Chosen Components:

Vector Store:

We decided to go with ChromaDB because it is open-source, does semantic search, NLP and integrates with LangGraph. It has a user-friendly API. It performs well with our collected data, supports our embedding model (Sentence Transformers), has long term storage, runs entirely off memory. And works both local and client-server deployments

Embedding Model:

We chose Sentence transformers because they can perform semantic similarity between sentences and paragraphs. They encode entire sentences or text fragments into vectors, keeping contextual meaning and enhancing NLP application performance. They are good with comparing articles to queries using cosine similarity. Sentence Transformers works well with BERT.

Text Splitter:

We have decided to go with sentence chunking to maintain context.

LLM:

We have picked llama3.2:1b-instruct-q4\_0 as it is completely free, small, open-source and is easier to integrate into our pipeline than GPT4 or Gemini. allows developers to modify and integrate the model into unique use cases like language translation tools and efficient chatbots.