Title: LLM-as-a-Judge

URL: https://en.wikipedia.org/wiki/LLM-as-a-Judge

PageID: 79975879

Categories: Category:Large language models, Category:Machine learning stubs, Category:Natural language processing stubs

-----

LLM-as-a-Judge or LLM-based evaluation is a conceptual framework in natural language processing (NLP) that employs large language models (LLMs) as evaluators to assess the performance of other language-based systems or outputs.

Instead of relying solely on human annotators, the approach leverages the general language capabilities of advanced language models to serve at automated judges.

LLM-as-a-Judge may be more cost-effective and may be added to automated evaluation pipelines.

Unlike traditional automatic evaluation metrics such as ROUGE and BLEU , which rely on transparent, rule-based comparisons with surface-level n-grams, LLM-as-a-Judge relies on the opaque internal reasoning of large language models. The LLM-based evaluations likely incorporate deeper semantic understanding, but at the cost of interpretability.

Beyond the interpretability there may be other issues with LLM evaluators. [1] For instance, if an LLM has generated an output, the evaluation of the output with the same LLM may yield a distorted evaluation, "LLM narcissism". [1] [2]

Typically, a more powerful LLM is employed to evaluate the outputs of smaller or less capable language models—for example, using GPT-4 to assess the performance of a 13-billion-parameter LLaMA model. [3] Recent research has also explored leveraging multiple LLM evaluators to improve fairness and scalability, [4] and the idea of “LLM juries” has been proposed as a practical mechanism to mitigate bias. [5]

References

This large language model -related article is a stub . You can help Wikipedia by expanding it .

v

t

e

This machine learning -related article is a stub . You can help Wikipedia by expanding it .

v

t

e