Name= Shuvranshu Halder(MS RAP)
Project= RAG system

I have worked on" **hallucination**" of the RAG system.
Hallucination metric: how much of the generated answer is not from retrieved context.
This gives a value from 0 to 1.
1= highest hallucination(Worst case)
0= best case
I have achieved the goal in 3 stages by updating the model in every step.

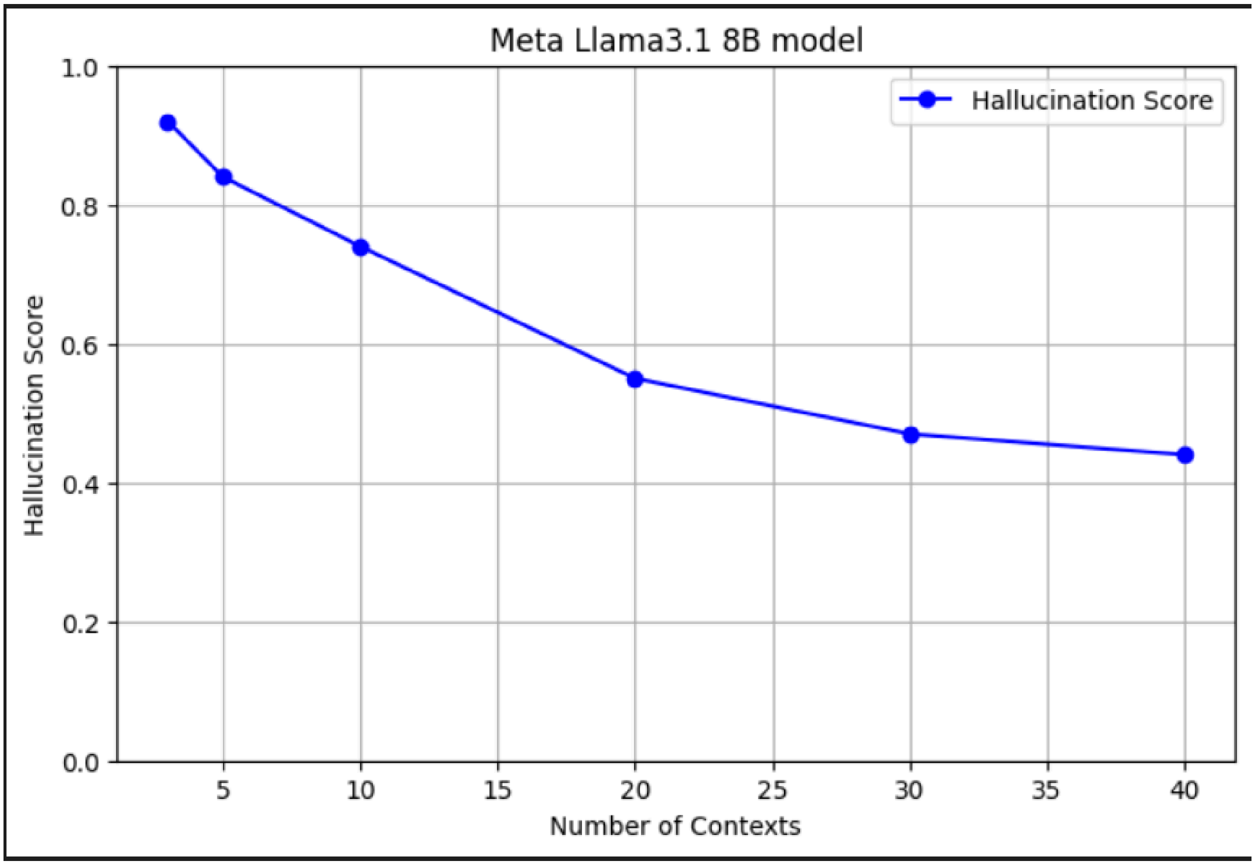
Stage 1: vanilla RAG

Base model: Meta Llama3.1 8b model
Embedding model: all-MiniLm-L6-V2
Benchmark datasets: SQUAD-V2(stanford question answers)

Workflow: user query-> retriever retrieves relevant documents from chroma db-> gives it to LLM along with the query->LLm generates the answer

The average hallucination score is tested on 100 queries.

No of contexts	**Avg Hallucination score**
3	0.92(very high)
5	0.84
10	0.74
20	0.55
30	0.47
40	0.44



Conclusion:
Initially very high hallucination. It depends on no of contexts retrieved. If we increase no of contexts , the hallucination score reduces.Although 40 contexts for every query is not practically possible.

Stage 2: Knowledge-Graph

I have read below research papers on hallucination.

Hallucination Mitigation for Retrieval-Augmented Large Language Models: A Review

Can Knowledge Graphs Reduce Hallucinations in LLMs? : A Survey

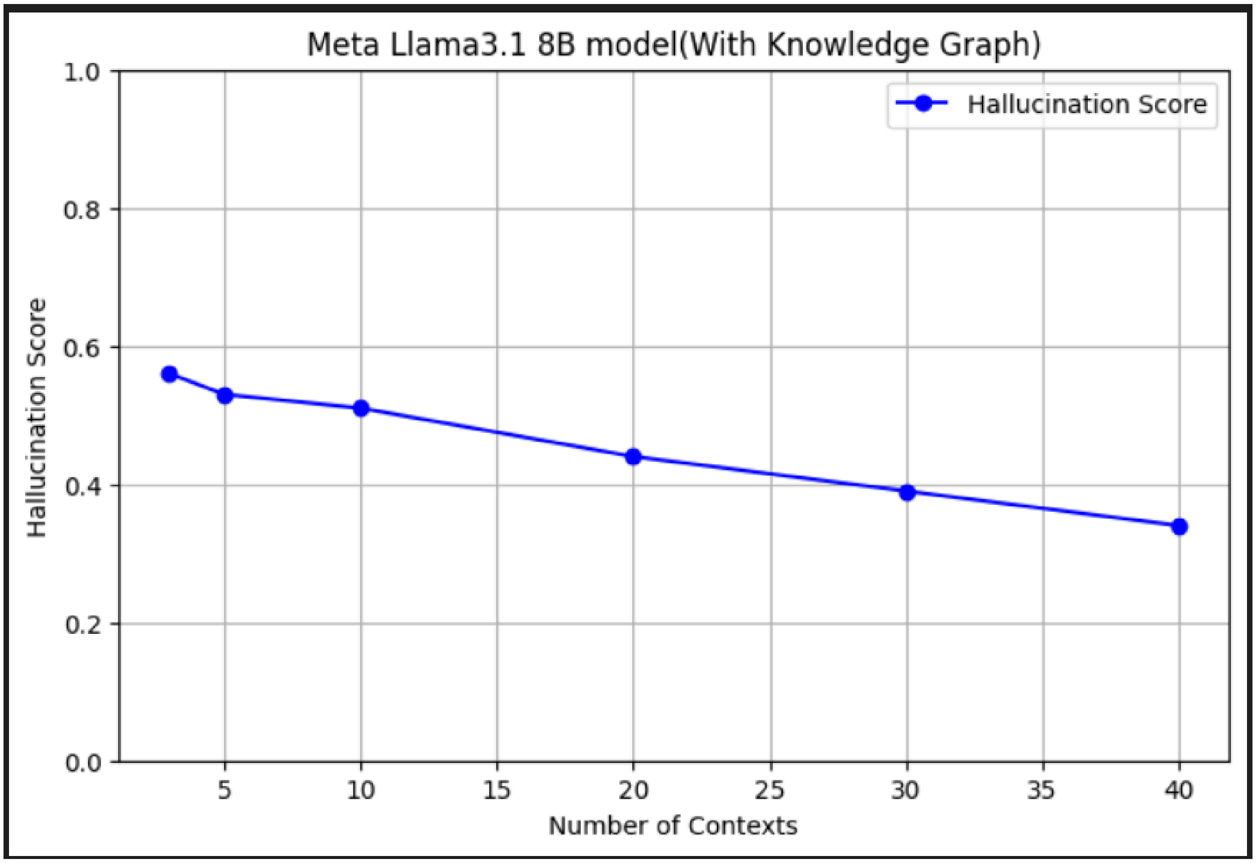
Found out that, one of the biggest reason of hallucination is, not properly formatted database. So i have implemented knowledge graph in the dataset so that relevant information retrieval become easier.

Implementation: knowledge graph stores relationship as a triple.
ex=the sentence “newton has discovered gravity” will be stored as (newton,discover,gravity).
I have used a small nlp model “en_core_web_sm” from “spacy” library that tokenizes the dataset and converts it into a knowledge graph.

Workflow:user query->retriver retrievs the relevent documents along with relevant triples-> gives it to llm->llm generates answer

This reduces hallucination score further.

No of contexts	**Avg hallucination score**
3	0.56
5	0.53
10	0.51
20	0.44
30	0.39
40	0.34



Conclusion: The result we previously achieved using 40 contexts has now been achieved using only 20 contexts.
Still not satisfactory but Much better.

Stage 3:chunking

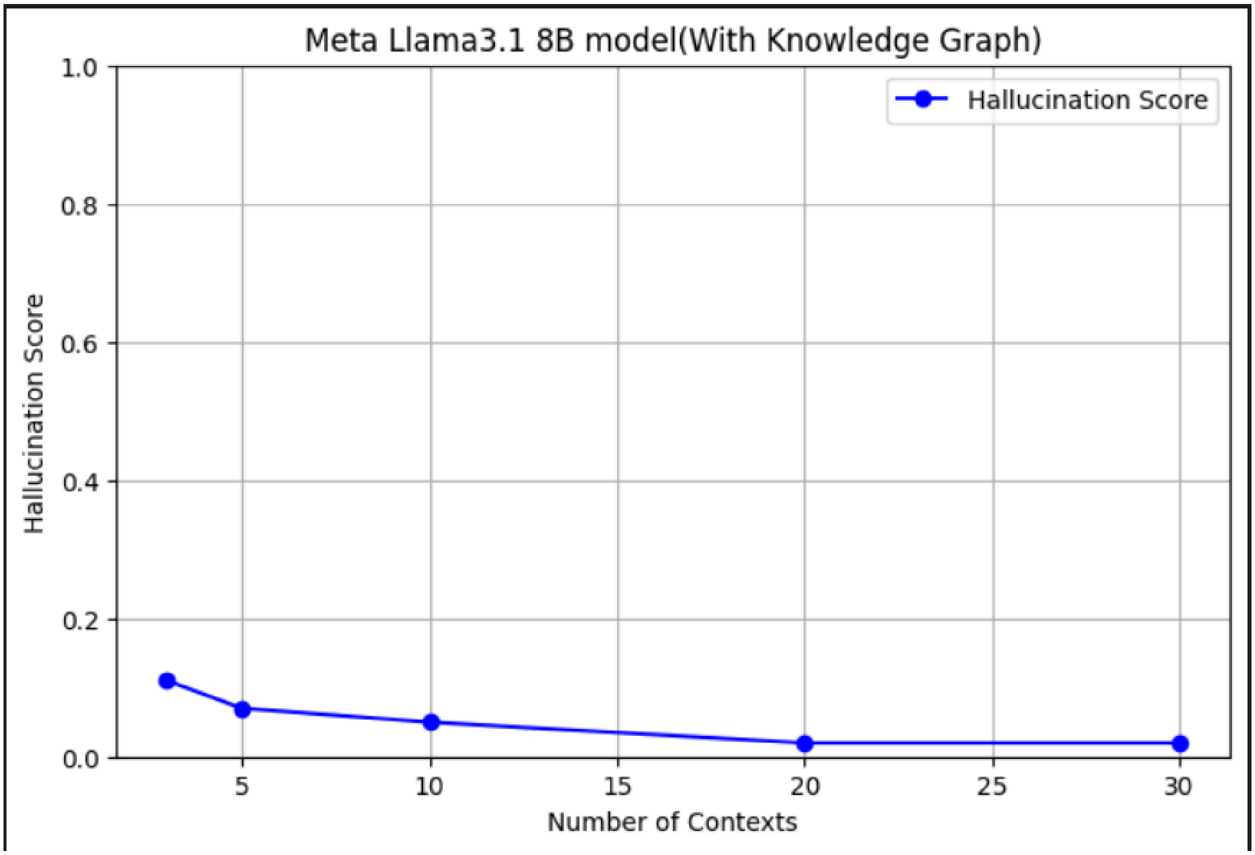
I have tried to improve it further. So i have read the below paper and tried to implement it.
Passage Segmentation of Documents for Extractive Question Answering

implementation= the document is chunked into paragraphs of 200-400 tokens and 50 overlaps, before storing into the database.
Rest all same as knowledge graph retrieval.

Workflow: same as before

This small update reduces the hallucination score significantly.

No of contexts	**Avg hallucination score**
3	0.11
5	0.07
10	0.05
20	0.02
30	0.02



Conclusion: now with just 5 contexts we have achieved hallucination score of 0.07 which is almost negligible.