Ranking Large Language Models with LMArena

Overview

LLM (Large Language Models) are capable of doing amazing things! But how do we evaluate their performance? Essays written by humans also have subjective and stylistic aspect - how do we even evaluate these LLMS?

These introductory questions would let us familiarize ourselves with LMArena before we dive into implementing models. Skim and read through the paper and the platform below, and try answering the questions below.

Paper: Arena and VibeCheck

Paper Questions

1. What are some challenges of evaluating a Large Language Model's performance? How is it different from some other evaluation tasks?
2. What are some challenges that Chatbot Arena tries to address, and how were they done?
3. What is a Bradley-Terry Coefficient, and why is this involved in the chatbot arena scoring?
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)?
5. What are some limitations that still exist within Chatbot Arena?
6. What is the purpose of Vibecheck, and how is it measured?
7. How can VibeCheck be used to address some limitations with Chatbot Arena?

Other Practical Resources

Machine Learning - Participating in Kaggle Competitions

💡 Here, I will introduce a specific competition in more detail! This is a great opportunity to see both deep learning and machine learning-based solutions.

Kaggle operates like a game with tiers. Usually, many machine learning competitions are held, and finishing in the top ranks allows you to earn medals and increase your tier. This is also recognized in the industry! I also learned a lot about machine learning through Kaggle initially, and I've often found that my competition experience has been helpful both technically and in my career.

💡 There are cases where ML Engineers from non-cs or liberal arts background becoming a Kaggle Grandmaster! This shows how Kaggle is a great platform to get hands-on experience with ML.

One of the biggest advantages of Kaggle is that despite being a competition, there is a **culture of sharing codes and solutions among participants**, making it very useful for learning.

As an entry-level competition, I recommend the Titanic competition, which is the most famous and common one in Kaggle.

https://www.kaggle.com/competitions/titanic

After following the above link, you can check out the various codes shared by users in the code section. Studying these codes and understanding them one by one can be a quick way to improve your skills in a short period.

Some other competitions that might be more closely related to the homework is

https://www.kaggle.com/competitions/lmsys-chatbot-arena https://www.kaggle.com/competitions/wsdm-cup-multilingual-chatbot-arena (Multilingual Chatbot)

https://www.kaggle.com/competitions/llm-detect-ai-generated-text/leaderboard

https://www.kaggle.com/competitions/llm-classification-finetuning/overview