Amazon Internal Chatbot

For VectorDB choices, refer to [WIP] AG-RAG VectorDB Deep Dive by Shreyash Iyengar.

Retrieval using Queries: it represent the queries into the same embedding space and sea

Comments: existing embedding models may not be powerful enough on some long-tail

3.3 LLM Inference

In this section, we concatenate the retrieved top k documents with the query and feed it
An LLM refines answers by incorporating information from relevant document snippets.

LLMs selection: most opensource LLMs can be accessed via the huggingface library. Sele

3.4 Evaluation

Evaluation is not a mandatory component for AutoRAG, but it would be beneficial to off

3.5 Model Fine-tuning (Ongoing Research Project)

In the future, we expect the users can easily finetune the embedding model as well as th

Haoyang Fang is working on this in parallel with this project

Message chatbot...