Homework 2 Paper: Ranking LLM

Due: Friday, October 17th at 11:59 pm

Deliverables. Submit a PDF of your write-up to Gradescope *HW2 Paper*

Overview

Large Language Models (LLMs) are capable of doing amazing things! But how do we evaluate their performance? Unlike standard classification tasks, evaluation of LLM outputs involves nuance, style, and human judgment. This homework will help you explore two recent approaches to evaluation: **Chatbot Arena** and **VibeCheck**.

You are encouraged to skim related works, but focus on the problems, current solutions, contributions, methods, and limitations.

This assignment is not about memorizing details, but about developing the ability to *read research papers methodically.* Most research papers follow a common structure: they first motivate a **problem**, then describe **current solutions** and their limitations, followed by the **proposed solution and key insights**, the **methods** used, and finally a discussion of **limitations**. This homework is designed to help you practice and internalize this process of critical reading as you answer the questions. We have also provided pointers to relevant section that you might want to pay more attention to for each question.

- Paper: Chatbot Arena (arXiv:2403.04132) VibeCheck (arXiv:2410.12851)
- Platform: VibeCheck

For each question, we are looking for high-level concise answers, no need to write essays.

Paper Questions

Chatbot Arena

Q1. [**Problem**] What problem is Chatbot Arena aiming to solve? Why is evaluating generative models challenging compared to classification tasks? [**Section: 1. Introduction**]

Q2. [**Current works**] What are current approaches to solving this problem and why are they not sufficient? [**Section: Related Works, LLM Benchmark**]

Hint: What are the categories of LLM benchmarks outlined in the paper, and what are the limitations of each category?

Q3. [**Proposed solution**] What are the inputs and outputs of Chatbot Arena?

Hint: outputs exist both per-battle and aggregated.

Q4. [**Key insight / contributions**] Why does the Chatbot Arena evaluation approach address the issues of prior benchmarks? What are their main contributions? (These could be methods, software artifacts, formalization of a problem, datasets, etc.)

Q5. [**Method details**]

(a) What is a Bradley–Terry coefficient, and how is this used in the Chatbot Arena scoring?

(b) **Section 4** of the Chatbot Arena paper mentions that the model scoring system of the original Chatbot Arena interface used Elo scores instead of Bradley–Terry coefficients because they are "better for the purpose of statistical estimation." Why are Bradley–Terry coefficients more appropriate for this setting?

Hint: Elo is an online metric, meaning that the model scores are updated after every battle, while Bradley–Terry is offline—it fits all match outcomes at once to find coefficients that best explain the entire dataset. Think about what assumptions Elo makes about the models.

Additional resource: "As a starting point, we show that the Elo score provably fails to extract the transitive component of some elementary transitive games." (Bertrand et al., 2023)

Q6. [**Limitations**] What are some limitations of Chatbot Arena? What are general limitations of human preference benchmarks? [**Section: 8. Discussion**]

VibeCheck

Q7. [**Problem**] What problem is VibeCheck aiming to solve? [**Section: 1. Introduction**]

Q8. [**Current works**] What are existing approaches to evaluation in this space, and why are they not sufficient? [**Section: 2. Related Work**]

Q9. [**Proposed solution**] What are the inputs and outputs of VibeCheck? [**Section: 3. Vibe-Based Evaluations, Briefly 5 for Examples**]

Q10. [**Method details**] How are vibes quantified? What is the metric of success (i.e., what numbers in the results section do we want to be high)? [**3. Vibe-Based Evaluations**]

Q11. [**Key insight / contributions**] Why does VibeCheck address the issues present in current approaches? How does VibeCheck address some limitations of Chatbot Arena? What are their contributions? (These could be methods, software artifacts, formalization of a problem, datasets, etc.) [**Section 6. Application, 8. Conclusion**]

Q12. [**Limitations**] What are some limitations of VibeCheck? Would there be circumstances where the results may be untrustworthy? [**Section 7. Limitation**]

Other Practical Resources

Machine Learning Competitions: Kaggle

Kaggle hosts many ML competitions with a tier/medal system valued by industry. Despite being competitive, there's a strong culture of code sharing that accelerates learning.

Recommended Entry-Level Competition

- Titanic: Machine Learning from Disaster

Other Relevant Competitions

- Chatbot Arena (LMSYS)
- WSDM Cup Multilingual Chatbot Arena
- LLM Detect AI-Generated Text
- LLM Classification and Fine-Tuning