Problem Statement for AI Tinkerer Hackathon

Launch date of Problem Statement : 26th September
Actual demo day : 24th October

Context

In the fast-paced world of AI, **Large Language Models (LLMs)** have become indispensable across various industries. They generate human-like text, answer questions, and perform complex language tasks. However, leveraging their full potential isn't just about deploying these models—it's about meticulously **evaluating and refining** them to meet specific needs.

Challenges in LLM Evaluation

Traditional evaluation methods—such as n-gram matching, semantic similarity metrics, or comparisons to gold-standard references—are often **ineffective** at distinguishing high-quality responses from mediocre ones.

While building evaluation datasets using **human annotators** gives great results, it requires significant effort and high-quality labeled data, making it **difficult to scale**. Using LLMs for evaluation can be **slow, expensive, and hard to scale**.

An emerging solution is using **LLM Evaluators**, also known as **"LLM-as-a-Judge"**—LLMs that evaluate the quality of another LLM's response to an instruction or query. However, aligning an LLM Judge with **human judgments** is often challenging, with many implementation details to consider.

Hackathon Goal

In this hackathon, let's collaborate to build and improve LLM-judge together. Some ideas are:

- **Productionizing the Latest LLM-Evaluator Research**: Implement cutting-edge research findings into practical, scalable solutions.

- **Enhancing Existing LLM Judges**: Improve the alignment of LLM evaluators with human judgments.
- **Develop human-in-the-loop LLM-judge**: Create prototypes of platforms that enable real-time collaboration between humans and AI for evaluating LLM responses.

Dataset

You are free to use **any datasets** as long as you can demonstrate how your LLM Judge **improves** by using them. This flexibility allows you to tailor your approach to the data that best suits your solution. Here's how you can leverage datasets effectively:

Demonstrate Improvement: Show how your LLM Judge improves using these datasets through quantitative metrics (e.g., accuracy, F1 score) or qualitative analyses (e.g., examples of better alignment with human judgments).

Baseline Comparison: Compare your LLM Judge's performance against existing evaluation methods or baseline models like GPT4 to highlight the enhancements you've achieved.

If you don't know where to start, LMSYS-Human-Preference-55k is a good place to start. It contains over **55,000 human-annotated preferences** between language model responses, enabling your LLM Judge to closely mimic human evaluations.

Evaluation Criteria

- **Creativity**
  Anything from creative prompting, to system design and/or UX for llm-as-a-judge projects
- **Utility / Usefulness**
  How does this project affect the real world

- **Technical Implementation / Execution**
  High level of technical ability, implementation of existing eval research
- **Presentation**
  Team concisely delivers their project during presentation, github is open, weave dashboards and traces included, etc