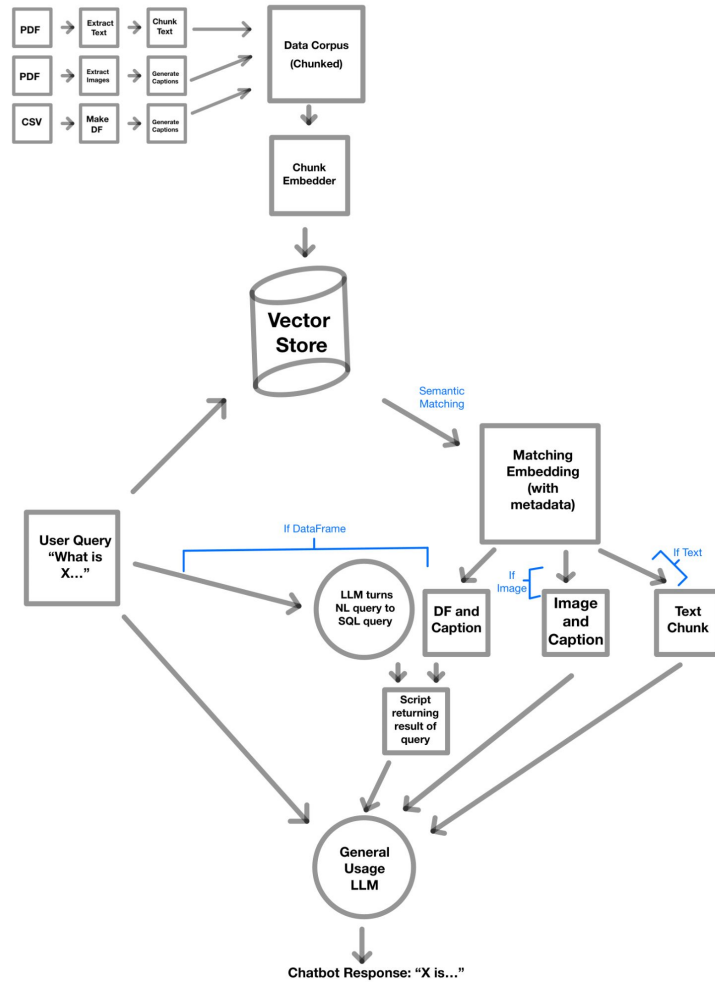# ChatGBeeT Pricing Options

+ System Architecture

1. Data Ingestion and Preprocessing

This stage involves extracting and preparing data from PDFs and CSVs.

A. PDF Text Extraction

Tool	Description	Cost
PyPDF2	Library for extracting text from PDFs.	Free (open-source)
PDFPlumber	Library for extracting text and images from PDFs.	Free (open-source)

B. PDF Image Extraction

Tool	Description	Cost
PDFPlumber	Library for extracting images from PDFs.	Free (open-source)
PyMuPDF	Library for extracting images and text from PDFs.	Free (open-source)

C. CSV Processing

Tool	Description	Cost
Pandas	Library for converting CSVs into DataFrames.	Free (open-source)
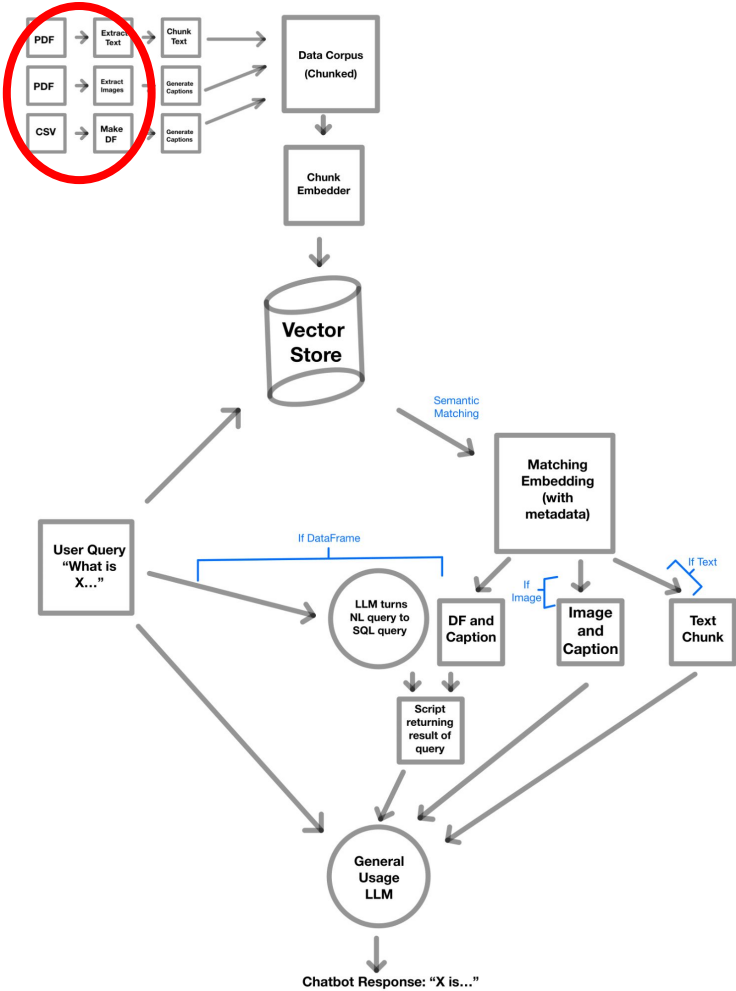
2. Data Chunking and Captioning

This stage involves chunking text, generating captions for images and DataFrames, and preparing the data for embedding.

A. Text Chunking

Tool	Description	Cost
LangChain	Framework for chunking text and managing embeddings.	Free (open-source)

B. Image Captioning

Tool	Description	Cost
GPT-4 Vision (GPT-4V)	Generates detailed captions for images.	$0.03/1k tokens
BLIP-2 (Hugging Face)	Open-source model for image captioning.	Free (open-source)

C. DataFrame Captioning

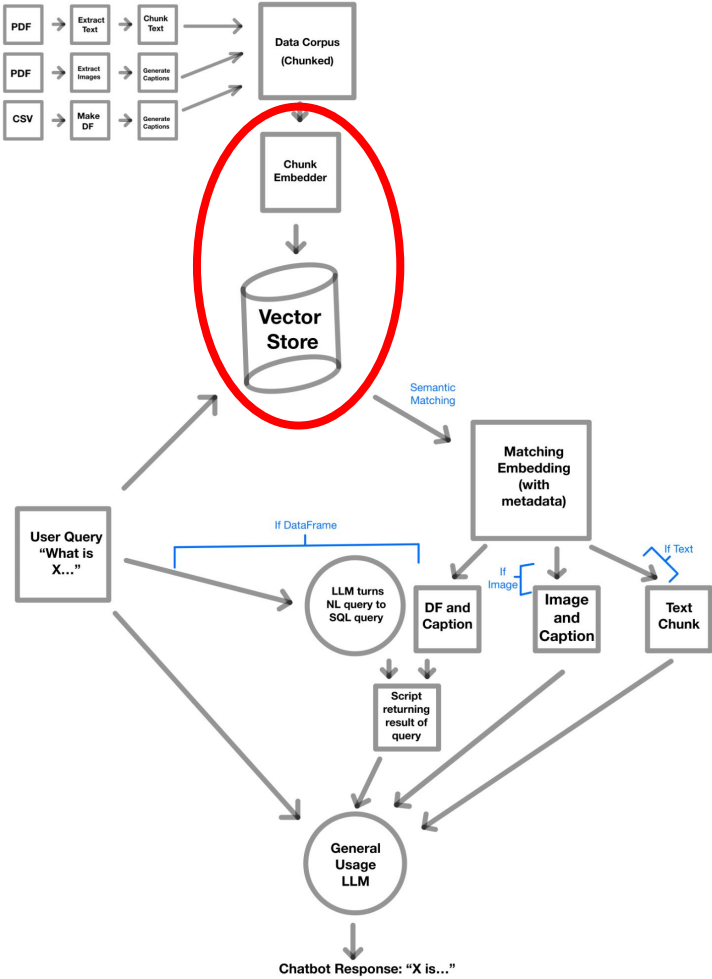| Tool	Description	Cost
GPT-4	Generates captions describing the content of DataFrames.	$0.03/1k tokens (input), $0.06/1k tokens (output)
Hugging Face	Open-source models (e.g., Llama-2) for generating	
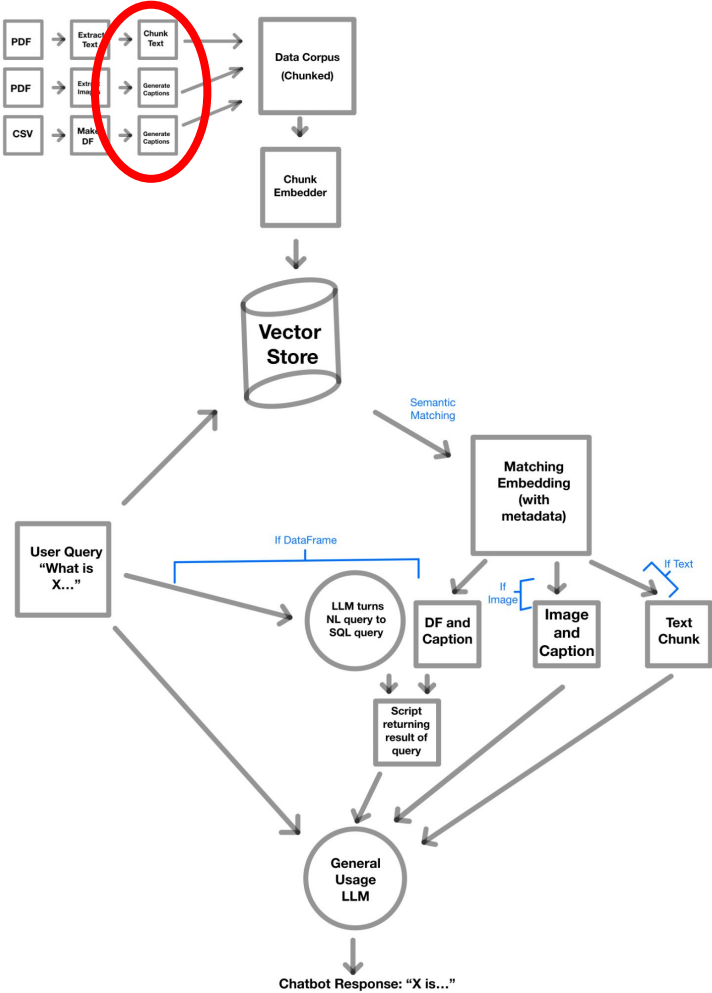
3. Embedding and Vector Storage

This stage involves embedding the data and storing it in a vector database.

A. Text and Caption Embedding

Tool	Description	Cost
OpenAI Embedding API	High-quality embeddings for text and captions.	$0.0001/1k tokens
Cohere Embed	Competitive embeddings for text and captions.	$0.40/1k embeddings
Hugging Face Transformers	Open-source embedding models (e.g., `all-MiniLM-L6-v2`).	Free (open-source)

B. Vector Storage

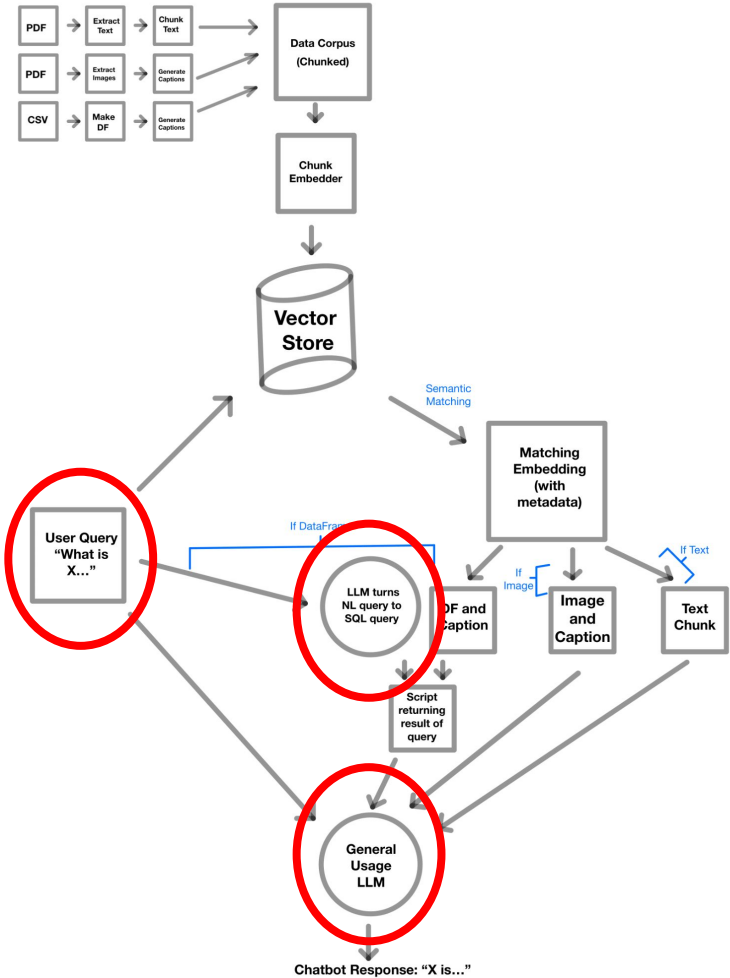| Tool	Description	Cost
Pinecone	Fully managed vector database for large-scale similarity search.	$70/month (5M vectors)
Weaviate	Open-source vector database with hybrid search capabilities.	$25/month (managed service)
FAISS	Open-source library for similarity search (requires self-hosting).	Free (open-source)

4. Query Processing and Response Generation

This stage involves processing user queries, retrieving relevant data, and generating responses.

A. Natural Language Query to SQL Query

Tool	Description	Cost
Fine-tuned LLM	Translates natural language queries into SQL queries for DataFrame retrieval.	Depends on LLM used (e.g., GPT-4)

B. General Usage LLM

| Tool	Description	Cost
OpenAI GPT-4	Generates high-quality responses to user queries.	$0.03/1k tokens (input), $0.06/1k tokens (output)
Anthropic Claude	Competitive LLM for generating responses.	$0.01/1k tokens (input), $0.03/1k tokens (output)
Hugging Face LLMs	Open-source LLMs (e.g., Llama-2) for generating responses.	Free (open-source)

5. User Interface

This stage involves building an interface for users to interact with the chatbot.

Tool	Description	Cost
Streamlit	Library for building interactive web interfaces.	Free (open-source)
Gradio	Library for building interactive web interfaces.	Free (open-source)

PDF → Extract Text → Chunk Text →

PDF → Extract Images → Generate Captions →

CSV → Make DF → Generate Captions →

Data Corpus (Chunked)

Chunk Embedder

Vector Store

Semantic Matching

Matching Embedding (with metadata)

User Query "What is X…"

If DataFrame

LLM turns NL query to SQL query

DF and Caption

If Image

Image and Caption

If Text

Text Chunk

Script returning result of query

General Usage LLM

Chatbot Response: "X is…"

6. Summary of Costs

Stage	Estimated Cost	Notes
Text Embedding (PDFs + TXT)	Negligible (<5$)	One-time cost.
Image Captioning (PDFs)	Negligible (<5$)	One-time cost.
CSV Captioning	Negligible (~1$)	One-time cost.
Total Embedding Cost	Negligible (~5$)	One-time cost.
Vector Storage	~$25/month	Ongoing cost.
Query Processing and Response Generation	~$30/month (LLM API calls)	Usage-based cost.
User Interface	Free (open-source tools)	
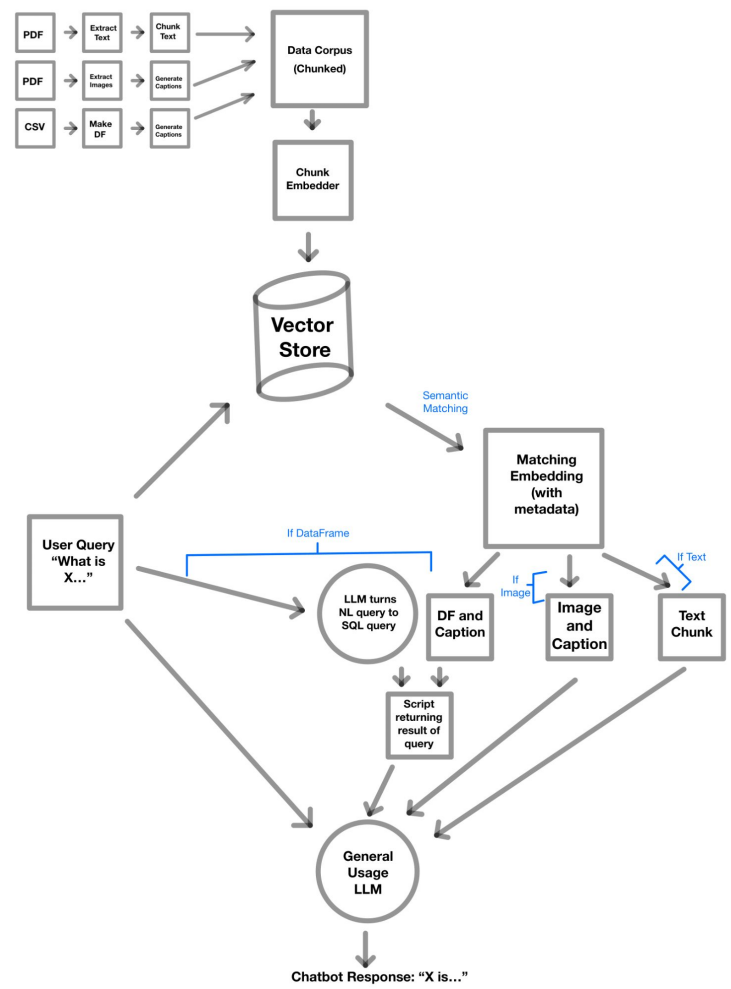
- **Questions and Concerns**
  - **Sorting which data to return and when**
  - **Keeping conversational context**

- **More Resources:**
  - **Open Source Models: https://github.com/eugeneyan/open-llms**
  - **UCSB Virtual Machines: https://www.library.ucsb.edu/dreamlab/virtual-machines-and-remote-software-access**
  - **https://csc.cnsi.ucsb.edu/get-started**