

DocSage: chat with your documents locally

<https://github.com/KedoKudo/DocSage>

Control Panel

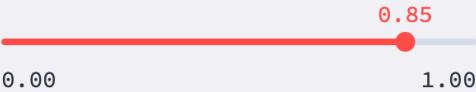
Select the Model

codellama:latest

▼

Size/GB: 3.56

Temperature



Upload a file for RAG

Drag and drop files here

Limit 200MB per file • PDF, TXT

Browse files

Select the knowledge base

None

▼

DocSage

What is DocSage?

- A simple RAG (Retrieval-Augmented Generation) model for local document search and chat.
- Backend:
 - **Ollama**: a local-host LLM server that allows users to run different LLMs models directly.
 - **Langchain**: a mature framework that allows users to send context embeddings to LLM service for context aware text generation.
- Frontend:
 - **Streamlit**: a low-coding Python library that allows users to spin up a simple web applications.

Ollama(🦌)



<https://ollama.com>, <https://github.com/ollama/ollama>

```
brew install ollama  
brew services start ollama  
ollama run llama2
```

Langchain

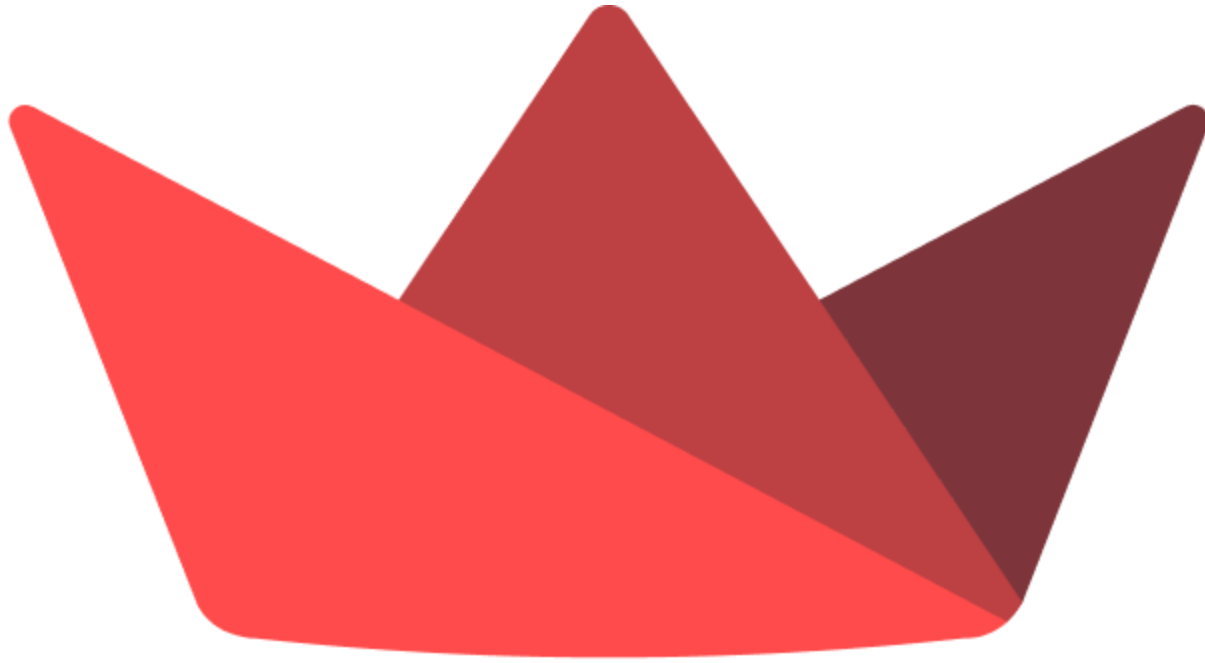


LangChain

<https://python.langchain.com>, <https://www.langchain.com>

```
pip install langchain
```

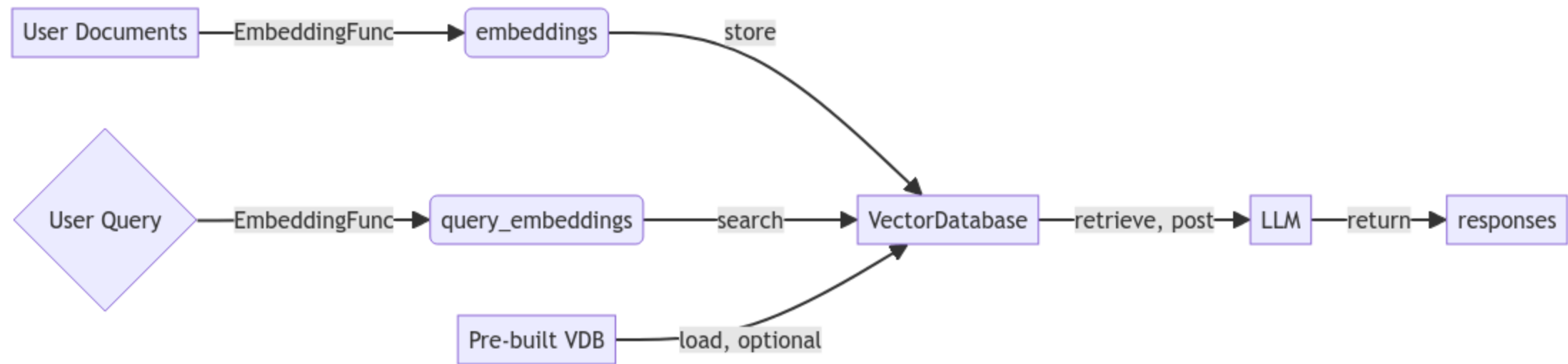
Streamlit



<https://streamlit.io>, <https://github.com/streamlit/streamlit>

```
pip install streamlit
```

What is RAG?



Demo

Case 1: simple interaction with DocSage

- testing prompt: "What is Mantid?"
- temperature: 0.1, 0.5, 0.9

Case 2: chat with your documents

- testing prompt: "What is ornl-next?"
- temperature: 0.1, 0.5, 0.9

Case 3: chat with knowledge base

- testing prompt: "Provide a python script to calibrate PowGEN detector with mantid"
- temperature: 0.1, 0.5, 0.9
- knowledge base: Mantid

Q&A