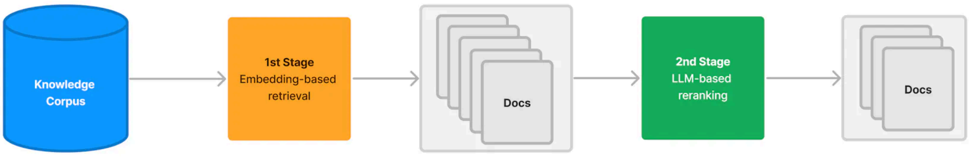
Knowledge Corpus
1st Stage Embedding-based retrieval
Docs
2nd Stage LLM-based reranking
Docs

Jerry Liu • May 17, 2023

Using LLM’s for Retrieval and Reranking

Artificial Intelligence Machine Learning Large Language Models Llamaindex NLP

Summary

This blog post outlines some of the core abstractions we have created in LlamaIndex around LLM-powered retrieval and reranking, which helps to create enhancements to document retrieval beyond naive top-k embedding-based lookup.

LLM-powered retrieval can return more relevant documents than embedding-based retrieval, with the tradeoff being much higher latency and cost. We show how using embedding-based retrieval as a first-stage pass, and second-stage retrieval as a reranking step can help provide a happy medium. We provide results over the Great Gatsby and the Lyft SEC 10-k.