LLM-as-a-judge

Mokrov Semen
Amsterdam University of Applied Science
Amsterdam, The Netherlands
semen.mokrov@gmail.com

Deep Kayal
Amazon Research
London, The United Kingdom
_

Riccardo Pinosio
Amsterdam University of Applied Science
Amsterdam, The Netherlands
_

Abstract

Here the abstract

CCS Concepts

• **Do Not Use This Code → Generate the Correct Terms for Your Paper**; *Generate the Correct Terms for Your Paper*; Generate the Correct Terms for Your Paper; Generate the Correct Terms for Your Paper.

Keywords

Do, Not, Us, This, Code, Put, the, Correct, Terms, for, Your, Paper

ACM Reference Format:
Mokrov Semen, Deep Kayal, and Riccardo Pinosio. 2025. LLM-as-a-judge. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 2 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Here the introduction

2 Related Work

2.1 Large Language Models as Evaluators

2.1.1 Introduction to LLM Evaluators. Large Language Models (LLMs) are increasingly being used as automated evaluators for open-ended text generation tasks, providing a scalable alternative to slow and costly human judgment. Recent benchmarks and leaderboards for chatbots and text generation often rely on powerful LLMs (such as GPT-4) to rank or score model outputs in place of humans. For instance, the Vicuna Chatbot evaluation pipeline employs GPT-4 to compare responses in a pairwise manner, achieving evaluation agreements on par with human annotators on many queries. Researchers have even built interactive frameworks like *EvalGen* [7] that leverage LLMs to suggest evaluation criteria and generate test queries for model outputs, with a human-in-the-loop to refine these criteria. These approaches demonstrated the potential of LLM-as-a-judge systems: a strong LLM can approximate human evaluation reasonably well (e.g., GPT-4's judgments agree with human preferences roughly 75–80% of the time on multi-turn dialog tasks) [10]. As a result, LLM evaluators have quickly become a central tool for comparing models and tuning them to human-like performance.

2.1.2 Evaluation Paradigms. LLM-as-a-Judge evaluations mostly employ two paradigms: pointwise and pairwise evaluations.

Pointwise Evaluation. In pointwise evaluation, an LLM assesses a single response independently, assigning a score based on predefined criteria such as relevance, coherence, or factual accuracy. This method is straightforward and allows for fine-grained analysis of individual responses. However, it may suffer from inconsistencies due to the subjective nature of scoring and the lack of comparative context [7].

Pairwise Evaluation. Pairwise evaluation involves presenting the LLM with two responses to the same prompt and asking it to determine which one is better according to specific criteria [6]. This comparative approach can mitigate some biases inherent in pointwise evaluations and often aligns more closely with human judgment. In score-based pairwise comparison, the LLM assigns each response a numerical score based on evaluation criteria such as coherence, helpfulness, or relevance, and then ranks them by comparing the scores [5]. This method allows for fine-grained differentiation between responses while retaining the benefits of comparative judgment.

2.1.3 Benchmarks Utilizing LLM-as-a-Judge. Several benchmarks have been developed to standardize the evaluation of LLM outputs using the LLM-as-a-Judge paradigm.

MT-Bench. MT-Bench is a multi-turn benchmark designed to assess the conversational abilities of LLMs. It evaluates models on their capacity to maintain context, follow instructions, and provide informative responses over multiple dialogue turns. MT-Bench employs GPT-4 as the judge, leveraging its advanced understanding to score model responses. The benchmark includes an Elo rating system derived from pairwise comparisons, offering a dynamic leaderboard that reflects the relative performance of various models[10].

AlpacaEval. AlpacaEval focuses on single-turn instruction-following tasks, evaluating models based on their ability to generate helpful and accurate responses. It uses GPT-4 as an automatic evaluator to compare model outputs against reference responses. AlpacaEval provides win-rate statistics and supports a leaderboard that ranks models accordingly. The benchmark has been validated against a substantial set of human annotations, ensuring its reliability and relevance[3].

Chatbot Arena. Chatbot Arena is a platform that combines human and LLM-based evaluations to rank conversational agents. It utilizes pairwise comparisons, where human annotators or LLM

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

judges select the better response between two options. The results contribute to an Elo rating system, providing a comprehensive leaderboard that reflects both human and automated assessments. Chatbot Arena serves as a valuable resource for understanding the comparative performance of various chatbots in real-world scenarios[10].

2.2 Problems in Current Systems

2.2.1 Position Bias. Position bias refers to the tendency of LLM-based evaluators to favor responses based on their position in a prompt during pairwise comparisons. Studies have shown that models like GPT-4 often prefer the first response presented, regardless of content quality[8]. This bias can compromise the fairness and reliability of evaluations. To address this, frameworks such as PORTIA have been developed, which align similar content across candidate answers to mitigate position bias effectively[4].

2.2.2 Length Bias. Length bias occurs when LLM evaluators disproportionately favor longer responses, associating verbosity with higher quality. This can lead to inflated evaluations for unnecessarily lengthy outputs. Research has demonstrated that longer responses often receive higher preference scores, even when shorter responses are equally or more informative[2].

2.2.3 Cost Problem. Employing state-of-the-art large language models (LLMs) as evaluators in LLM-as-a-Judge frameworks introduces significant computational costs. High-performance models like GPT-4 entail substantial expenses per token, rendering large-scale evaluations financially heavy. Simpler queries are handled by smaller, cost-effective models, while more complex tasks are escalated to larger, more expensive models. This strategy can significantly reduce overall evaluation costs without compromising accuracy[1]. At the same time, ARJudge optimizes the trade-off between evaluation quality and computational expense by dynamically allocating resources based on the prompt characteristics[9].

References

[1] Lingjiao Chen, Matei Zaharia, and James Zou. 2023. FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance. arXiv:2305.05176 [cs.LG] https://arxiv.org/abs/2305.05176

[2] Zhengyu Hu, Linxin Song, Jieyu Zhang, Zheyuan Xiao, Tianfu Wang, Zhengyu Chen, Nicholas Jing Yuan, Jianxun Lian, Kaize Ding, and Hui Xiong. 2024. Explaining Length Bias in LLM-Based Preference Evaluations. arXiv:2407.01085 [cs.LG] https://arxiv.org/abs/2407.01085

[3] Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. AlpacaEval: An Automatic Evaluator of Instruction-following Models. https://github.com/tatsu-lab/alpaca_eval.

[4] Zongjie Li, Chaozheng Wang, Pingchuan Ma, Daoyuan Wu, Shuai Wang, Cuiyun Gao, and Yang Liu. 2024. Split and Merge: Aligning Position Biases in LLM-based Evaluators. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 11084–11108. doi:10.18653/v1/2024.emnlp-main.621

[5] Yinhong Liu, Han Zhou, Zhijiang Guo, Ehsan Shareghi, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2024. Aligning with Human Judgement: The Role of Pairwise Preference in Large Language Model Evaluators. In *First Conference on Language Modeling*. https://openreview.net/forum?id=9gdZI7c6yr

[6] Adian Liusie, Potsawee Manakul, and Mark J. F. Gales. 2024. LLM Comparative Assessment: Zero-shot NLG Evaluation through Pairwise Comparisons using Large Language Models. arXiv:2307.07889 [cs.CL] https://arxiv.org/abs/2307.07889

[7] Shreya Shankar, J. D. Zamfirescu-Pereira, Björn Hartmann, Aditya G. Parameswaran, and Ian Arawjo. 2024. Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences. arXiv:2404.12272 [cs.HC] https://arxiv.org/abs/2404.12272

[8] Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. 2025. Judging the Judges: A Systematic Study of Position Bias in LLM-as-a-Judge. arXiv:2406.07791 [cs.CL] https://arxiv.org/abs/2406.07791

[9] Kaishuai Xu, Tiezheng Yu, Wenjun Hou, Yi Cheng, Liangyou Li, Xin Jiang, Lifeng Shang, Qun Liu, and Wenjie Li. 2025. Learning to Align Multi-Faceted Evaluation: A Unified and Robust Framework. arXiv:2502.18874 [cs.CL] https://arxiv.org/abs/2502.18874

[10] 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. arXiv:2306.05685 [cs.CL] https://arxiv.org/abs/2306.05685