Outline for the dissertation

1. Introduction.

1.1. Project Overview:

Talks about the general structure, Describes the two experiments being performed in this, simple words, no technical terms

1.2. Motivation:

Rapidly changing space, lots of research taking place, requirement for LLMs to be assessed and talked

1.3. Project Breakdown:

Chapter wise breakdown of the project, what each chapter entails

2. Background/Literature Review

2.1. Background:

What is a llm, Current state of LLMs, papers on Transformer model that started everything (2), evolution of llm(), metrics used previously for testing previous word and semantic models() ,current ways of evaluating LLMs().

2.2. Current state of LLM evals,

Large datasets being tested (MMLU, OpenQA, HumanEVAL, etc.)

2.3. RAG (Retrieval Augmented Generation):

What is RAG, when it was introduced, What does it do, architecture of RAG, Example of how it works.

Fine tuneing discussed, Why fine tuneing is not preferred in this situation(attach paper)

2.4. Metrics around RAG and how they can be helpful -> BERTScore, Semantic Search, BLEURT, CHRF, MOVERScore

Describe these metrics listed above and how they are used for evaluation with their formulas and use cases.

3. Experiment1 Summarisation Tasks:

3.1. Why summarization task chosen? ->

Understand comprehension capabilities of LLM -> tested via BERTScore, BLEURT score and MOEVERScore.

3.2. Setting up the Summarization Tasks

3.2.1. WITH RAG

How RAG is being set up, storage of document embeddings, ,

3.2.2. WITHIN CONTEXT (WARS)

3.3. QUESTIONS Asked and Responses Received

Here talking about the nature of question being asked, why they are being asked and considerations that went behind putting up the questions

4. Experiment 2 Knowledge Based Tasks

4.1. Why Knowledge based Tasks Chosen

4.2. Setting up the Experiment

4.2.1. WITH RAG

4.2.2. Within CONTEXT(WARS)

4.3. Questions asked and Responses received.

Here talking about the nature of question being asked, why they are being asked and considerations that went behind putting up the questions

5. Evaluations

5.1. Summarization Tasks

5.1.1. WITH RAG -> Ngram and semantic

5.1.2. NON RAG-> ngram and Semantic

5.2. Knowledge Based Tasks:

5.2.1. WITH RAG -> Ngram based and Semantic

5.2.2. NON RAG -> Ngram based and Semantic

5.3. Comparisons

6. Conclusion

6.1. Conclusion:

Upon conducting the experiments, a superiority of embeddings based approaches like BERTScore, BARTScore and BLEURT is observed. By leveraging contextual embeddings, these metrics demonstrate a better semantic understanding of the content. Classical NLP metrics such as rouge and BLEU which focus predominantly on surface level word overlap fall short in capturing the depth of meaning and subtleties inherent in LLM generated text which do not correlate with human judgement on knowledge based and summarization tasks. Additionally an (improvement/reduction) in score is observed in embeddings as well as NLP based metrics for a subject whose context is added via RAG. Finally, embedding based evaluation metrics aligns more closely with inherent capabilities of LLM's and their verbose nature. (improvement/reduction) in ability is observed with the introduction of new Context.

6.2. Future Work