Reliability of LLM-as-a-Judge Evaluations Using MT-Bench Benchmarks and Telemetry Data

Bryan Falkowski
bf2349@rit.edu

***Abstract*— Large language models are used as evaluators for comparing model-generated outputs such as text and code. This project uses MT-Bench as a dataset and collects telemetry data while rerunning the dataset across multiple models and llm-as-a-judge evaluators. The resulting data will then be analyzed. A web-based dashboard will also be developed to visualize the results.**

Keywords—LLM evaluation, LLM-as-a-judge, telemetry, MT-Bench

I. INTRODUCTION

Large language models are used to generate text and code, however, based on the non-deterministic nature of the responses, evaluating the output is a challenge. Human evaluation does not scale well, and traditional techniques of string comparisons are not possible. As a result, recent work has explored using LLMs themselves as evaluators. Using LLMs as evaluators is commonly referred to as *LLM-as-a-judge* [1].

One popular benchmark is MT-Bench [1] which supports LLM-as-a-judge by using structured prompts with human-labeling to evaluate the responses. The MT-Bench benchmark is generally used to compare models.

This project proposes using MT-Bench as an existing, public dataset [2] to analyzes the consistency of LLM-as-a-judge evaluations. MT-Bench contains 3.3K expert-level pairwise human preferences for model responses generated by 6 models in response to 80 MT-bench questions [2]. Consistency between runs will be evaluated by rerunning benchmark prompts across multiple models. Telemetry data will then be collected during execution to review variation in evaluation behavior. Results from these repeated runs and telemetry data will then be summarized through a web-based dashboard.

II. PROBLEM

A. Problem Definition

Automated evaluation using large language models is gaining acceptance and can be used to compare text and code outputs. Once issue is that the consistency of these evaluations, across repeated executions, is difficult to evaluate. Since the same prompt and evaluation can produce different results, it raises the questions about the accuracy of using LLM-as-a-judge [4]. This project examines automated evaluation using MT-Bench, telemetry, and repeated benchmark execution as the basis for analysis.

B. Significance of Problem

LLM-as-a-judge evaluation is widely used in model development, benchmarking, and system comparison because it can scale better than human evaluation. Developers will frequently try to obtain metrics to analyze drift, or the difference between models. If tests cannot be repeated consistently, it questions the validity of the tests. With a better understanding of how consistent automated judgments are, developers can better interpret evaluation result. These results can be used to decide when automated evaluation is trustworthy [4].

III. PRIOR WORK

Evaluating open-ended outputs from large language models has traditionally relied on human evaluation which has been used successfully, but is difficult to scale. Several studies have shown that LLM-based evaluators can almost replace human scoring for many tasks, which, if true, makes them valuable for large-scale evaluation [3].

MT-Bench [1] was introduced to support this automated judging by providing structured prompts and human-labeled data. MT-Bench has been widely adopted in the industry as a test dataset. This project builds on the existing MT-Bench dataset and evaluation work by examining the consistency of LLM-as-a-judge evaluations using repeated benchmark execution.

IV. PROPOSED METHODOLOGY

A. Plan

The goal of this project is to analyze the consistency of LLM-based automated evaluation by testing how evaluation outcomes vary when prompts are executed multiple times. MT-Bench will be used as an initial dataset because it provides structured prompts along with human-labeled data.

The project will begin by selecting a small set of representative language models (TBD based on availability and cost) and run the MT-Bench prompts against them. These responses will then be evaluated using the LLM-as-a-judge approach. Each prompt will be executed multiple times to collect variance.

During each job execution, telemetry data will be collected. Telemetry data would include llm reponses, character lengths, execution time, token usage, and any other relevant metadata retuned from the llm. This data will be cleaned up for analysis.

Once data collection is complete, the analysis will be performed to examine patterns. I intended for this analysis to focus on how often evaluation outcomes change across runs and what factors can influence that.

Finally, the results of the analysis will be presented through a simple web-based dashboard or application. The dashboard will allow results to be summarized and reviewable (design TBD)

B. Challenges or Barriers

I think one challenge I will have is working with the non-deterministic nature of llm outputs. I expect that this variability will be seen across repeated executions, however, I don't expect it to be extreme. Another potential challenge is the cost of executing prompts multiple times across different models. To manage this, the scope will be a subset of models that are accessible and cost efficient. Finally, the content of telemetry data might vary across executions (and between models). The greater the variance in the model responses, the more difficult it could be to compare the telemetry data.

C. Project Deliverables

The project will produce the following deliverables:

- A dataset derived from repeated execution of MT-Bench prompt. This would include the LLM-as-a-judge evaluation outcomes and telemetry data.
- An analysis of the above data which examines consistency in LLM-as-a-judge evaluation across runs.
- An online dashboard or application. The dashboard will allow results to be summarized and reviewable.
- A report documenting the steps, results, and conclusions of the project.
- A presentation summarizing the project and the analysis.

REFERENCES

[1] L. Zheng et al., "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena," arXiv preprint arXiv:2306.05685, 2023.

[2] LMSYS Org., "MT-Bench Human Judgments Dataset," Hugging Face, 2023. (accessed Jan. 28, 2026)

[3] Y. Liu *et al.*, "G-Eval: NLG Evaluation Using GPT-4 with Better Human Alignment," *Proc. EMNLP*, 2023.

[4] J. Tan *et al.*, "JudgeBench: A Benchmark for Evaluating LLM-Based Judges," *arXiv preprint arXiv:2410.12784*, 2024.

[5] T. Dubois *et al.*, "AlpacaEval: An Automatic Evaluator of Instruction-Following Models," 2023.

[6] W.-L. Chiang et al., "Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference," arXiv preprint arXiv:2403.04132, 2024.

[7] B. H. Sigelman *et al.*, "Dapper, a Large-Scale Distributed Systems Tracing Infrastructure," *Communications of the ACM*, vol. 54, no. 4, pp. 57–66, Apr. 2010.