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**]

Solution. Chatbot Arena addresses the difficulty of evaluating free-form generative responses, where correctness is not binary but depends on human preference, style, and nuance.

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?

Solution. Existing approaches to LLM evaluation fall into several categories (Figure 1 of the paper):

(1) **Static, ground-truth–based benchmarks** (e.g., MMLU, GSM-8K, BigBench, HumanEval, AGIEval). – Limitation: subject to contamination, saturation, and overfitting; may not reflect real human preferences.

(2) **Safety and comprehensive suites** (e.g., ToxicChat, HELM). – Limitation: valuable coverage, but still static and do not capture interactive use cases.

(3) **Open-ended benchmarks with human or LLM judgment** (e.g., MT-Bench, AlpacaEval, GPT-4-as-judge). – Limitation: human evaluation is expensive and does not scale; LLM-as-judge may introduce bias.

(4) **Live benchmarks and human interaction** (e.g., Codeforces problems, RLHF studies, live user feedback). – Limitation: often restricted to specific organizations and lack openness/scale.

Overall, these benchmarks either fail to capture alignment with human preferences, are vulnerable to contamination and overfitting, or are costly and unscalable. This motivates Chatbot Arena's large-scale, crowdsourced, human-preference framework.

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

Hint: outputs exist both per-battle and aggregated.

Solution. Input: prompt + two candidate model responses. Outputs: (i) battle outcome (human vote between A/B), (ii) aggregated leaderboard via pairwise comparisons. **[Section: 3, 4]**

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.)

Solution. Arena turns subjective human preference into structured pairwise data, enabling large-scale crowdsourced evaluation. Contributions include: the Arena platform, the leaderboard, and the formalization of head-to-head model evaluation.

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)

Solution. (a) The Bradley–Terry model assigns each model a latent "strength" parameter and uses pairwise win/loss outcomes to estimate these strengths. In Chatbot Arena, this

produces a probability distribution over which model will win a head-to-head match, which can then be aggregated into a leaderboard.

(b) The authors switched from Elo to Bradley–Terry because Elo updates scores sequentially after each battle, which introduces noise and makes rankings depend on the order of matches. Bradley–Terry, by contrast, performs a joint fit using the entire dataset of outcomes.

Bradley–Terry coefficients are more appropriate because they provide statistically consistent estimates that best explain all pairwise results simultaneously. Elo allows for changing skill differences and can overweight recent matches or small sample sizes, which is problematic in crowdsourced, noisy human-preference data. Bradley–Terry avoids these issues by treating all battles together in a unified model.

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

Solution. The paper highlights several key limitations:

– The user base, while extensive, is skewed toward LLM hobbyists and researchers, leading to a potentially biased distribution of votes.

– Prompts are primarily drawn from the Arena's online chat interface, which may not reflect real-world or domain-specific usage, creating a skewed prompt distribution.

– The evaluation focuses on helpfulness, not safety, overlooking an important dimension of LLM behavior.

More broadly, human preference benchmarks face challenges of user bias, uneven coverage, lack of standardized evaluation contexts, and scalability constraints.

VibeCheck

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

Solution. VibeCheck tackles scaling issues of human eval by introducing LLM-as-a-judge and a consistent "vibes" framework.

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

Solution. Existing approaches fall into three main directions:

– *Aspect-based benchmarks:* static metrics like BLEU or ROUGE, later extended with criteria such as fluency, factuality, or conciseness. These are limited because they still define the axes on what makes something correct beforehand.

– *Human and pairwise comparisons:* interactive tools (e.g., AutoSxS, LLMComparator) and crowdsourced preference studies provide richer insights. While these efforts focus more on the user experience, it does not provide an interpretable view of exactly why these users prefer one output over the other

– *Qualitative/trait discovery:* recent HCI and ML work explores discovering separable traits in unstructured data (e.g., "set A contains more X"). However, these efforts lack comprehensive evaluation metrics and consistent validation.

Together, these approaches supply valuable foundations but fall short of providing scalable, unbiased, and systematically verifiable evaluations for LLMs.

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

Solution. Input: triples (p, o_p^A, o_p^B) of prompts with two model outputs, plus preference labels and vibe annotations from a judge (human or LLM).

Output: ordered "vibes" (axes like formal → friendly) with associated scores. Vibes are quantified by:

- *Well-definedness* (agreement, Cohen's κ),

- *Differentiation* (separability score, model-matching accuracy),

- *User alignment* (logistic regression predicting human preference). Together these yield a set of validated vibes with interpretable coefficients and predictive power.

- *Example from Section 5*: Language and Tone. Professional, straightforward tone TO Enthusiastic, friendly tone.

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**]

Solution. (a) Vibes are quantified along three axes:

"We define 3 key criteria of a useful vibe; it should be well-defined, differentiating, and user-aligned."

– *Well-definedness:* measured by inter-annotator agreement (Cohen's κ).

– *Differentiation:* measured by separability score and model-matching accuracy.

– *User alignment:* measured by logistic regression prediction accuracy for human preferences.

Higher values for κ, separability, and preference accuracy indicate stronger vibes.

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**]

Solution. VibeCheck provides reproducible, rubric-based judgments via LLMs, reducing cost while adding consistency. It complements Arena by calibrating and diagnosing biases.

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

Solution. Limitations include reliance on rubric quality, bias of judge LLMs, domain shift sensitivity, and reduced transparency compared to full human evaluation.

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