Sourcery.info

Your private, secure investigative journalism AI

Introduction

Sourcery.info lets you interview a cache of documents using natural language, all within a secure environment. It helps you find answers to questions you didn't know to ask.

Large language models combined with source documents has created a potential treasure trove for investigative journalists, OSInt researchers, or anyone who needs to glean insights from a document cache. However, the process typically involves uploading the documents to third parties, which is inherently insecure and in many cases dangerous.

With Sourcery.info, you run your own large language model, process the documents locally, and query them locally. You can run it on your own network, on your PC, on a virtual PC, or even in an air gapped environment.

Sourcery.info uses an AI methodology called retrieval-augmented generation (RAG) to combine both your own unique knowledge base (the document cache) with the power of a large language model.

Apart from the ability to search through documents using natural language, the RAG-based AI method has the potential of surfacing interesting, useful information to researchers who don't know exactly what they're looking for. This can be particularly useful for large caches of document leaks, for instance.

Sourcery.info takes the complexity out of RAG by optimising the generation of embeddings (the LLM codes that turn your documents into vectors), deciding on chunk size, using strategies like small-to-big, and selecting the appropriate language model for your use case.

The user interface is simple without losing the power to inspect the documents manually. It includes features to ingest documents, create the embeddings, query the documents using natural language, and view the original sources along with the results.

It will:

- Ingest a cache of documents;
- OCR and extract the data as necessary;
- Generate embeddings locally;

- Store the embeddings in a local vector database;
- Use local LLMs in a RAG model to query the documents using natural language;
- Present results with the associated sources for easy reference to the sources.

It won't:

- Send any information out of the system;
- Load any external dependencies;
- Censor or refuse to answer questions that could be sensitive;
- Store any usage information, unless required for auditing purposes (with a specific "opt-in").