Demonstration:

Research Objectives->1. Exploring Q&A capabilities of different LLMs

• Model cards of different models being used to check their training cutoffs

• performance on the MMMLU benchmark and Stanford QA benchmark

• Semantic differences in questioning -> build question and observe metrics

• Creating a dataset

2. Bridging the Gap between NLP and LLMs

• Why does hallucinations happen and how to quantify/ detect them -> any NLP way of doing it?

Demonstration flow 1.

-> Display MMLU Results

-> GSM8K

-> HellaSwag

2. Talk about different individual score and show them, how they look for the following models -> Llama 7B , Mixtral 8x-7B and Mistral 7B.

-> in that, talk about different metrics and the problems associated with that metric

-> Can highlight the issue with MMLU where some questions aren’t correctly marked in the dev set even and causes incorrect eval (minor issue)

3. Now focusing on one of the problems, create a RAG Pipeline where papers are added to the code, stuff is retrieved. Looking at one of the metrics of MMLU

-> draw up suggestion

4. Contrasting provide NLP evaluations

-> BLEU

-> ROGUE

# Responses, Questions and References:

Experiment 1: NO Context fed into the system, Let’s take the following examples.

• adjust temp and top-k, not adjust temp and top k

-> fairly recent event(out of training window)

-> Asking about celebrities

-> asking one fact question.

-> asking double fact questions.

Experiment 2: Given context through RAG Performance measurement:

• adjust temp and top-k, not adjust temp and top k

-> Give context in form of vectorized documents for QA

-> compare the results