Ranking Large Language Models with LMArena

Paper Questions Answers

1. What are some challenges of evaluating a Large Language Model's performance? How is it different from some other evaluation tasks? (Answer: very subjective, hard to have a simple rule to quantify - imagine a book translated by two different translators)
2. What are some challenges that Chatbot Arena tries to address, and how were they done? (Answer: try to suggest a solution for the subjective nature of LLM evaluation noted in question 1 with quantitative methods)
3. What is a Bradley-Terry Coefficient, and why is this involved in the chatbot arena scoring? (Answer: it allows us to formulate a setup to learn the strength and probability a certain model wins another model)
4. What are each of the heatmaps in Figure 2 denoting? Which battle has the highest win-rate (write out model A vs model B)? (Answer: gpt-4-turbo vs mistral-7b-instruct)
5. What are some limitations that still exist within Chatbot Arena? (Answer: Biased users)
6. What is the purpose of Vibecheck, and how is it measured? (Answer: basically LLMs-as-judge)
7. How can VibeCheck be used to address some limitations with Chatbot Arena? (Answer: can find biases which can be taken into account when computing the leaderboard)