Big Goal

Build a web-based PDF chat app

1. User logs in
2. User uploads a PDF
3. We generate embeddings
4. User asks question about the PDF
5. We find relevant documents
6. Put question + docs into a LLM
7. Show user the output
8. User can 'like' or 'dislike' the answer

Err...Didn't we *just* build this?

There are some interesting challenges that come up when adding text generation into a web app!

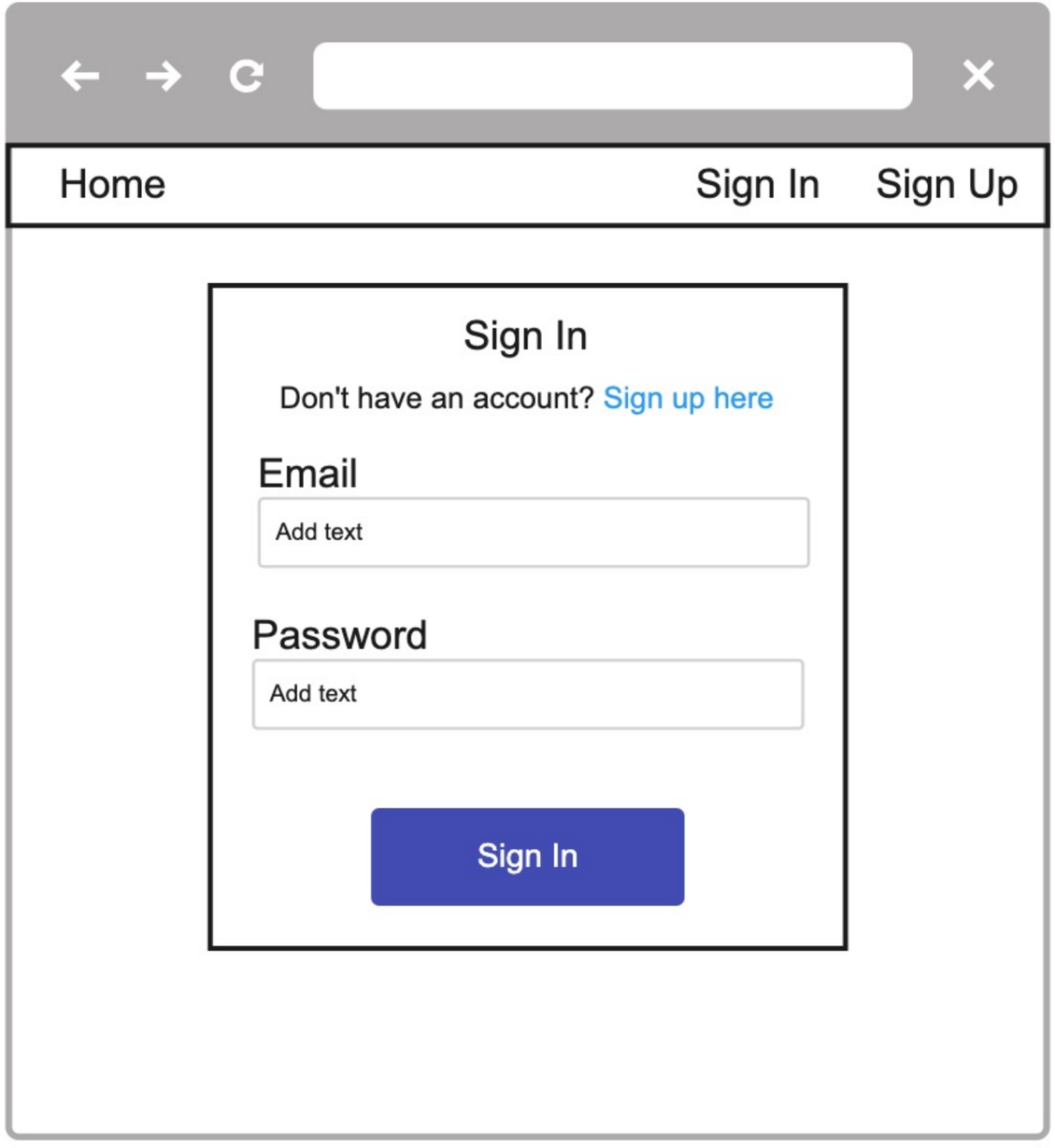
Plan

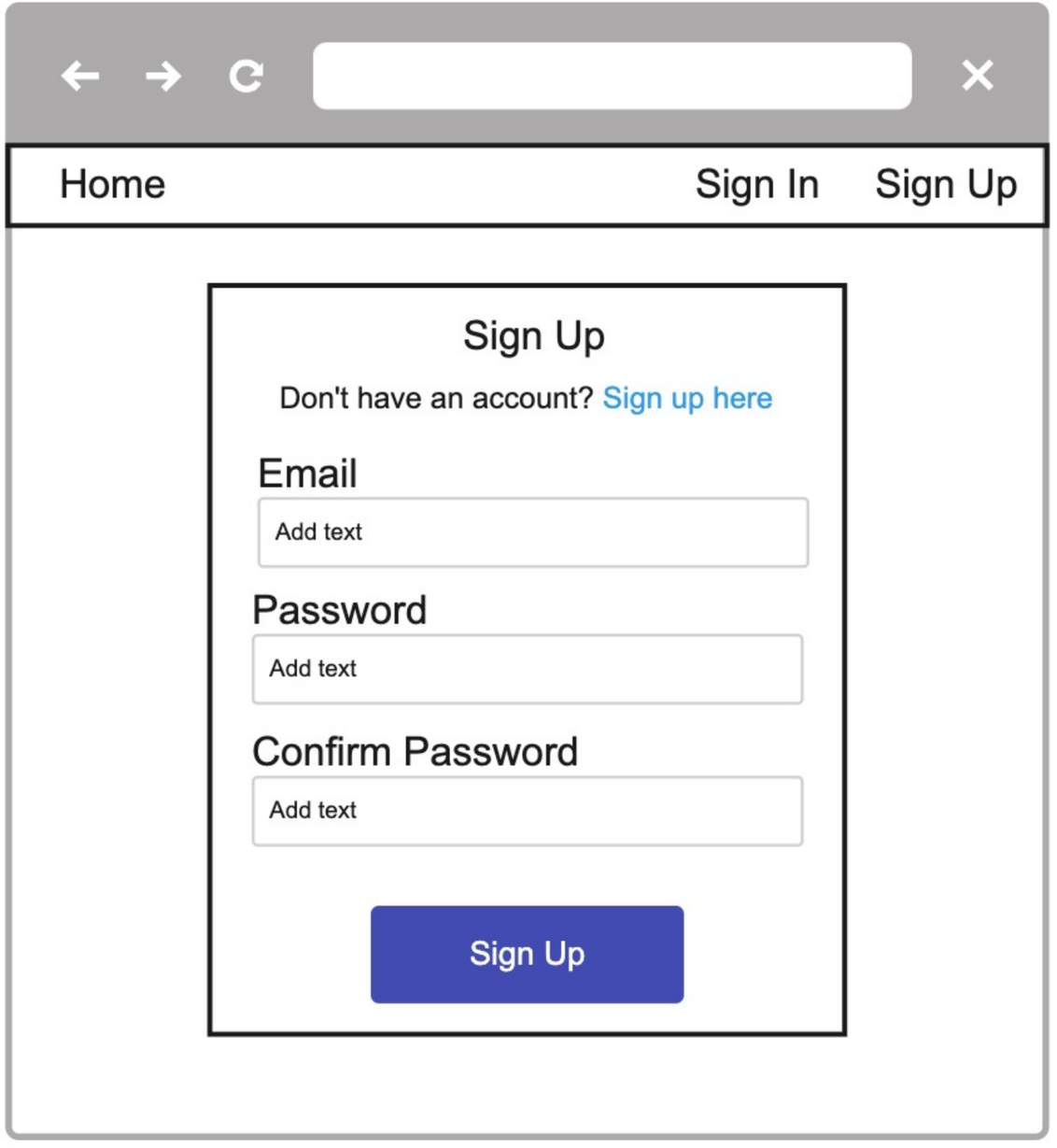
Look at some mockups to understand exactly what we're building

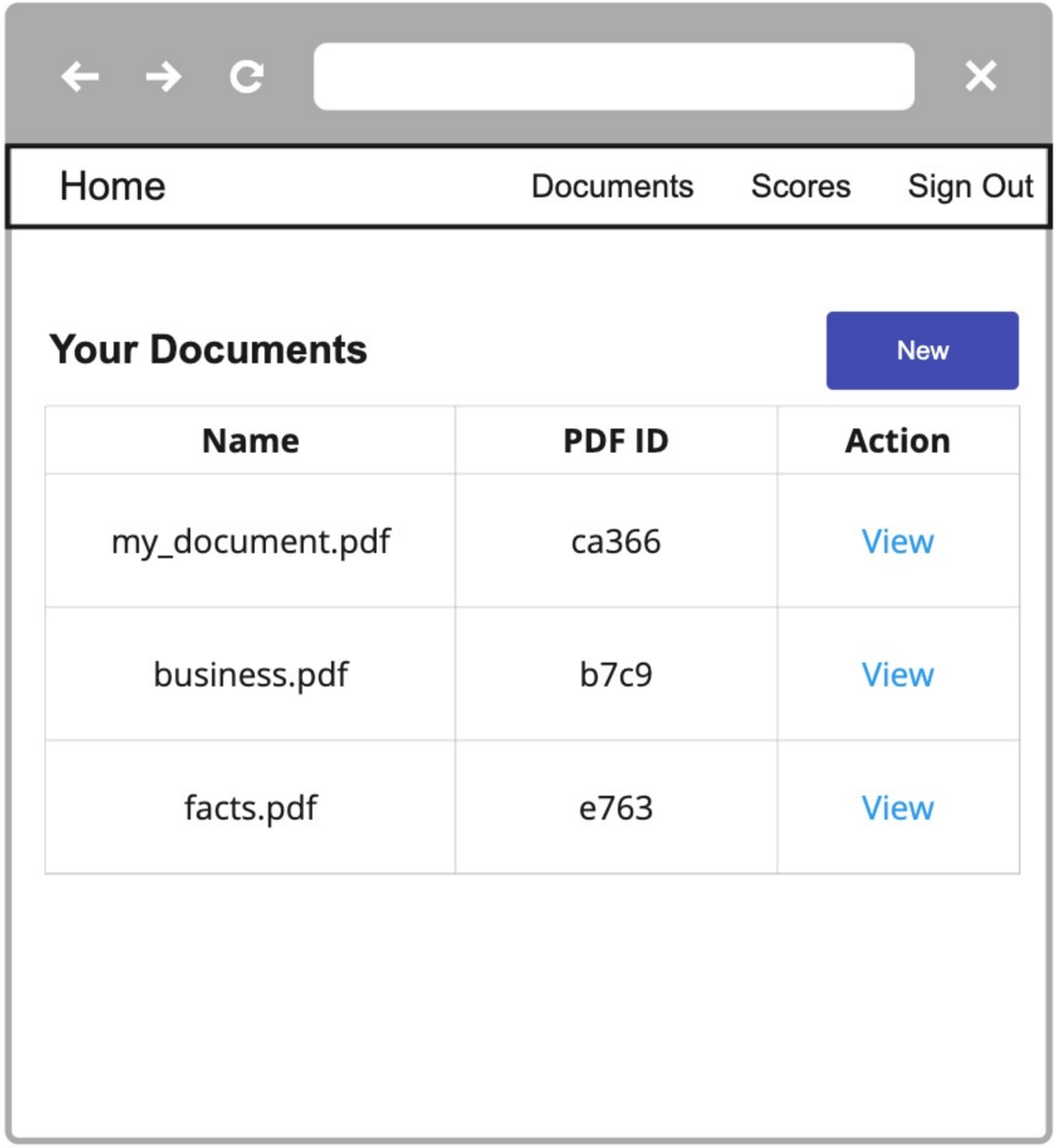
Download + setup some boilerplate I wrote for us

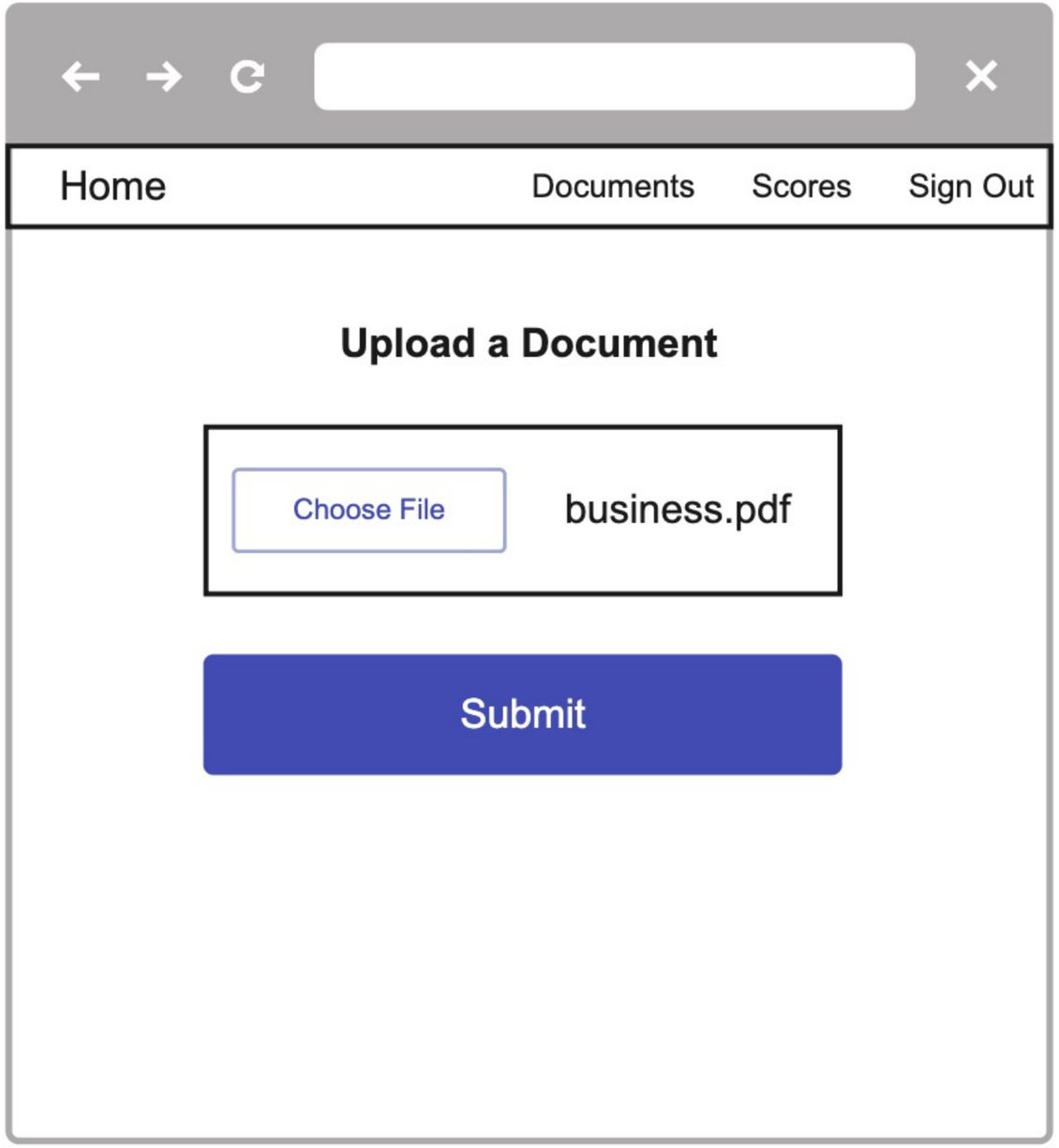
Focus on LangChain stuff

Most of the non-langchain code is already written


Home
Sign In
Sign Up
Sign In
Don't have an account? Sign up here
Email
Add text
Password
Add text
Sign In


Home
Sign In
Sign Up
Sign Up
Don't have an account? Sign up here
Email
Add text
Password
Add text
Confirm Password
Add text
Sign Up


Home
Documents
Scores
Sign Out
Your Documents
New
Name
PDF ID
Action
my_document.pdf
ca366
View
business.pdf
b7c9
View
facts.pdf
e763
View


Home
Documents
Scores
Sign Out
Upload a Document
Choose File
business.pdf
Submit

Home

Documents Scores Sign Out

History New Chat

What is this PDF about?

This PDF is about the design of transistors for use in computer circuitry

How is a transistor made?

A computer transistor is made by doping semiconductor materials to create junctions that can control electrical current.

alskdfjlaksjdflkasjdf

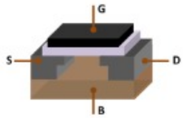
WIKIPEDIA
The Free Encyclopedia

Transistor

A **transistor** is a semiconductor device used to amplify or switch electrical signals and power. It is one of the basic building blocks of modern electronics.[1] It is composed of semiconductor material, usually with at least three terminals for connection to an electronic circuit. A voltage or current applied to one pair of the transistor's terminals controls the current through another pair of terminals. Because the controlled (output) power can be higher than the controlling (input) power, a transistor can amplify a signal. Some transistors are packaged individually, but many more in miniature form are found embedded in integrated circuits. Because transistors are the key active components in practically all modern electronics, many people consider them one of the 20th century's greatest inventions.[2]

Size comparison of bipolar junction transistor packages, including (from left to right): SOT-23, TO-92, TO-126, and TO-3

Physicist Julius Edgar Lilienfeld proposed the concept of a field-effect transistor in 1926, but it was not possible to construct a working device at that time.[3] The first working device was a point-contact transistor invented in 1947 by physicists John Bardeen, Walter Brattain, and William Shockley at Bell Labs; the three shared the 1956 Nobel Prize in Physics for their achievement.[4] The most widely used type of transistor is the metal–oxide–semiconductor field-effect transistor (MOSFET), invented by Mohamed Atalla and Dawon Kahng at Bell Labs in 1959.[5][6][7] Transistors revolutionized the field of electronics and paved the way for smaller and cheaper radios, calculators, computers, and other electronic devices.



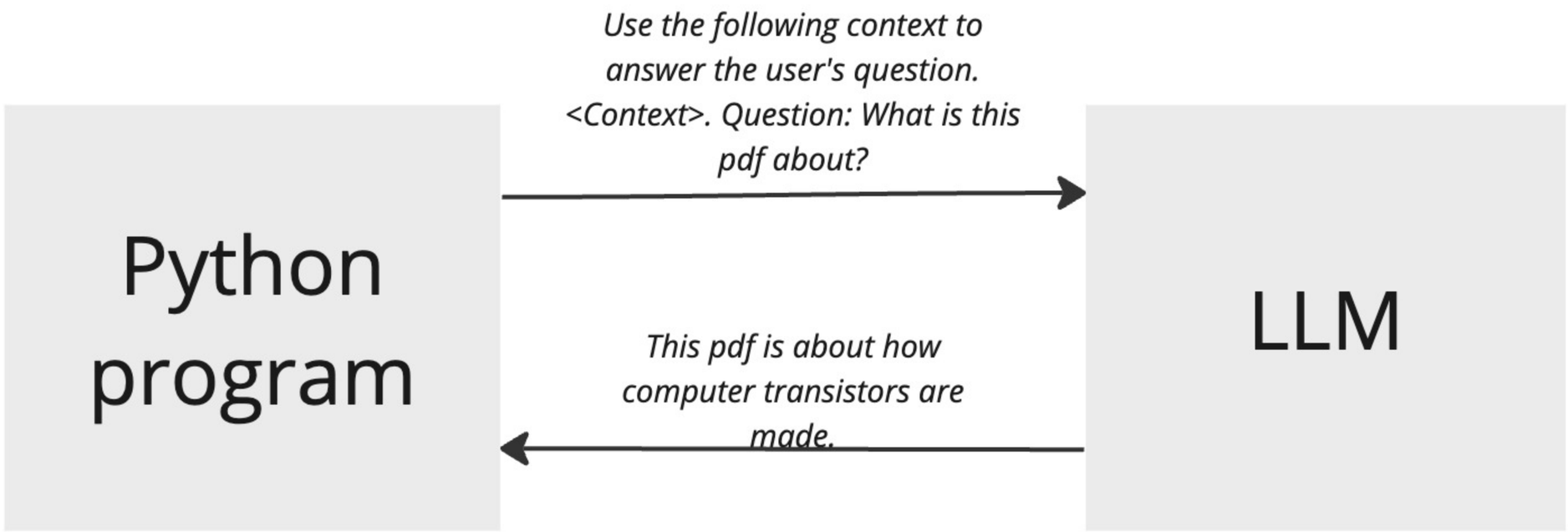
Metal–oxide–semiconductor field-effect transistor (MOSFET), showing gate (G), body (B), source (S) and drain (D) terminals. The gate is separated from the body by an insulating layer (white).

Most transistors are made from very pure silicon, and some from germanium, but certain other semiconductor materials are sometimes used. A transistor may have only one kind of charge carrier in a field-effect transistor, or may have two kinds of charge carriers in bipolar junction transistor devices. Compared with the vacuum tube, transistors are generally smaller and require less power to operate. Certain vacuum tubes have advantages over transistors at very high operating frequencies or high operating voltages. Many types of transistors are made to standardized specifications by multiple manufacturers.

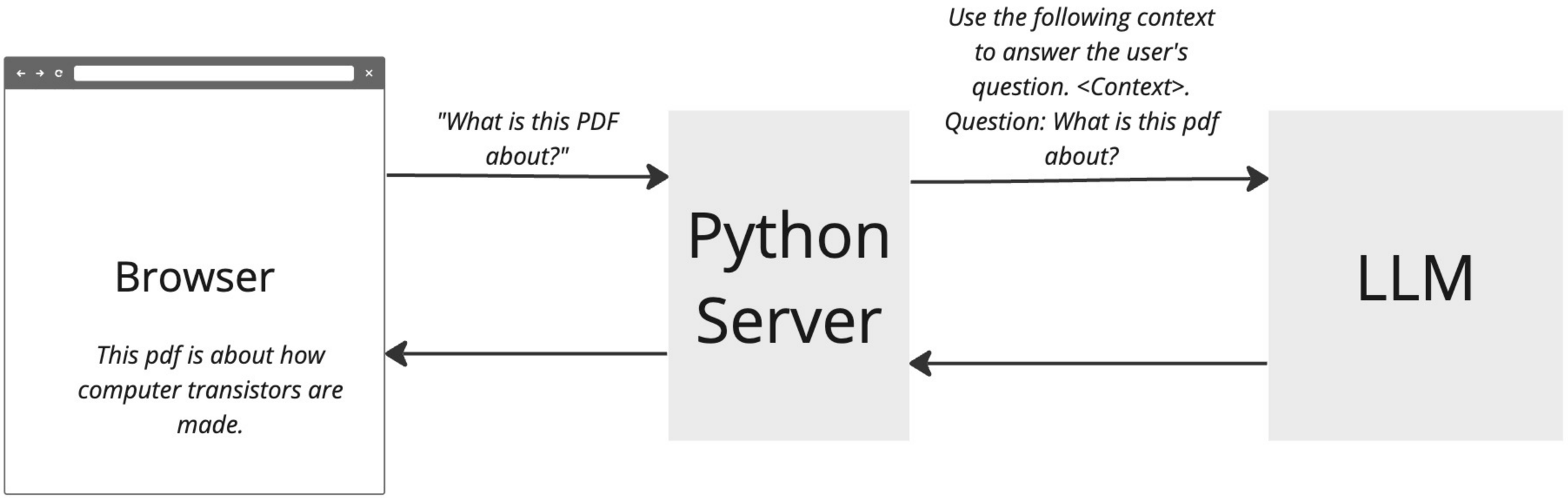
History

The thermionic triode, a vacuum tube invented in 1907, enabled amplified radio technology and long-distance telephony. The triode, however, was a fragile device that consumed a substantial amount of power. In 1909, physicist William Eccles discovered the crystal diode oscillator.[8] Physicist Julius Edgar Lilienfeld filed a patent for a field-effect transistor (FET) in Canada in 1925,[9] intended as a solid-state replacement for the triode.[10][11] He filed identical patents in the United States in 1926[12]

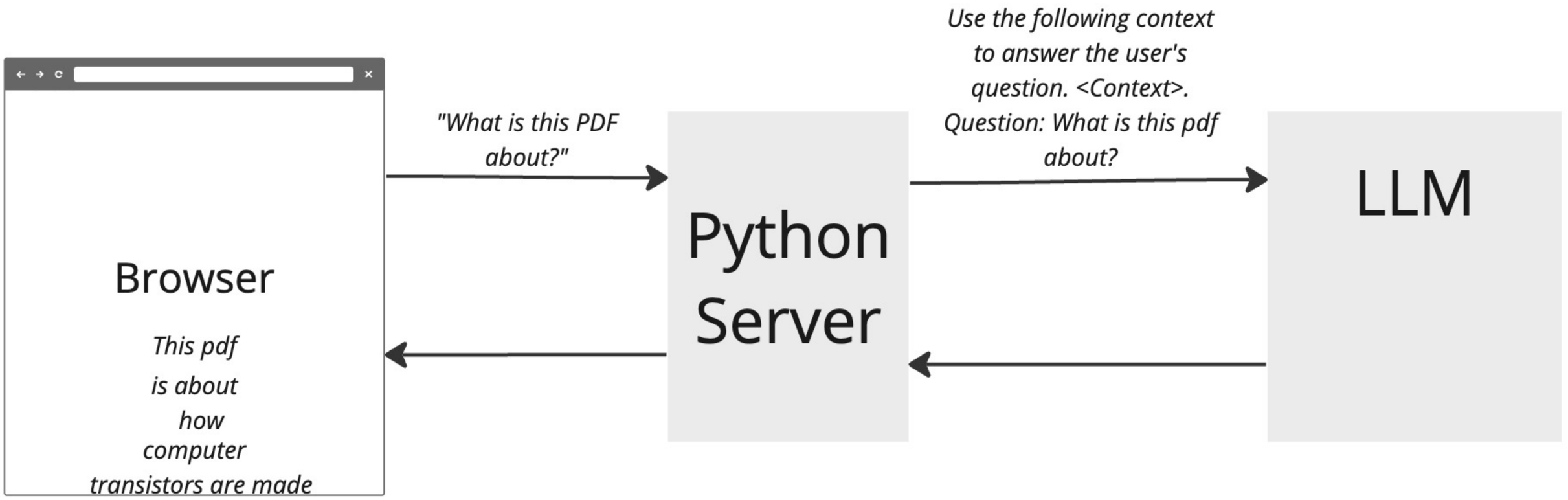
Bulk Text Generation



Bulk Text Generation



Streaming Text Generation



What combination of components parts makes for the best chain?

RetrievalChain

*retriever that fetches **2** relevant documents*

*Memory that remembers **all** previous questions about a pdf*

GPT-3.5-turbo

Retrievers

*retriever that fetches **2** relevant documents*

*retriever that fetches **3** relevant documents*

Memory

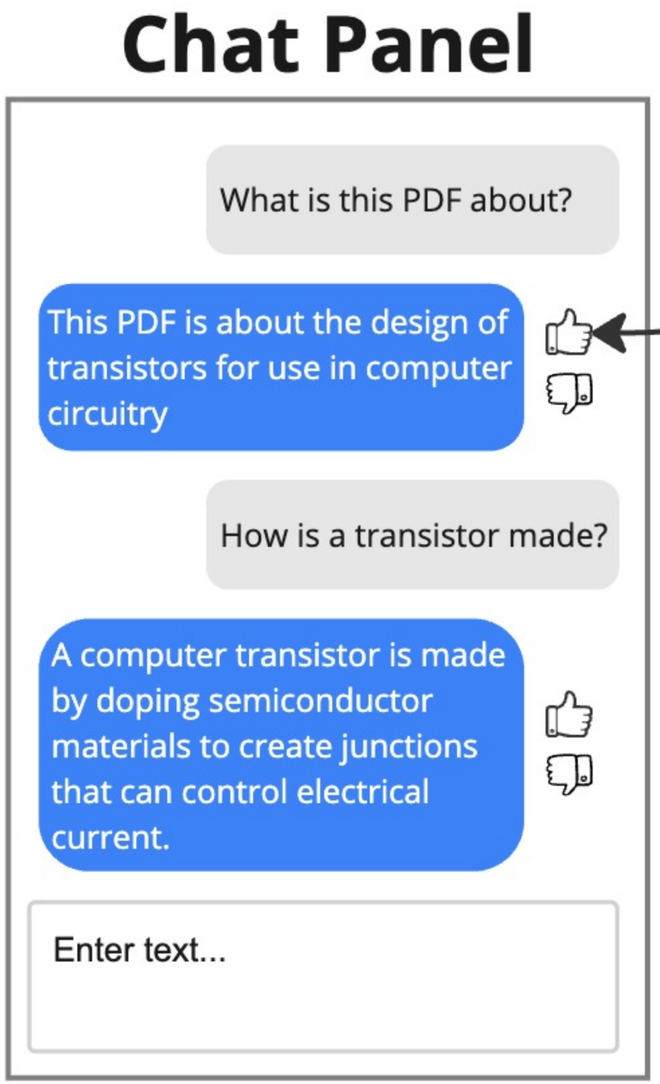
*Memory that remembers **all** previous questions about a pdf*

Memory that doesn't remember any previous conversations

LLM

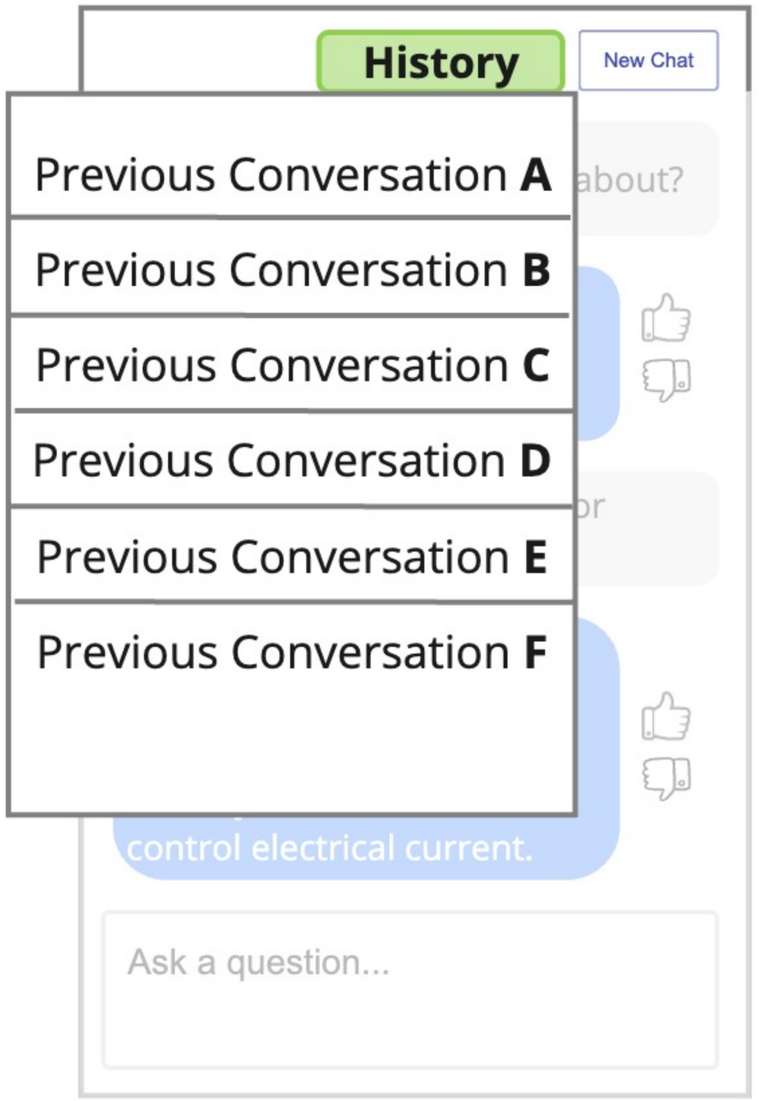
GPT-3.5-turbo

GPT-4



How do we keep track of user's "liking" or "disliking" an answer?

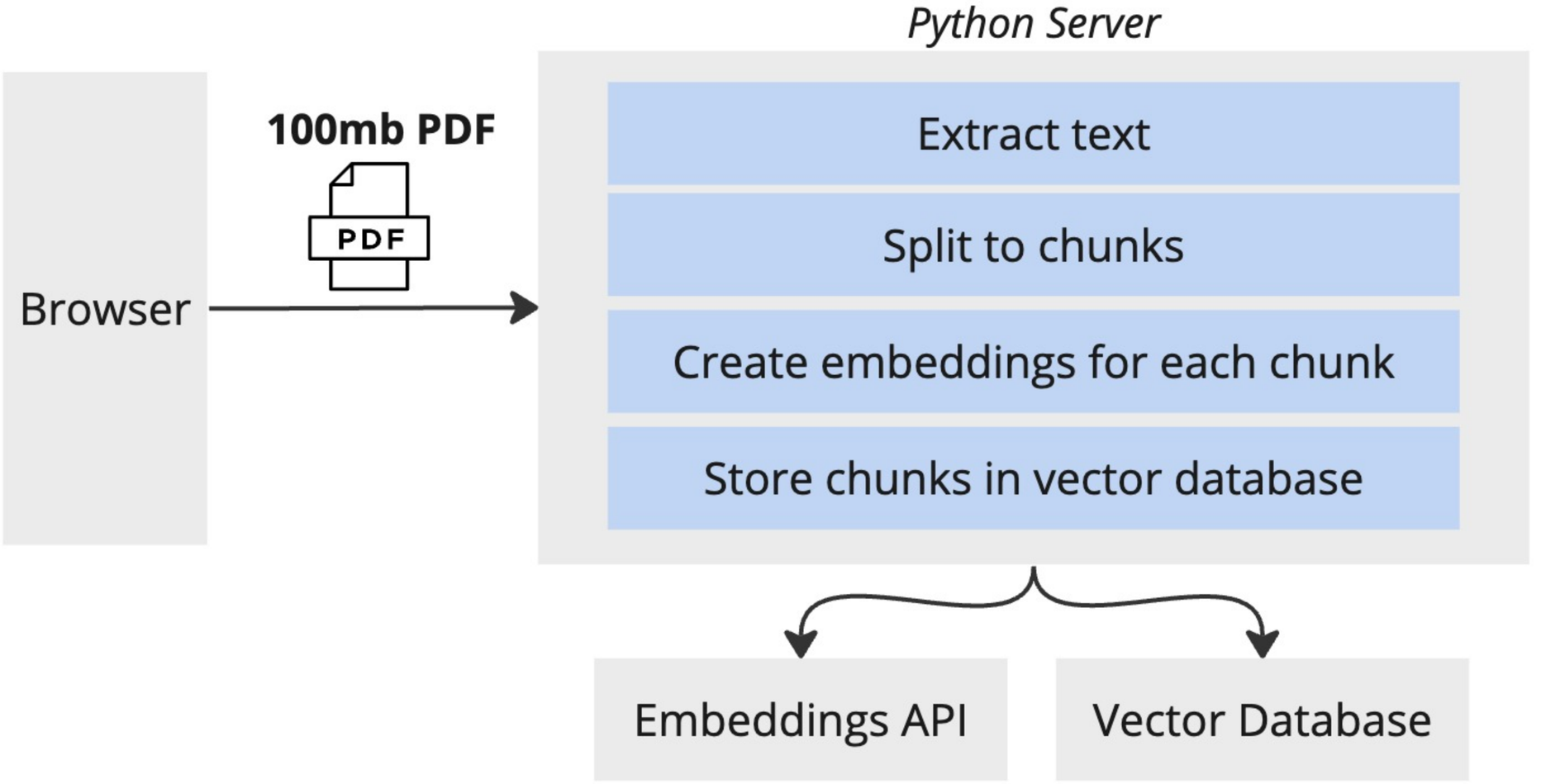
How do we keep track of what combination of parts was used to generate an answer?



How do we keep track of separate conversations?

How do we allow a user to select an old conversation and extend it with additional messages?

PDF upload + parsing can be computationally *intense*. How do we do this performantly while still responding to other incoming requests?



Our Project

Flask

Python web framework

LangChain

LLM framework

Celery

Delayed job processing

SQLAlchemy

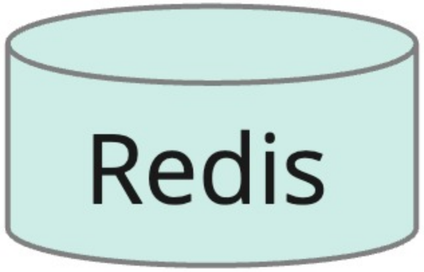
Data storage

Svelte

Frontend Javascript app (already written)

Data Storage

General data storage



Job processing

Outside Services

OpenAI

LLM Provider

File Server

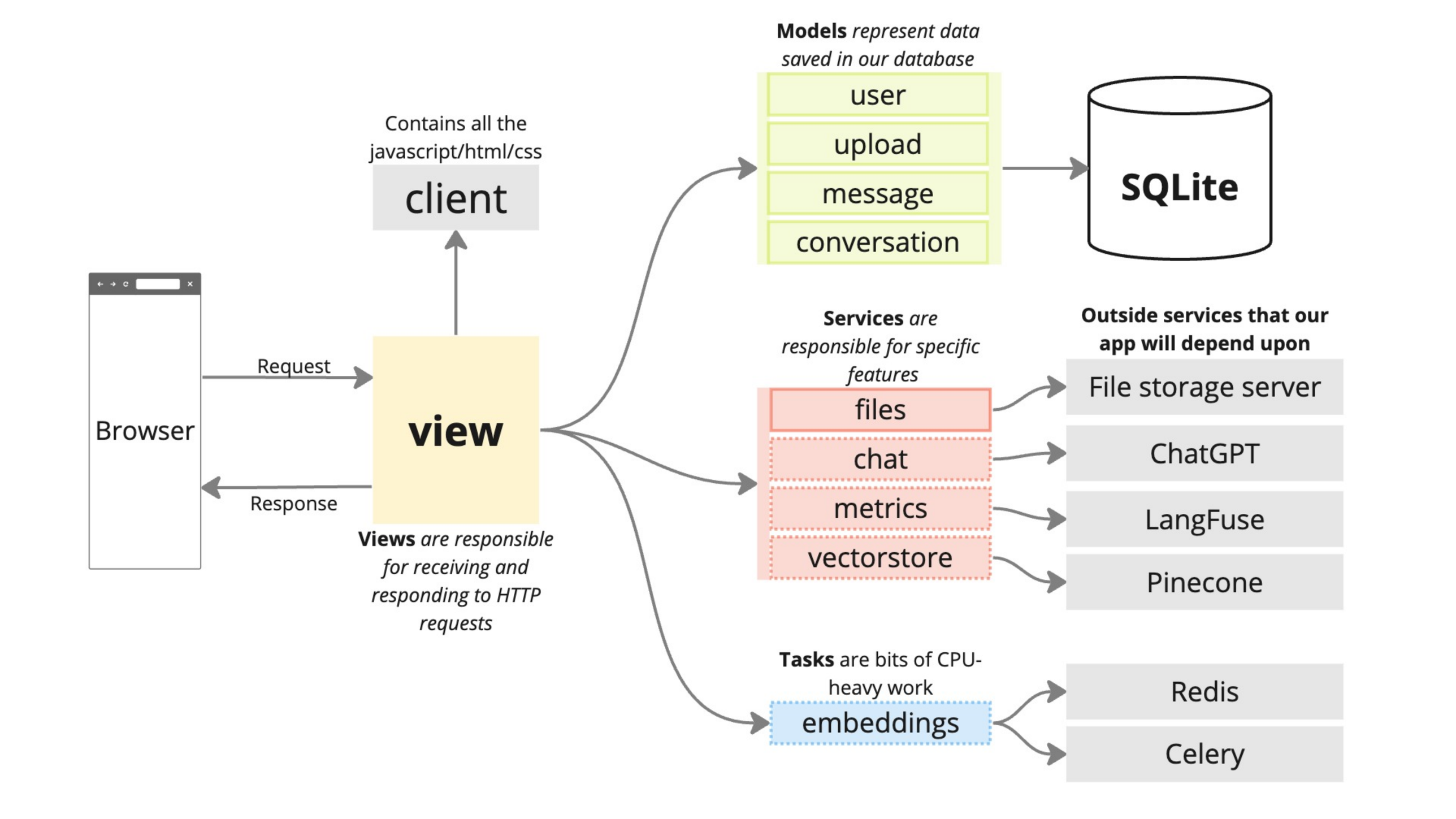
File storage server for PDF uploads (made and hosted by me)

Pinecone

Embedding storage + lookup

LangFuse

Analytics + tracing provider


Models represent data saved in our database
user
upload
message
conversation
SQLite
Contains all the javascript/html/css
client
Browser
Request
Response
view
Views are responsible for receiving and responding to HTTP requests
Services are responsible for specific features
files
chat
metrics
vectorstore
Outside services that our app will depend upon
File storage server
ChatGPT
LangFuse
Pinecone
Tasks are bits of CPU-heavy work
embeddings
Redis
Celery