LLM as Judge

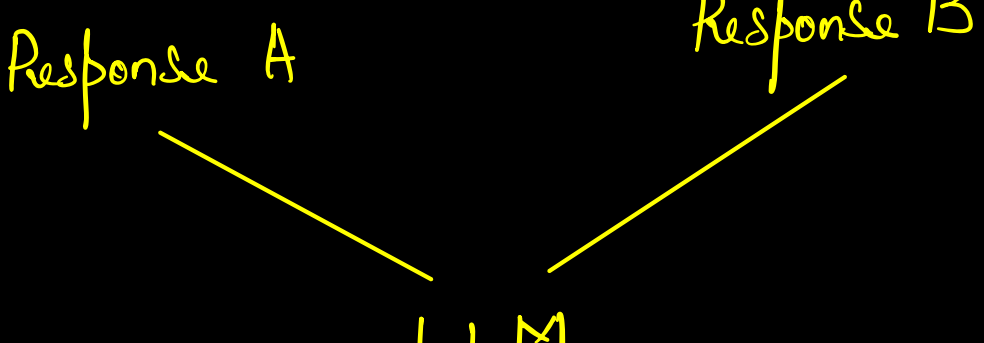
Uses of LLM to evaluate AI-generated text based on custom criteria defined in a prompt.

Evaluation

AI Text :-
1) Politeness
2) Bias
3) Sentiment
4) Tone
5) Hallucinations

Types of LLM Judge

1) Pairwise Comparison
   - Give LLM 2 responses
   - Ask to choose which is better



LLM

Which one is better?

Criteria

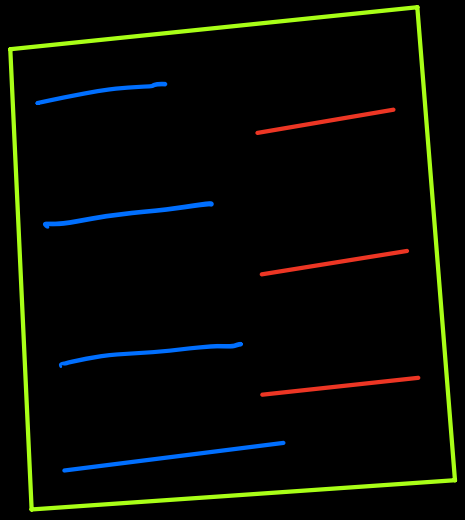
2) Evaluation by Criteria

Response → LLM → Awesome

LLM → Not Awesome

Evaluate based on criteria

Conversation Evaluation



1) Detecting Denial
2) Identify Repitions
3) Detecting Emotion
4) Frustation Detection
5) User Query Resolution

3) Reference based Evaluation S.T/G.T = Reference

(i) LLM Answer + Reference Answer

(ii) LLM Answer + Question
Response Addressing the question or not.

* (iii) LLM Answer + Reference + Question
Answer + Retrieved
Context

*
RAGAS = { user_input, response, reference, retrieved_contexts }
*