Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Question and Answering capabilities of LLMs

Ayush Gupta

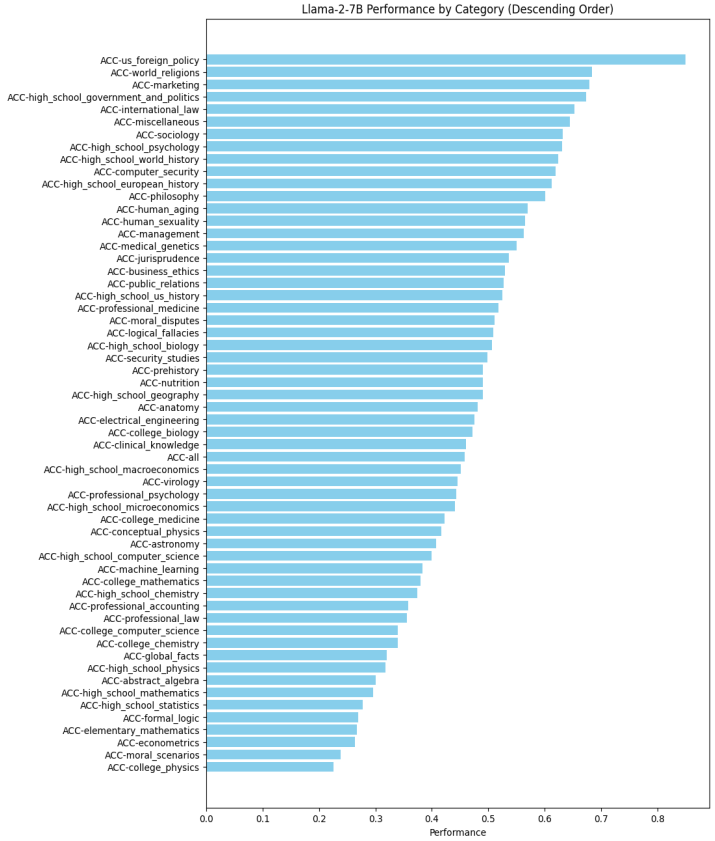
Date 22/03/2024

Overview

- Motivation
- Experiment 1
- Observations from experiment 1
- Experiment 2
- Inferences from experiment 2
- Conclusions

Motivation

- Introduction of LLMs have taken the world by a storm.
- LLM's have good breadth of knowledge, but how do they perform on new knowledge?
- MMLU -> Massive Multitask Language Understanding.
  - Measure knowledge of pretrained models
  - 57 subjects including STEM, History, Politics, etc.
  - Most used benchmark.
- Benchmarking LLMs reliably is a very hot topic currently, with no concrete set of metrics.



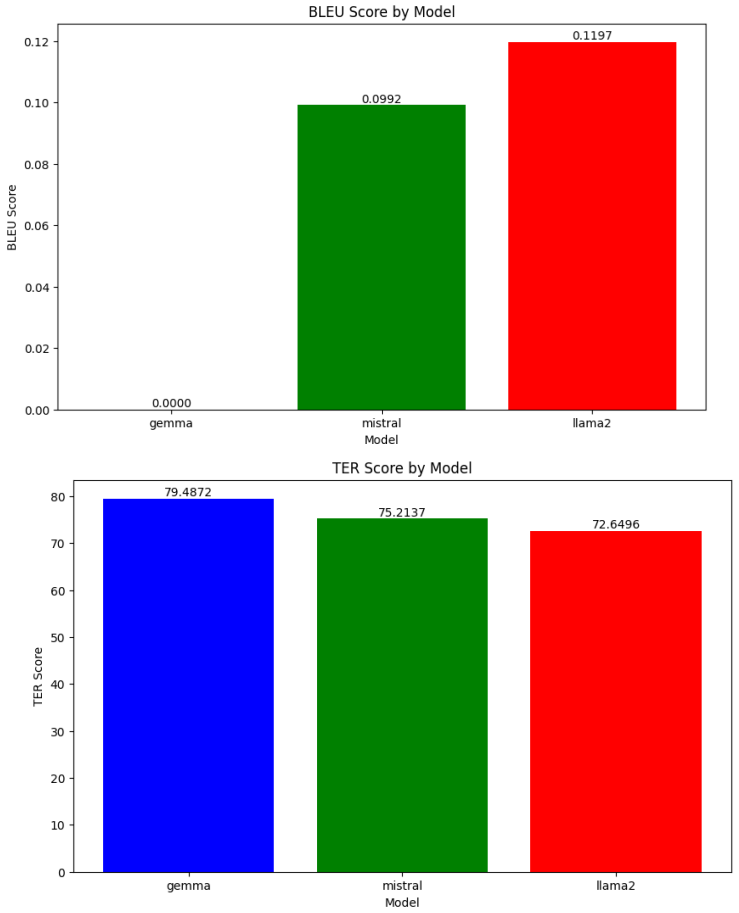
MMLU score and performance of llama2

Experiments

- Main Goals
  - In this presentation we explore the Q&A abilities of open-source Large Language Models.
    - We look at their performance with and without relevant context.
    - Measure their performance through traditional NLP practices.
    - Find alternative methods to the above NLP methods.
- Experiments:
  - 1. Measuring n-gram based NLP metrics for an LLM.
  - 2. Measuring Q&A capabilities of LLM when given a specific context on which it hasn't been trained via RAG.

Experiment 1: n-gram based tests

- Setting up the Experiment:
  - Prompt the LLMs with a specific question.
  - Have a reference answer text against which the LLM's responses will be tested.
  - Question -> *Tell me about the Battle of Boyne in 100 words.*
  - BLEU, ROUGE and TER scores are calculated
    - BLEU -> quality of machine translated text based on overlap of n-gram phrases.
    - ROUGE -> evaluating machine translation using overlap of n-grams, word sequences, and word pairs b/w reference and response.
    - TER -> amount of editing required to a machine translation to match reference text.
- The performance is quite poor.

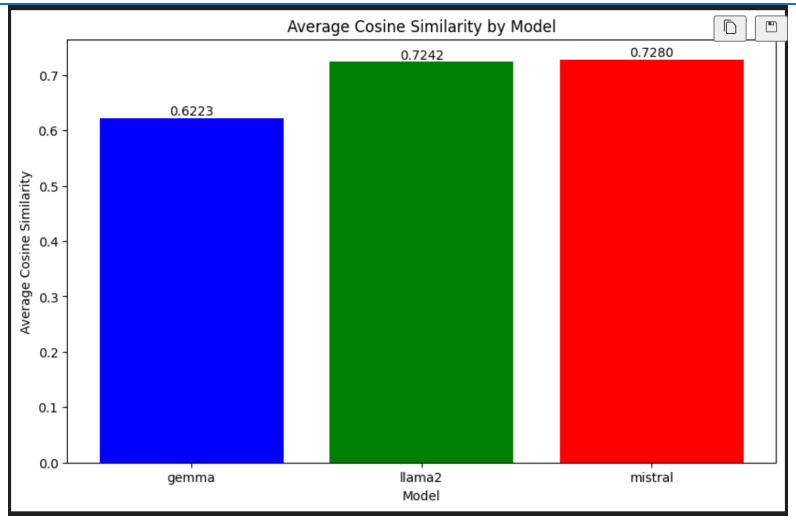


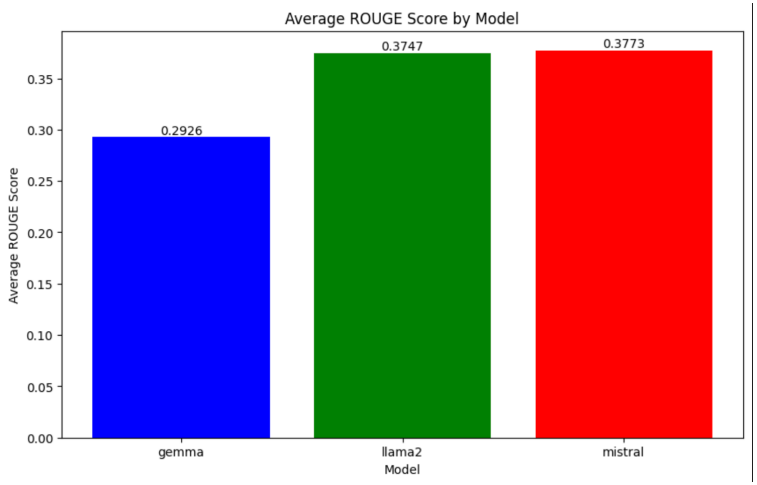
Performance on NLP metrics

N –Gram based Testing

Observations:

- BLEU scores are very low -> expected since summarization tasks generally perform poorly.
- ROUGE score indicates higher overlap with reference text's n-grams.
- TER score is high across the board, indicating poor performance. Again, expected due to summarization nature and metric is used for translation.
- Cosine similarity is relatively high which indicates good semantic alignment.





Experiment 2: Introduction of RAG

- RAG -> retrieves relevant information from corpus
  - How it works:
    - Input query is converted into an embedding(high dimensional vector); easier to work with vectors
    - Above query is used to search for pre-processed queries.
    - Retrieved query is used to provide context for generating response.
- Movies released in 2023 are taken as RAG stores
    - Barbie
    - Oppenheimer

Tests Performed

- We perform the following test:
  - BERTScore -> uses BERT's contextual embeddings to compute similarity between words in reference and generated text via cosine similarity.
  - Semantic search -> uses meaning and context to find the most relevant information.
- Testing methodology:
  - 10 questions each are formed about the movie along with corresponding reference answers.
  - Questions are categorise based on knowledge/ prose.
  - LLM's responses are tested against the reference text, before and after giving access to context and RAG.

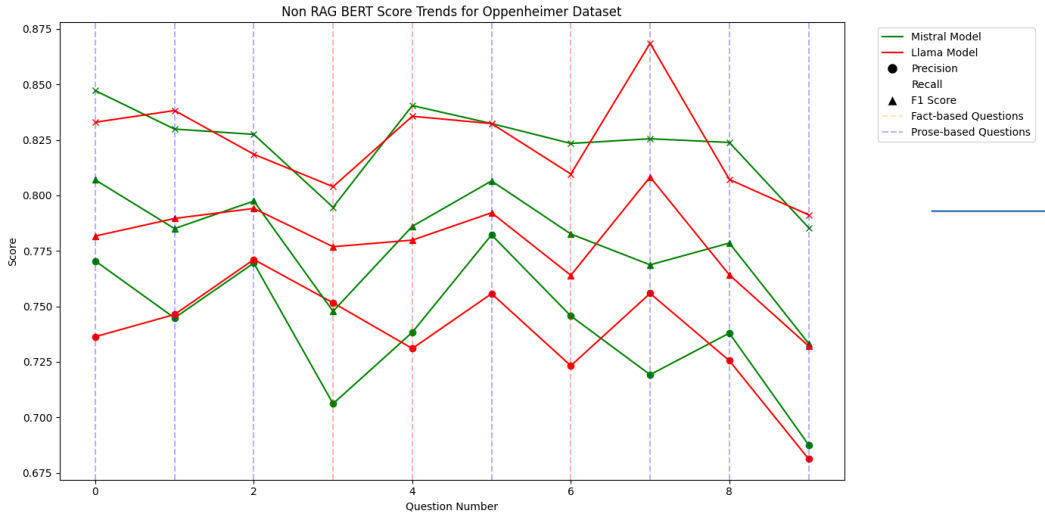
Question	Reference Text	Question_Kind
Who tries to breakdance at Barbie Margot's party?	During the big block party hosted by Barbie Ma...	Fact
What does Barbie Margot think about dying duri...	In the midst of the vibrant party atmosphere, ...	Prose
What is Ken Ryan Gosling's idea of his house?	Ken Ryan Gosling, asserting dominance over the...	Fact
What does the Mattel CEO prioritize over busin...	The Mattel CEO passionately declares that his ...	Prose
How does Barbie Margot feel after her party?	Following the joyous and lively party, Barbie ...	Prose
What unusual event happens to Barbie in the mo...	Barbie Margot experiences a series of bizarre ...	Fact
What change does Barbie Margot notice about he...	While enjoying time at the beach, Barbie Margo...	Fact
What happens when Barbie Margot visits Weird B...	Upon visiting Weird Barbie's abode, Barbie Mar...	Fact
How does Ken Ryan Gosling react to the concept...	Ken Ryan Gosling, after learning about the con...	Fact
What does Barbie Margot find challenging in th...	Barbie Margot finds navigating the real world ...	Prose

Reference Barbie Question set

Question	Answer	Question_Kind
What were the consequences of Oppenheimer's pe...	Oppenheimer's legal challenges and personal co...	Prose
How did the public perception of Oppenheimer c...	After the atomic bombings, Oppenheimer became ...	Prose
What was Truman's response to Oppenheimer's co...	Truman dismissed Oppenheimer's concerns about ...	Prose
How did Strauss's actions contribute to Oppenh...	Strauss played a crucial role in the revocatio...	Fact
What role did Lewis Strauss play in the post-w...	Lewis Strauss's influence extended beyond Oppe...	Fact
What were the effects of the H-bomb developmen...	The development of the H-bomb and the associat...	Prose
How did Oppenheimer's past affiliations and ac...	Oppenheimer's past affiliations with communist...	Fact
What was the impact of the atomic bomb on Oppe...	The atomic bomb had a profound impact on Oppen...	Prose
In what ways did Oppenheimer attempt to influe...	Oppenheimer attempted to influence US atomic p...	Prose
How did the security hearing against Oppenheim...	The security hearing against Oppenheimer was c...	Prose

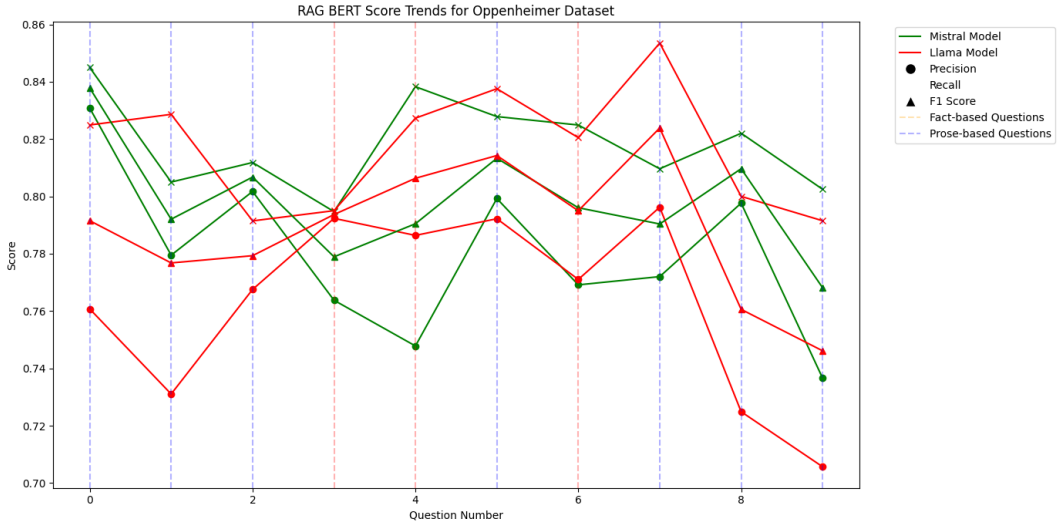
Reference Oppenheimer Question set

No RAG v/s RAG Performance (Oppenheimer)

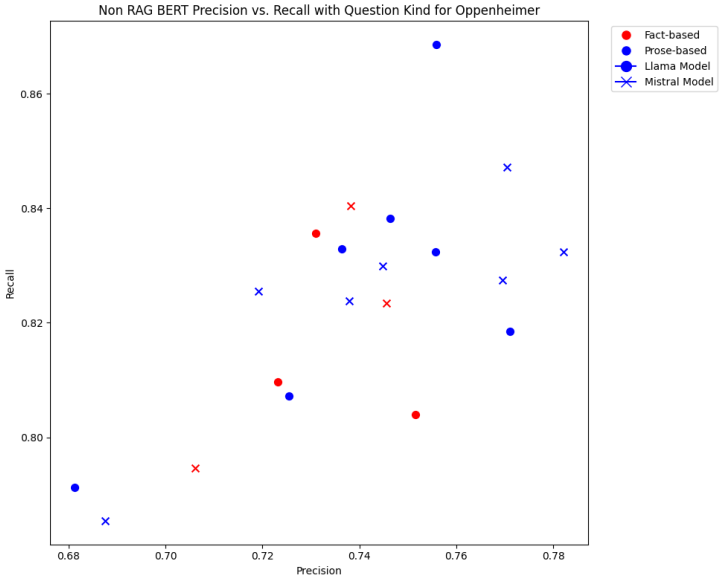


- Noticeable fluctuation -> inconsistent performance
- Performance is good could be since Oppenheimer is very famous naturally and has good amount of information on him already.

- Less variance is observed in the scores, suggesting a stabilized baseline performance after RAG especially in Mistral.
- Similar trends are observed, and performance is affected by the question kind.

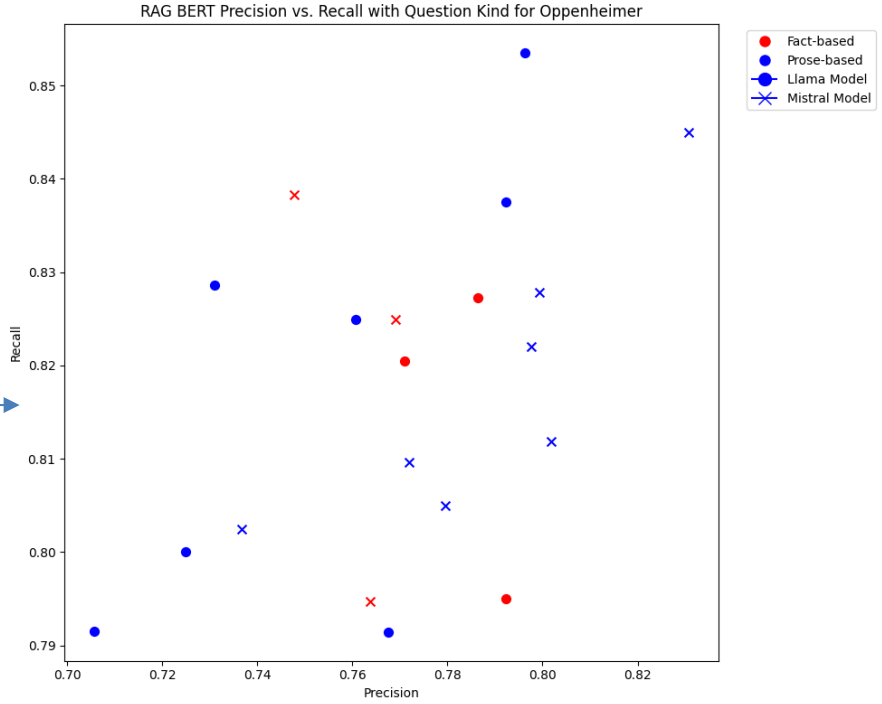


Precision vs recall (Oppenheimer)

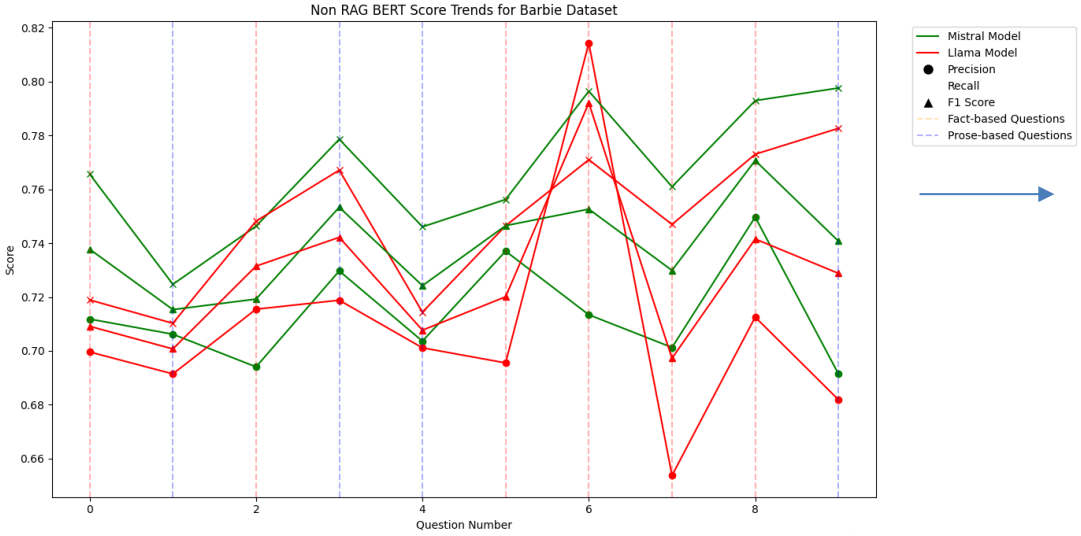


- a slight increase in precision is observed after adding the RAG context
- still, small changes are observed due to the nature of the questions being asked and Oppenheimer being quite famous and having material on him already.

- Prose based questions see lower performance due to translational complexity.
- Question asked as facts, are related to events in the movie and hence low precision and recall are observed.
- the variance in recall and precision for both the models is relatively the same

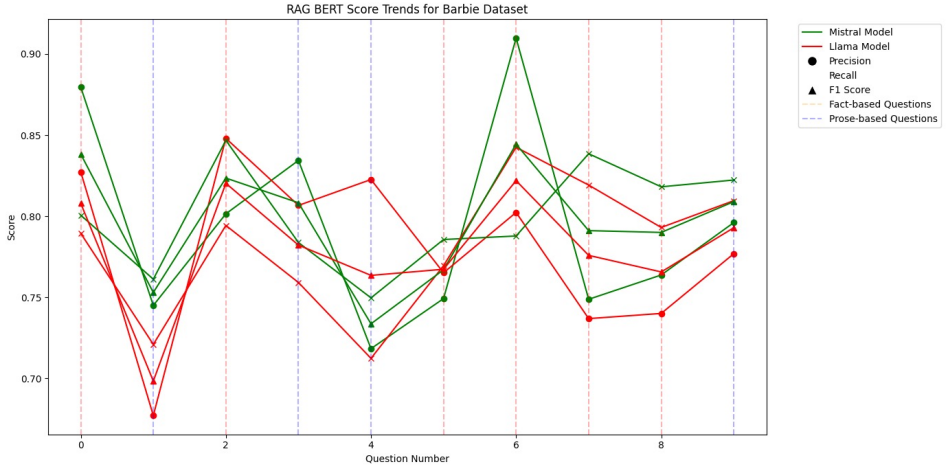


No RAG v/s RAG Performance (Barbie)

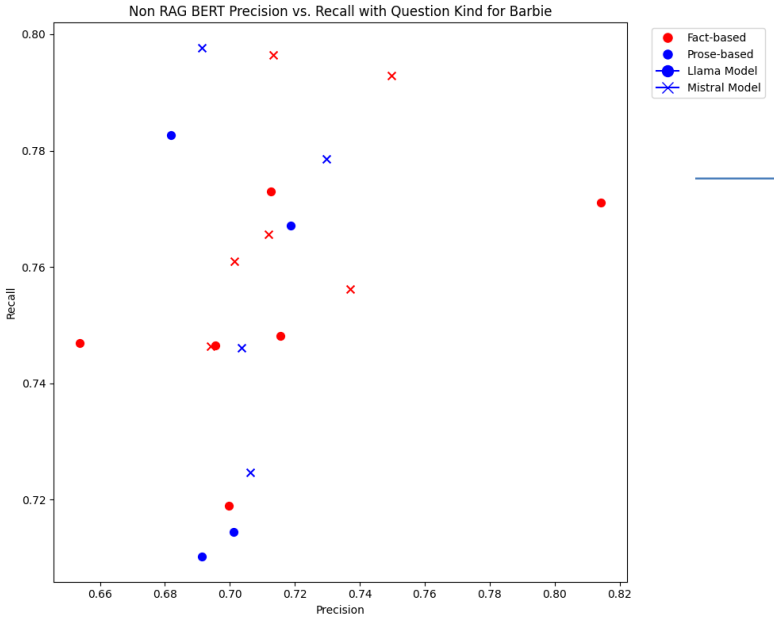


- Lower scores for precision and recall observed when compared to Oppenheimer
- expected since the movie is fictional and is not included in prior training data.
- still a decent performance in terms of not having seen the context.

- As observed earlier, the variance is notably low in RAG scores
- some higher peaks are observed in precision, generally, higher scores are observed in fact-based questions (red dotted vertical)
- similar but improved scores are observed in prose based Questions.

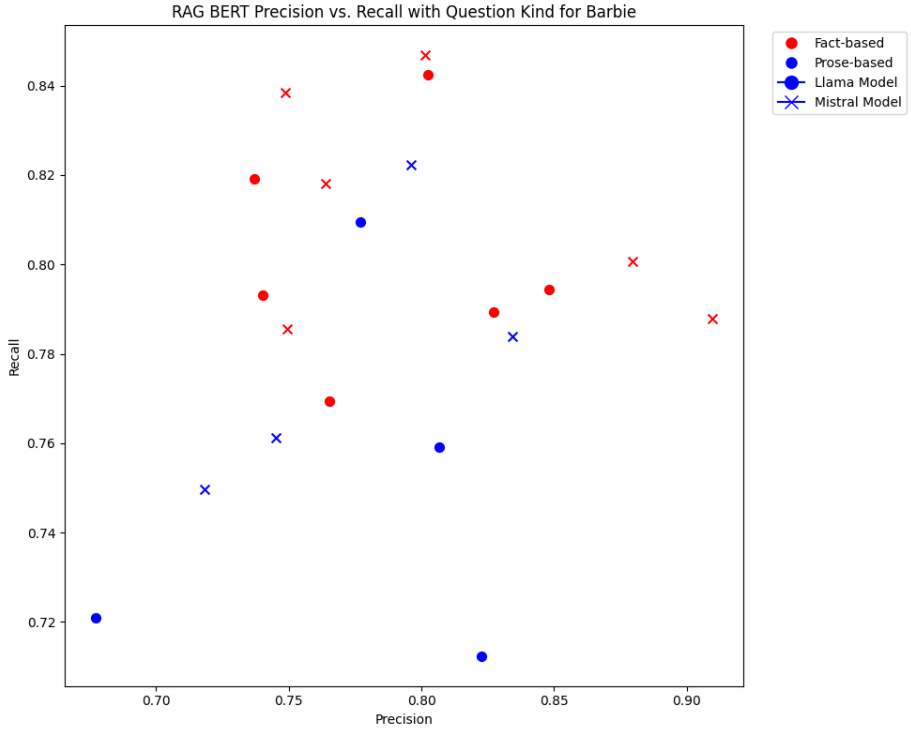


Precision vs recall (Barbie)

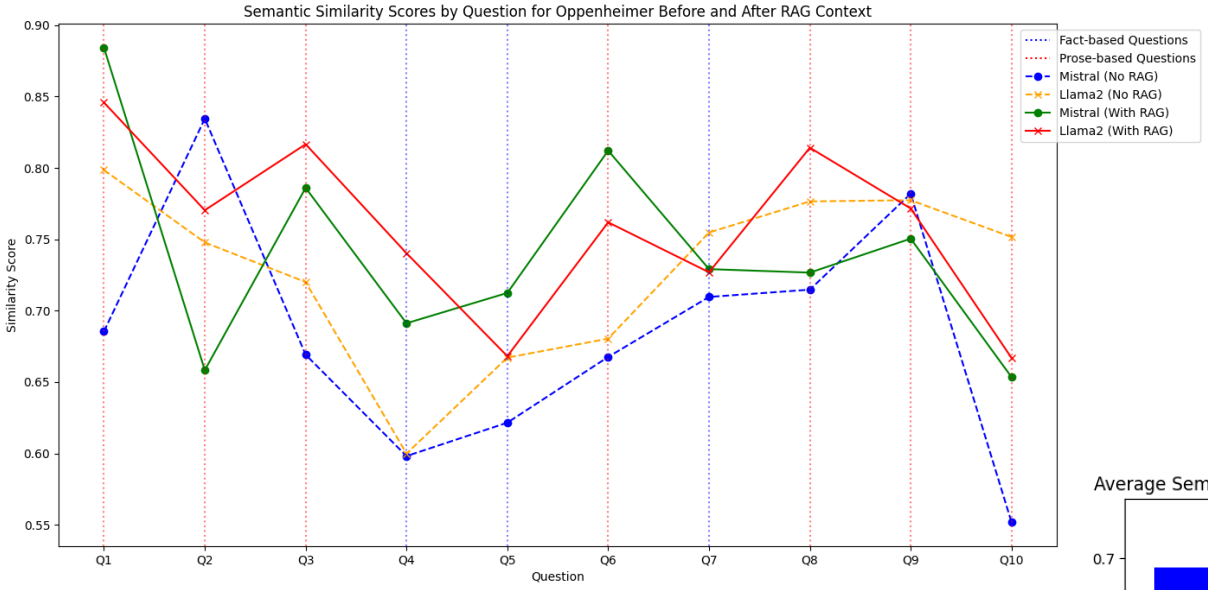


- Higher precision and recall scores are observed indicating RAG stabilizing the performance on questions.
- After RAG, Precision and Recall has improved relatively significantly across the board.
- lower variance in fact based questions is also observed.

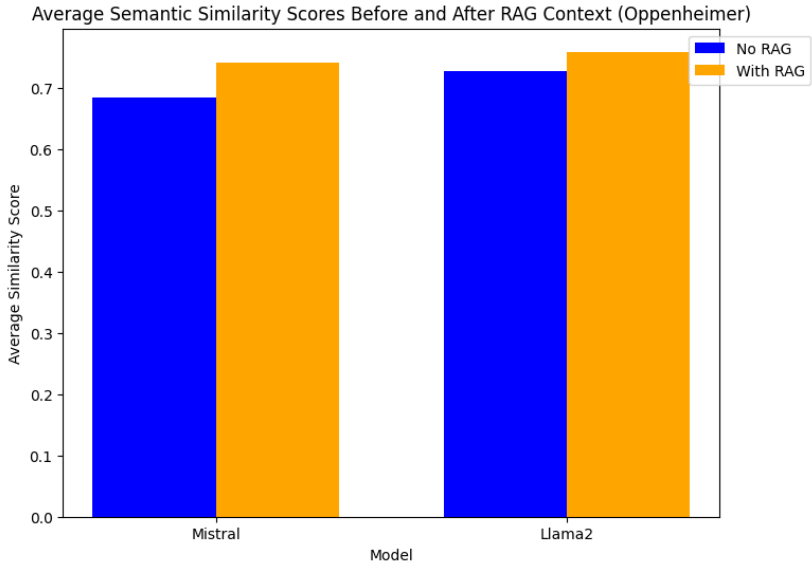
- Low precision scores are observed across the board due to fictional nature of the movie.
- for some questions, high precision and recall can be observed.
- High variance in precision is observed in fact-based question, again, expected due to fictitious nature of the movie and being out of context.



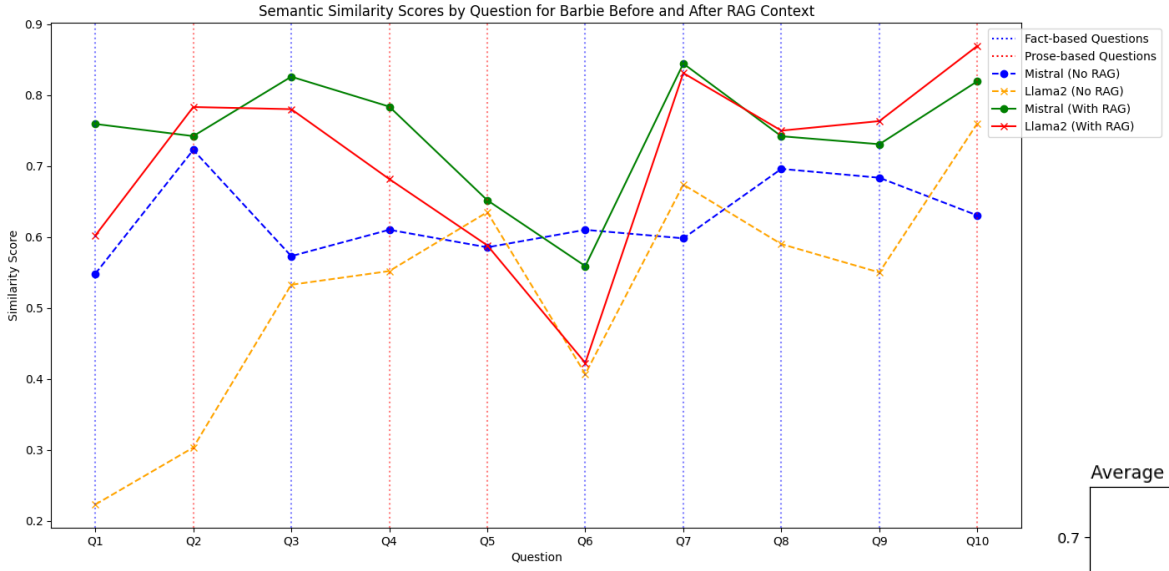
Semantic Similarity(Oppenheimer)



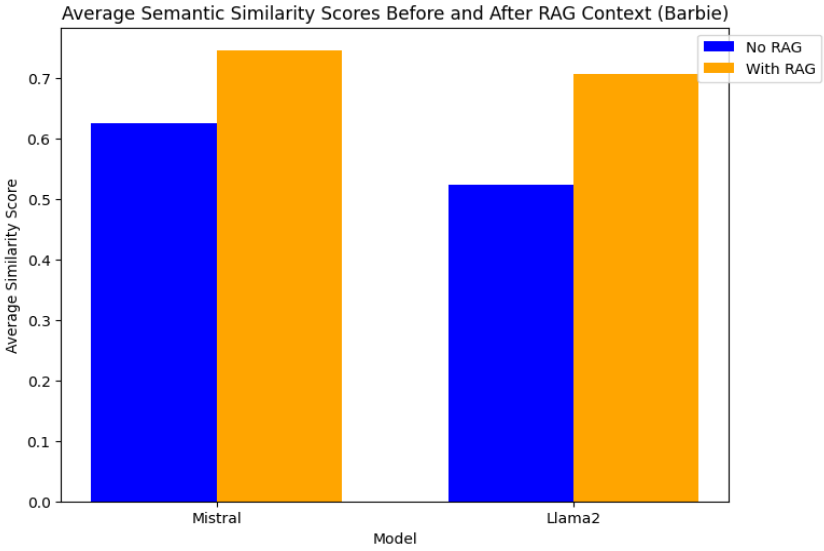
- Better performance is observed in semantic similarity after providing the RAG context.
- Scores are still erratic but after providing the context, both the models are more similar than they are dissimilar(except Q2,Q8).
- Only small increase in semantic similarity is observed.



Semantic Similarity(continued)



- After RAG, both models are very similar in performance, more tightly coupled.
- significant increase in semantic similarity is observed across the board, especially in fact-based questions.
- Prose based questions still perform relatively the same.



Inference from Experiment 2

- Performance is based on the kind of question being asked, poor performance on prose-kind questions observed before and after RAG.
- Relative improvement in scores observed in Fact-based questions after the introduction of the RAG.
- Q&A on Fictional movie (barbie) observed more improved performance when compared to Non-fictional movie based on a famous historical character(Oppenheimer).

Conclusions

- Traditional NLP doesn't capture the performance and performance improvement in LLMs very well.
- Alternative methods of assessment such as semantic search and BERTScore can capture the nuances of LLMs more efficiently.
- Fact based questions see significant improvement when an $RAG_{(context)}$ is fed into the system.
- Improvement in Q&A capabilities on fictitious material observed to be higher than non-fictitious material.

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Questions ?

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Thank You