Evaluation of chunking and storing strategies for increased Retrieval-Augmented Generation (RAG) accuracy

Harsh Vassaram, Computer Science
Mentor: Jia Zou, Assistant Professor
School of Computing and Augmented Intelligence

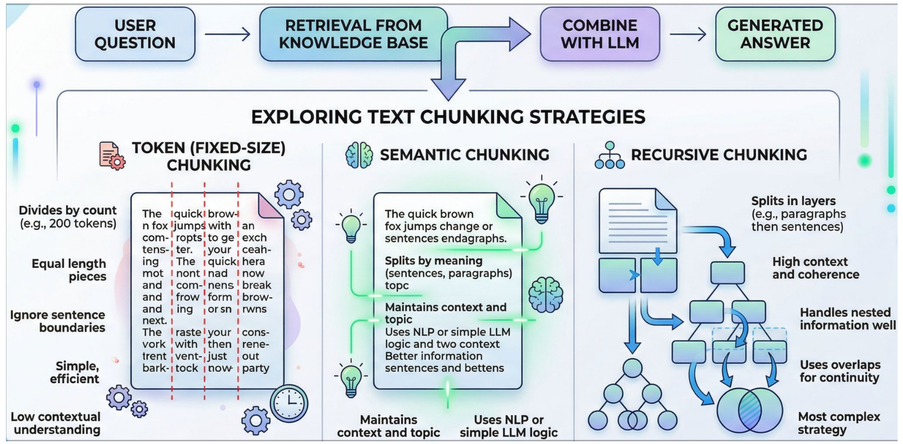
Introduction & Background

Large Language Models (LLMs) are severely limited by the context window of a user prompt, which relates to the size of information fed to the LLM, given the response accuracy is based on user query. Due to the limits of how much one can include in a prompt, incomplete queries lead to inaccurate responses from the LLMs. Therefore, using Retrieval-Augmented Generation (RAG) framework with LLMs, can aid in improved responses and reduced hallucinations (responding with no context).

Research Question

What is the impact of chunking strategies in the accuracy, latency & computing resources of Retrieval-Augmented Generation (RAG)?

Methodology

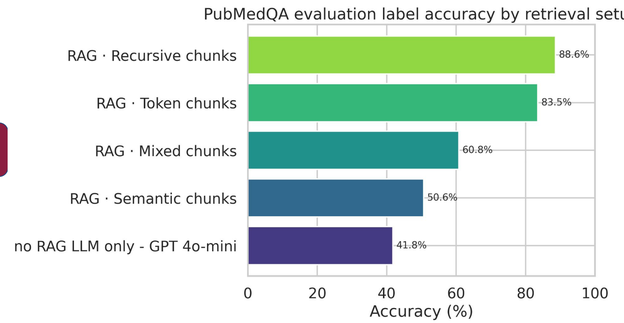


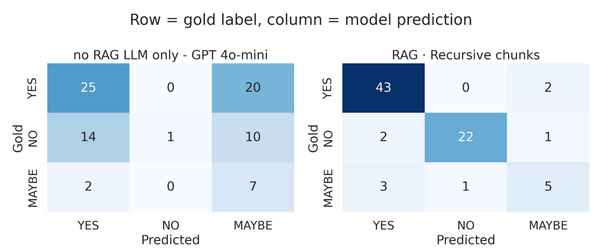
RAG comparison:
Baseline:
LLM only (no RAG - GPT 4o-mini)
mixed (top k, irrespective of chunking strategy)

Comparing chunking strategies:
Semantic (meaning based)
Token (fixed)
Recursive (hierarchical, structure based)

Data & Results

The chart below shows the accuracy percentages based on the type of chunking used in the RAG pipeline. Best accuracy; recursive chunking, due to the structured based article format, hence the increased accuracy.
The unexpected result of semantic chunking (meaning based), was hypothesized to have higher accuracy, but it may be due to the internal parameters of the semantic chunking that needs to be altered for this domain, and not use the default.





Conclusion

Discovered several different challenges, which led to conducting various experiments using an evaluation from a medical domain paper, for the current RAG implementation based on ingesting whole papers, and creating a Milvus lite vector database.
Evaluating different chunking strategies, given the domain specific corpus knowledge, is important in creating an improved RAG pipeline. The data ingested to the RAG pipeline, to be chunked and embedded into the vector database, checking for unstructured or structured data is initially vital, to ensure that the rest of the components do not intake incomplete or inaccurate data.

References

[1] G. Xiong, Q. Jin, Z. Lu, and A. Zhang, "Benchmarking retrieval-augmented generation for medicine," arXiv preprint arXiv:2402.13178, 2024.
[2] X. Wang, Z. Wang, X. Gao, F. Zhang, Y. Wu, Z. Xu, T. Shi, Z. Wang, S. Li, Q. Qian, and R. Yin, "Searching for best practices in retrieval-augmented generation," arXiv preprint arXiv:2407.01219, 2024.

Acknowledgements

I would like to express my gratitude to my mentor; Dr. Jia Zou, Dr. Arun Iyengar and fellow PhD student, Doug Oscarson, for guiding me throughout this research, in gaining deeper understanding to improve my skills, starting from a personal project to a research project.

FURI

Ira A. Fulton Schools of Engineering
Arizona State University