Main Research Questions:

- Explore Question and answer Capabilities of different LLMs
  - Question and answer Capabilities -> Different metrics such as Human Eval -> VARIOUS DIFFERENT METRICS -> look at them
  - How -> looking at Q&A of RAG Pipelines
- Looking at performance of Standard NLP tests and why do they fail to perform well on the LLM responses
  - Standard NLP metrics for evals -> BLEU, ROGUE, METEOR
  - Tasks were created for machine translation and summarization (cite)
  - Look at overlap in words, n-grams and syntactic structures.
  - LLMs have advanced beyond NLP perhaps but we can still look at them-> We can run MMLU, HumanEval(?), etc

- Tackle the GAP between quantitative metrics and Subject evaluation (real user experience) in LLMs.
  - Show how

Methodology behind asking a specific type of question: -> required to be discussed if validation set is made.

- Proposed theory can be:
  - Same question asked semantically in a different way:
    - What is the capital of France? Paris
    - Which city is the official residence of the president of France? Paris

  - Single knowledge base -> answering fact1-> boiling point of water at sea level is 100 degrees Celsius.
  - Double knowledge base -> answering fact1 and fact2 -> saltwater has higher boiling point than freshwater.
  - Reasoning built on single knowledge and double knowledge -> answer whose foundation lies in using both fact1 and fact2-> Q: determine cooking time of pasta at sea level. Answer: Saltwater boils at higher temp and hence reduced cooking time for pasta as compared to unsalted water at 100 degrees Celsius.

- Talk about HumanEval benchmark -> ability to write code (do we measure this? Perhaps not given the Q&A nature, but also at the same time its part of the Q&A)

Talk about Tokenization and how it effects the answers of the different LLMs based on their tokenization techniques -> don't know where to put this? Design? Literature? Is it required at all? Perhaps yes

Proposed Metrics/benchmarks looked at which can inspire our metrics:

I - > General Knowledge Benchmarks –

- Good for a general purpose LLM
- MMLU – set of 57 questions from STEM, humanities and History.
  - Assesses the language understanding in relatively good breadth.
  - Findings -> poor performance on morality, poor performance on calculations (can also talk about why modelling math is an issue in LLMs) -> all these for uncalibrated LLMs(not fine-tuned).
  - MMLU calculates average for how many questions it answered correctly for each of the 57 domains and averages the results from them to give a final quantifiable answer.

II -> Logical Reasoning Benchmarks:

- Good for interactable LLMs
- Popular tests are:
  -
  - AI2 Reasoning Challenge(ARC)
  - GSM8k -> Grade school Math challenge -> 8k questions.
  - Loads of prompt engineering required, can be one of the tests we can do on the validation set I create.

III -> Coding Benchmarks

- Human Eval -> by OpenAI -> coding performance is checked -> pre existing dataset -> I don't think its necessary.

Problems that can come with having such tests:

* no pipeline created for testing locally

Have to rely on NLP techniques for normal analysis

- Prompt engineering can help fix problem of getting mixed answers

Going forward now, I will implement the code for which the following would happen:

- Load in a RAG pipeline (set of documents on which we will Q&A)
- Load 3 instances of different LLMs and ask them a variety of questions
- Below the answer given by the prompt, show the various metrics which could be measure from the answer given by the LLM such as -> F1 score, BLEU score, ROGUE score.

Need help figuring out how to configure running the knowledge-based questions for testing? Do I have to create my own validation dataset?

If yes then, for the demonstration I can have a pre-run copy of the results of the answers on my validation dataset in case the code fails to run at that time. Based on the answers received by the different LLMs I can visualize the results and discuss the insights obtained from the test.

Finally, I have the introduction and conclusion written out in draft and it revolves around my main research question.