Alex Schell

COMP 5588

9/25/25

Advanced RAG Methodologies

As I was unfamiliar with RAG's for the most part until last week, I had a tough time understanding what these techniques were doing our data. Through this assignment and some extra research on my own I think I have a grasp on the topics. Reranking is a retrieval process that increases the quality of the retrieved documents in two parts. First part is the initial retrieval process, involving fast, broad search that retrieves many potentially relevant documents. The second part is the reranking, where a more sophisticated model scores and reorders these documents, selecting the most relevant ones for the final context. Context optimization involves techniques to make the most of the context window:

- Context compression: Removing redundant information while preserving key details
- Context selection: Choosing the most relevant passages rather than entire documents
- Context ordering: Strategically arranging retrieved content for optimal LLM processing

A multimodal RAG does the same thing as the traditional RAG, except it can handle multiple context types like text documents, images, videos, and audio. Because of this, there are now several options for retrieval:

- Text-to-text: Traditional semantic search
- Image-to-text: Finding relevant text based on image queries
- Text-to-image: Finding relevant images based on text queries
- Cross-modal search: Understanding relationships between different content types

Our datasets have tons of pictures that help guide the user during a certain project. The user will upload a picture of what they want fixed in their home and the model will give a response. Multimodal RAG allows for the user to query the database and have the model return with a set of images or diagrams along with some text to solve the issue.