Progress Report

Tags

Tasks That Are Completed

For this project, I am building a knowledge base store that is queryable via a chatbot. This would enable me to get capabilities like ChatGPT but grounded by knowledge that I have curated through my notes.

So far, I have managed to get most of the backend implementation working. I have an endpoint that lets me push documents (of free text) up to S3 and have it start an indexing process by Amazon Kendra. I have deployed LLaMA7b on Sagemaker that is used as an encoder for both these documents and queries. So when I ask a question through a POST endpoint, via Kendra, I am able to get back a list of documents that could be relevant.

I have also gotten these fetched fed to a Claude endpoint on Bedrock as Context and have it respond with an answer to my question.

Tasks That Are Pending

I have not started building any of the UI yet. That is still a large part that is very much pending, right now I am just testing via manually uploading doucments on S3 or using raw POST endpoints. This is a pretty bad experience, and at some point I will need to ensure that I back it via a chatbot like interface.

Challenges

The biggest challenge I am facing is on evaluation. There are so many models, both encoders and decoders, now that I am having a hard time knowing which model is good enough for my use case or if I need to keep exploring. It is also hard to get coverage on a wide variety of questions/answer pairs to ensure that I am actually able to say that this system performs accurately.

I need to maybe figure out how to get a golden dataset of questions answers over a wide variety of documents to ensure I am getting coverage over my entire set. It is also very hard to evaluate the encoder because it is always working in vector space.