**Tools, Libraries, and Frameworks**

The system is built using a combination of modern AI development tools and frameworks that facilitate modular design and rapid experimentation. The following libraries were integrated into the project:

* **LangChain**: For constructing the RAG pipeline and managing document loading, text splitting, vector storage, and LLM interaction.
* **ChromaDB**: Used as the vector store to perform semantic search on the knowledge base.
* **Sentence Transformers (via Hugging Face)**: Provides the embedding model (all-MiniLM-L6-v2) for document chunk vectorization.
* **Ollama**: A lightweight framework used to locally run open-source LLMs such as llama3.
* **AsyncIO**: Employed to manage asynchronous interactions, enabling a smooth, responsive user experience.
* **Matplotlib**: Used to visualize response time analytics for user interactions.

This combination of tools ensures a flexible and extensible architecture while keeping the system lightweight enough to run on a modest local machine