## Project Title: Intelligent Knowledge Base with Retrieval-Augmented Generation

Problem and Solution

Large language models (LLMs) like GPT-3.5 are powerful, but they are limited by the static training data they were exposed to. As a result, they often underperform in specialized, high-knowledge domains such as aerospace, biotech, or industrial diagnostics—domains where context evolves, and the latest knowledge isn't part of a frozen training set. To resolve these issues, supplemental information can be added to aid in knowledge transfer. This process is known as Retrieval-Augmented Generation (RAG).

Dataset and Technology

The primary dataset for this effort was built by leveraging Python with Beautiful Soup to gather information from a publicly available online forum, The data was processed through SentenceTransformer embeddings, FAISS vector search, and OpenAI GPT-3.5, ultimately served in a Streamlit web interface allowing users to explore the results.

Uses and Benefits

This project demonstrates how enterprises can enhance commercial LLMs with domain-specific knowledge. By implementing RAG, organizations can extend capabilities into specialized domains without retraining, improve response accuracy for technical queries, keep knowledge current by updating the corpus independently, preserve institutional knowledge, and reduce hallucinations by grounding responses in verified sources.

Drawbacks

RAG systems face several limitations: response quality depends on corpus accuracy, potentially propagating misinformation; semantic matching limitations when terminology differs between queries and sources; higher computational costs for both vector database management and commercial LLM APIs; increased architectural complexity compared to standard LLM implementations; and inheritance of biases present in the original community discussions.

Challenges

The initial challenge with this effort was Beautiful Soup programming specific to the forum platform (XenForo 2.3). Additionally, attempts to train a custom RAG model produced unusable results. Migrating to a semantic model combining vector database with a commercial generator yielded better outcomes. Finally, creating an inexpensive, ephemeral model proved difficult as commercial ML hosting solutions struggled with the implementation.

Results and Impact

The results show that the supplemental RAG corpus contained a richness of data not found in the LLM. This satisfies the original use case for this effort.

A close-up of a page

Description automatically generated

*\* Note that the OpenAI response references an airbag which is not a technology available in these older automobiles.*

Demo Resources

YouTube 2min: <https://youtu.be/Vf9RgCn8ow0>

YouTube 15m**:** <https://youtu.be/xAFkszSArU8>