CS-432 PROJECT

CONTEXT OVERLOADING

Group 3:
Hassan Ali
Muhammad Affan
Talha Tariq

MOTIVATION

Retrieval-Augmented Generation (RAG) pipelines often supply LLMs with large, relevant chunks of information. However, despite their relevance, the sheer volume of context can overwhelm the LLM, leading to missed information, incomplete reasoning, or hallucinations. The issue isn't retrieval quality — it's the LLM's inability to process long, dense context effectively. Our project is motivated by the need to optimize the RAG performance under high-context loads, enabling models to better utilize large retrieved inputs without losing accuracy or coherence.

PROBLEM STATEMENT

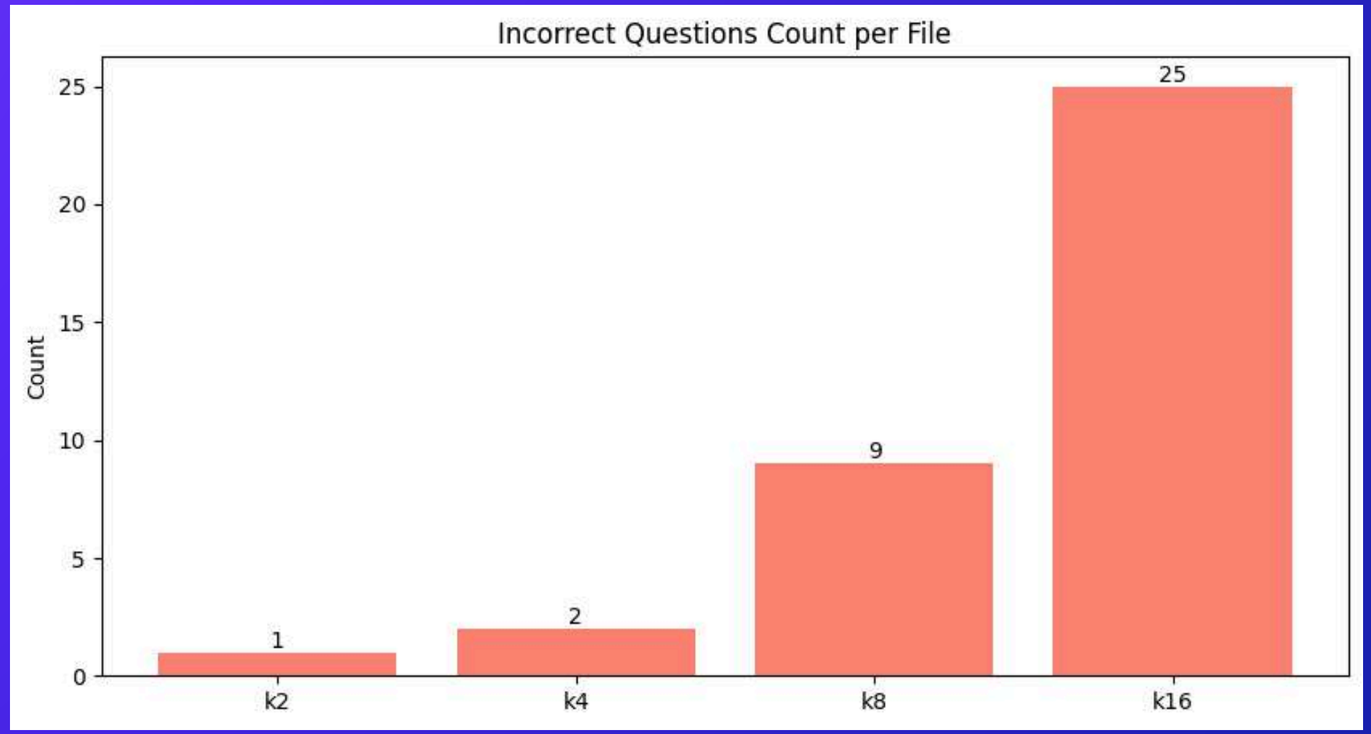
Context Overloading in RAG

⚠ Long Context ≠ Better Performance

- Even within the context window, LLMs struggle with very long inputs.
- Important info can be overlooked, leading to irrelevant or incorrect answers.
- LLMs do not "understand" — they rely on attention weights, which get spread thin in long inputs.

FINDINGS 1/4

- We noticed that indeed, the performance worsened as we increased the context paragraphs to extract answers from.
- K=16 means 16 paragraphs from SQuAD dataset were used and 5 questions from each paragraph
- As can be seen in the graph, K=16 (2256 tokens) marked 25/78 questions incorrect
- We used word matching with gold answers for evaluation.
- We found performance is acceptable around the 1000 token mark



FINDINGS 2/4

Testing Initial Approach

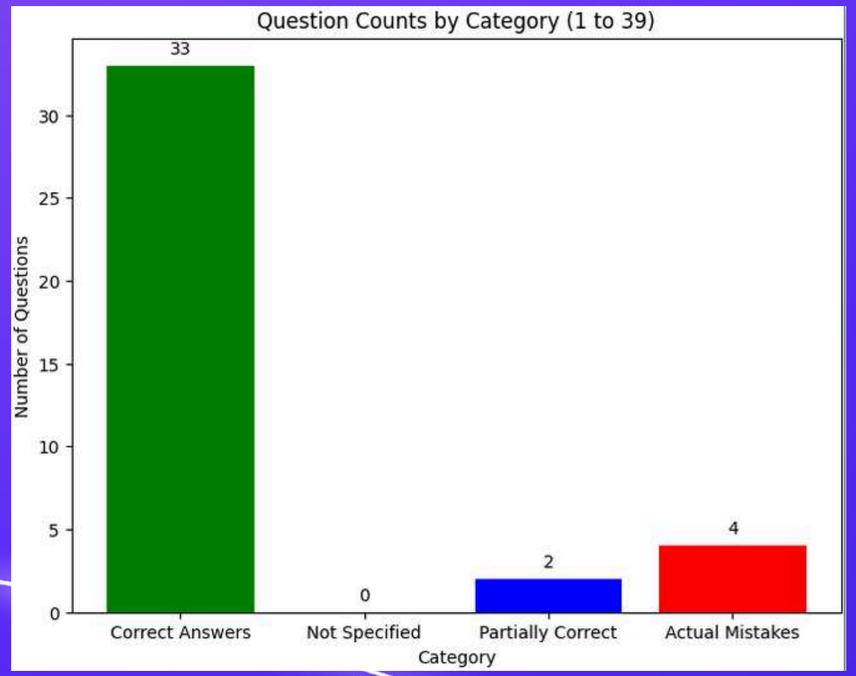
- Initial Approach was to split the context into 800 token chunks and pass all of the questions along with chunks and use prompt engineering to tell LLM to answer “Not specified” for the answers not found in contexts.
- However we found that the LLM we are using does not very often reply “Not Specified” to questions and it attempts to answer to its best ability from the context or sometimes even hallucinate the answer
- Therefore, we decided to revise our approach.

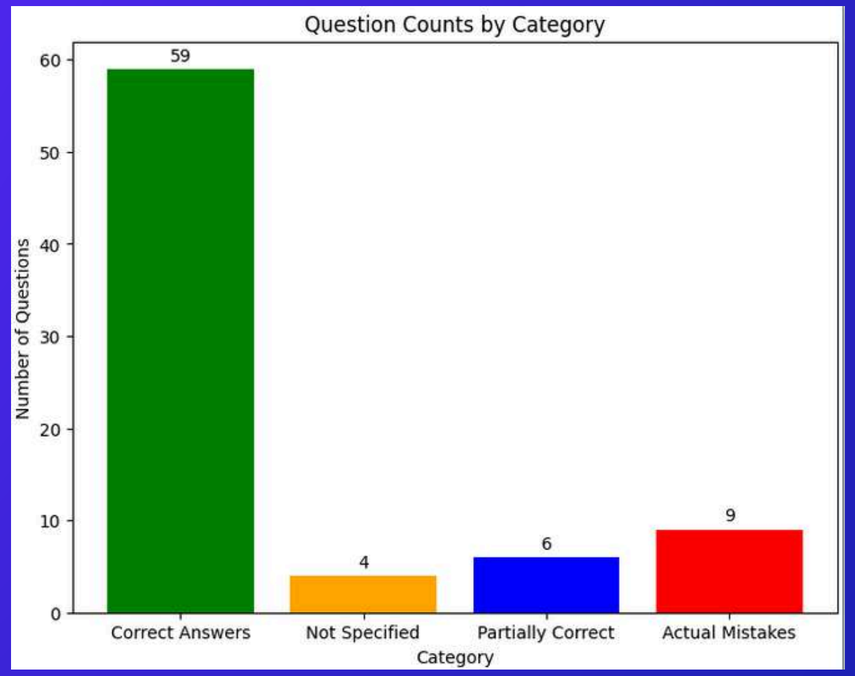
```
A9: The headquarters of the Congregation of the Holy Cross is Not specified.
A10: The primary seminary of the Congregation of the Holy Cross is Notre Dame Seminary in New Orleans, Louisiana.
A11: The oldest structure at Notre Dame is Not specified.
A12: The individuals who live at Fatima House at Notre Dame are Not specified.
A13: The number of BS level degrees offered in the College of Engineering at Notre Dame is Not specified.
A14: The College of Engineering at Notre Dame was formed in 1892.
A15: Similar studies were carried out at the College of Arts and Letters before the creation of the College of Engineering.
A16: There are 13 departments within the Stinson-Remick Hall of Engineering.
A17: The entity that provides help with the management of time for new students at Notre Dame is the First Year of Studies program.
A18: There are 4 colleges for undergraduates at Notre Dame.
A19: The First Year of Studies program was created at Notre Dame in 1962 to assist first year students.
A20: The First Year of Studies program at Notre Dame was declared "outstanding" by the American Council on Education.
```

FINDINGS 3/4

Testing Revised Approach (1/2)

- Revised Approach is to still chunk the context of the prompt into 800 tokens per chunk, but then, perform a pre-retrieval step of assigning questions to chunks which are likelier to have the answer. we used simple keyword matching for this objective.
- We observed that indeed, the performance does improve using this approach as there are reductions of 12% and 20% respectively for K=8 and K=16 in responding incorrectly.

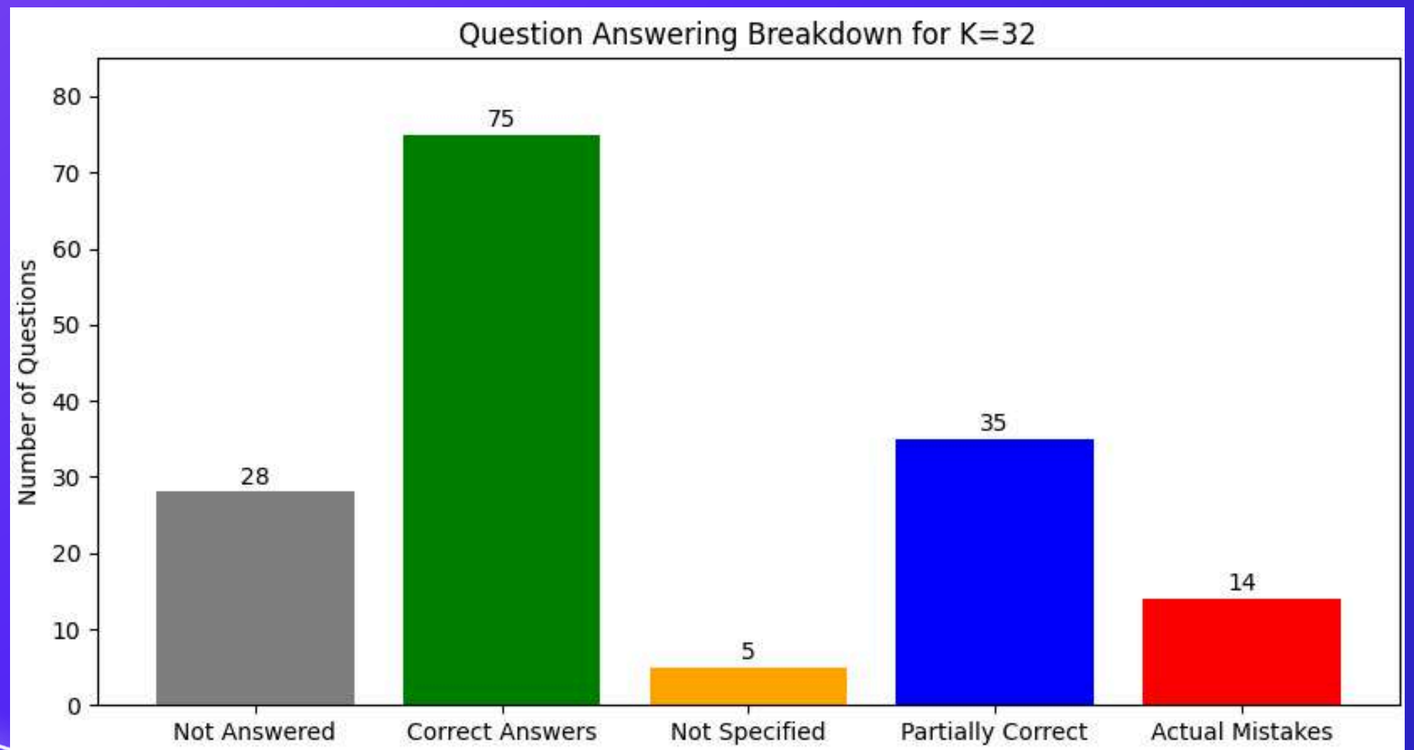




FINDINGS 4/4

Testing Revised Approach (2/2)

- For K=32, which is about 7000 tokens, passing the questions and context directly gave a timeout error on hugging face.
- However, using our revised approach of making chunks and assigning questions to them, we were able to get answers to 129/157 questions with an accuracy of 110/129 being correct (85%)



THANK YOU!

Group 3:
Hassan Ali
Muhammad Affan
Talha Tariq