DEMO SLIDES structure:

Slide 1 and 2

• Talk about objective -> research question

Slide 3 -> motivation -> bridging gap between LLM eval and NLP (making sense of LLM results)

Slide 4 -> what exists, where is the gap -> discuss MMLU

Slide 5 -> what I propose ->

Slide 6,7 -> methodology discussion.

Slide 8 -> results and conclusions

Demo begins:

• show MMLU benchmarks for different LLMs (slide 4)

• run experiment 1 and 2

• demonstrate results

• Q&A

////

Standardize reference -> homogenize -> Wikipedia 1st passage. -> con

Interest -> ask how it brings multiple facts together -> does it hit bag of references ->

-> performance on new -> see what entities it comes up with -> hallucination?