, Jiacheng Li, Ximing Li, Qizhe Zhang, Peter Liang, and Yejin Choi Zhou. Evaluating mathematical accuracy in language models. *arXiv preprint arXiv:2306.17bb35*, 2023.

[4] Xiang Xu, Peter Liang, and Yejin Choi Zhou. Chatbot arena: An open platform for evaluating large language models through human preference. `https://chat.lmsys.org`, 2023.

[5] Yue Zhang, Jindong Wei, Hao Ye, Kai Zhou, Xiaomin Yang, et al. Wildbench: A comprehensive evaluation framework for large language models in the wild. *arXiv preprint arXiv:2401.00595*, 2024.

[6] Lianmin Zheng, Wei-Lin Chen, Zhihong Jiang, et al. Mt-bench: A benchmark for multi-turn language model evaluation. *arXiv preprint arXiv:2306.05685*, 2023.

[7] Google. Gemini: A family of highly capable multimodal models. `https://blog.google/technology/ai/google-gemini-ai/`, 2023.

[8] Xiang Li, Dian Zhou, and William Yang Wang. Ifeval: A framework for evaluating instruction following in large language models. *arXiv preprint arXiv:2311.07911*, 2023.

[9] Laura Hanu. Detoxify: Toxic comment classification with transformers. `https://github.com/unitaryai/detoxify`, 2020.

[10] Llm extractum: A comprehensive collection of language model evaluations. `https://llm-extractum.org`, 2024.