LLM Evaluation Task

There are multiple statistical and non-statistical methods to evaluate LLMs.The statistical ones are generally metric based and can be compared.

One other upcoming strategy that is out performing all the other strategies are using a third LLM as a judge for the task this has better results than any other metrics

So i have used two models from Google's AI studio to create responses and the third to generate the response

https://github.com/xbhinxv54/Evaluate_task/tree/main

a. Evaluation Criteria

Have added all of the required metrics from the Assignment Documents
The model/judge has the capability to store previous history so it will verify against it for **Repeatability**

b. Comparison Criteria

The model is asked to score the LLMs based on the previous history the correctness of the current query
The judge will create mark both the LLMs on a scale of 10

c.Human Evaluation

If the judge is unable to make decisions it will mark the output for human verification after which a person can score both the models

Reference