Nuera-AI: The Cognitive Agent

Abstract:

This project is an interactive research assistant that integrates multiple advanced technologies to facilitate efficient knowledge management and retrieval. At its core, the system leverages a SQLite database for persistent storage of research material, while incorporating a FAISS index and SentenceTransformer to generate semantic embeddings. These embeddings enable fast, similarity-based searches, ensuring that relevant content is efficiently retrieved in response to user queries. The application further enhances its utility by incorporating an auto-update feature that fetches concise summaries from Wikipedia, thereby expanding the knowledge base with verified, authoritative information.

The research assistant employs a generative AI model—specifically, Google's FLAN-T5-large—to generate detailed and contextually relevant responses. The model is fine-tuned using prompt engineering techniques, combining user-provided context with tailored instructions to produce elaborate answers. This integration of retrieval-augmented generation (RAG) not only improves the quality and accuracy of the responses but also supports a dynamic interaction between stored knowledge and real-time user queries. The project is implemented using Python and Streamlit, providing an intuitive web interface that streamlines user interaction and knowledge exploration.

Overall, the project exemplifies a modern, agent-based approach to research and information retrieval. By combining persistent data storage, advanced embedding techniques, and state-of-the-art generative models, it provides users with a robust tool for managing and querying large volumes of research material. The system's modular design allows for incremental updates and scalability, making it a versatile solution for academic, corporate, or personal knowledge management applications.