

# Retrieval Augmented Generation.

## RAG Application.



by **Sriram Prakash**

# Project Overview

## Goal

Build a Streamlit app that lets a user upload any PDF and chat with it

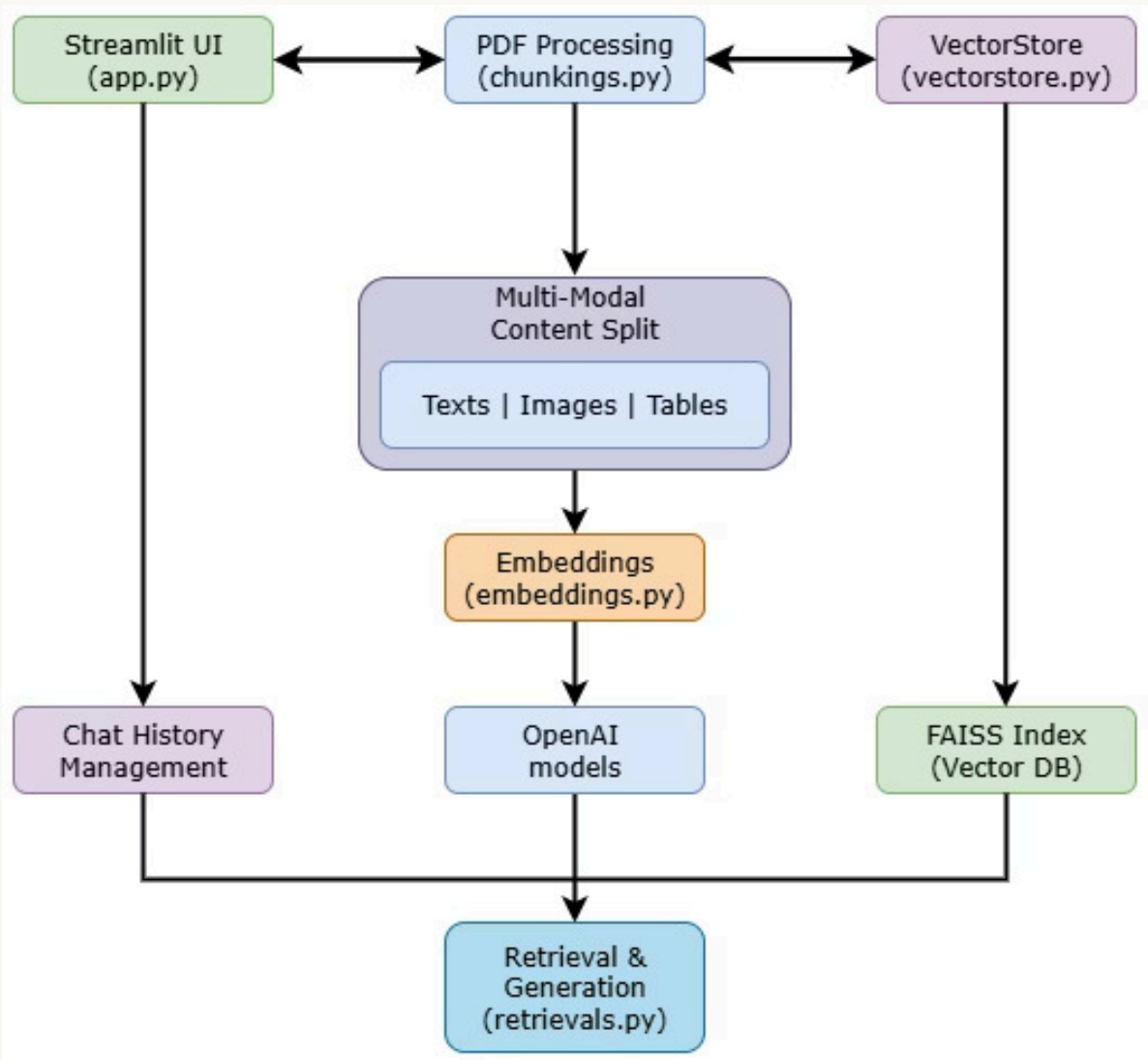
## Contents

System Architecture

Prompt Templates

Chats (Q&A) on app

# 1. System Architecture



## Streamlit UI (app.py)

Provides an intuitive web interface for PDF upload and Q&A interactions.

Features:

- Dark Theme UI: Modern, professional interface
- File Management: PDF upload with duplicate detection
- Chat Interface: Persistent chat history with follow-up suggestions
- Knowledge Base Management: Clear data, system diagnostics, file removal
- Citation Display: Interactive search results and source citations

## PDF Processing (chunkings.py)

Extracts text with line-level tracking Identifies and extracts images with position data.

Detects and extracts tables using fitz (PyMuPDF) table finder Maintains detailed metadata for citations

## Embedding Generation (embeddings.py)

Text embeddings using OpenAI's text-embedding-3-small  
Image descriptions using GPT-04-mini, Table embeddings with enhanced context Unified embedding approach for all content types

## Vector Store (vectorstore.py)

Separate FAISS indexes for text, images, and tables Efficient similarity search with metadata tracking Persistent storage and loading capabilities Statistics and monitoring features

## Hybrid Retrieval (retrievals.py)

Combines dense embeddings and keyword search Configurable weighting between search methods Detailed citation formatting Context-aware question answering

# 2. Prompt Library

These are the prompts used in cursor and the respective chat logs

You are an expert AI engineer tasked with building a production-ready, multi-modal Retrieval-Augmented Generation (RAG) Streamlit app that meets the following requirements:

Goal:

Build a Streamlit application where the user uploads any PDF (containing text, images, tables, or mixed layouts) and can chat with it. The system must extract, chunk, embed, and retrieve relevant content—including images and tables—with accurate citations (page/snippet). The chat must support follow-up questions using chat history.

Core Features:

1. Streamlit UI:

- PDF upload (single file for now)
- Chat interface with persistent history
- Each answer must show its citations (page number or snippet)
- Display retrieved image/table thumbnails alongside answers if relevant

2. Ingestion & Chunking Strategies (must be production-level and explicitly implemented):

- Use a robust PDF parser like MyPuPDF and others to extract:
  - \* Text with section/heading hierarchy
  - \* Images with captions using OpenAI vision models to extract summaries.
  - \* Tables as CSV/Markdown plus LLM-generated textual summaries
- Store each element (text, table, image) with metadata: source file, page number, section title, content type
- Semantic chunking strategy:
  - \* For text: chunk by semantic units (headings/paragraphs) with 50–100 token overlap, max chunk size 500–800 tokens
  - \* Keep tables and images as independent chunks linked to their captions
  - \* Store both a raw form and a summarized form for retrieval flexibility
- Ensure multimodal readiness:
  - \* Text embeddings → OpenAI text-embedding-3-large or BGE-large
  - \* Image embeddings → OpenAI CLIP or equivalent
  - \* Tables → embed both CSV text and its LLM-generated summary

3. Embeddings & Vector Storage:

- Use a vector database like FAISS
- Store separate indexes for text, images, and tables with unified metadata schema
- Allow hybrid retrieval (dense embeddings + keyword search) for better recall

4. Retrieval:

- On query, generate embeddings for text and optionally images
- Retrieve top-K chunks across all modalities
- Re-rank with a cross-encoder or RankGPT to ensure highest relevance
- Filter by metadata if needed (page range, content type)

5. Q&A Prompting:

- Pass retrieved chunks + chat history to the LLM
- Instruct: "Answer ONLY from the provided context. Include citations in (Page X, Snippet) format. If uncertain, say you don't know."
- If retrieved chunks contain images/tables, provide short descriptions and references to display in UI

6. Follow-up Support:

- Maintain chat history in session state and include in retrieval context
- Use a condensing step: summarize previous exchanges into a standalone query for better retrieval

7. Implementation:

- Modularize into:
  - processing folder
    - \*chunkings.py → Extracts text, images, tables from PDF
    - \*embeddings.py → Converts content to AI embeddings
    - \*vectorstore.py → Stores vectors in FAISS indexes
    - \*retrievals.py → Searches and generates answers
  - app.py → Orchestrates everything and provides UI
  - requirements.txt with all dependencies
  - README.md documenting model choices, chunking strategy, and limitations

Deliverables:

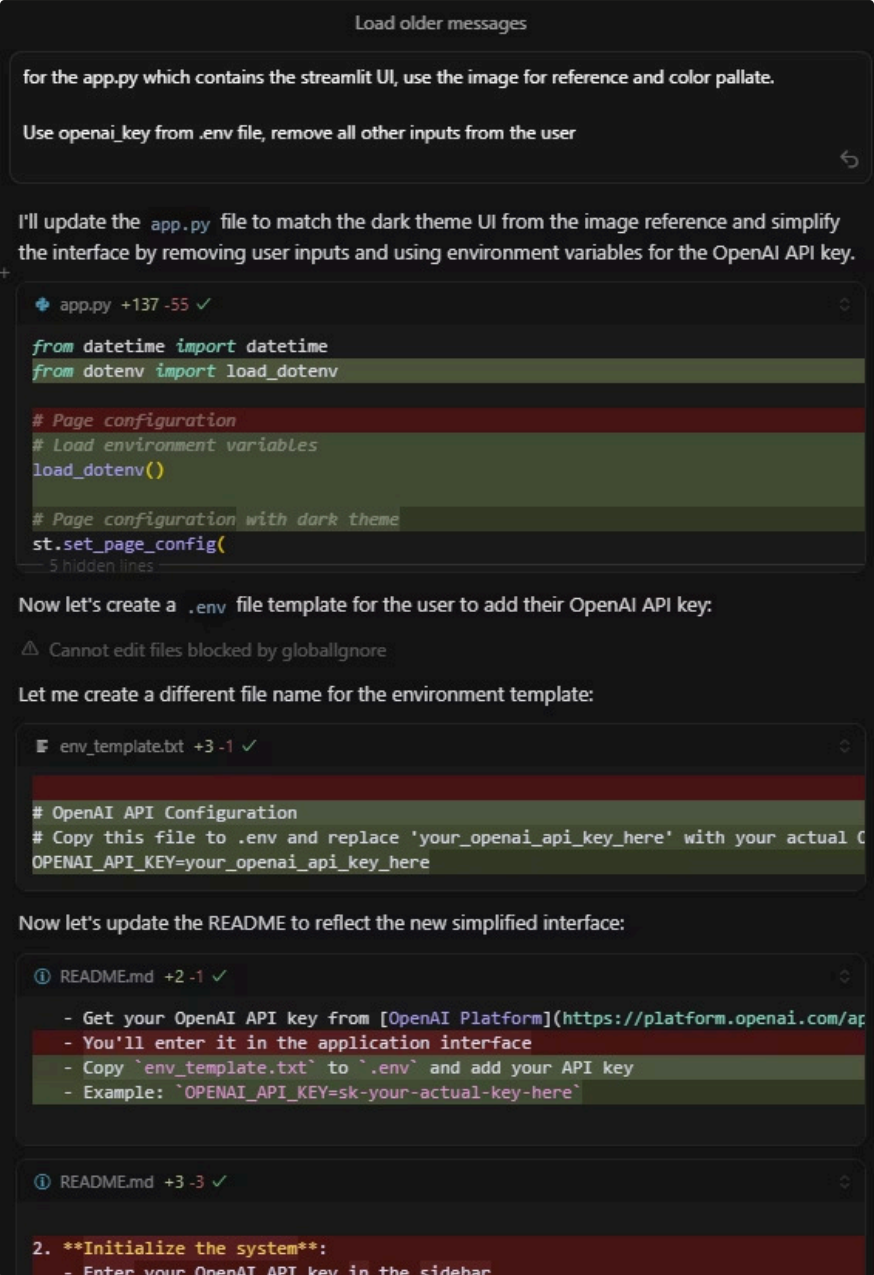
- Fully working app
- Citations for every answer
- Support for text, images, and tables in retrieval and display
- Production-level ingestion and chunking strategies implemented as described
- Code structured into clear modules

The above is the requirement, start with chunking and data ingestion only, once done with review we can go to the next one in a step by step manner

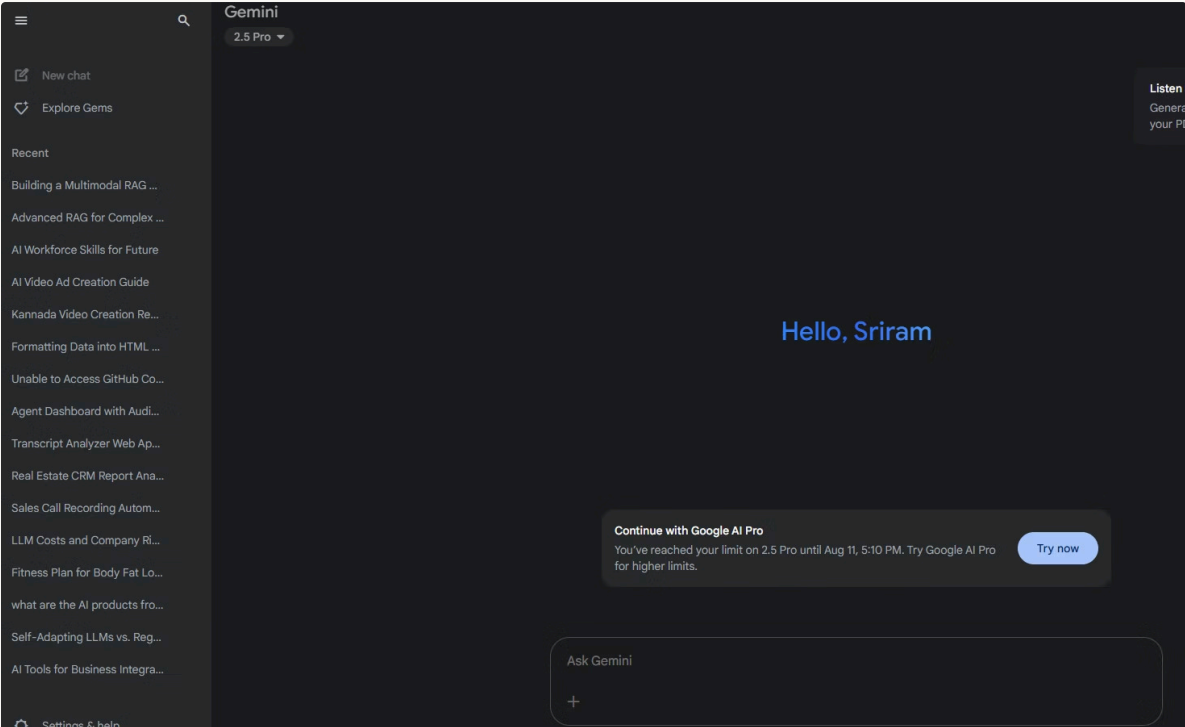


# 2. Chats (Cursor) Library

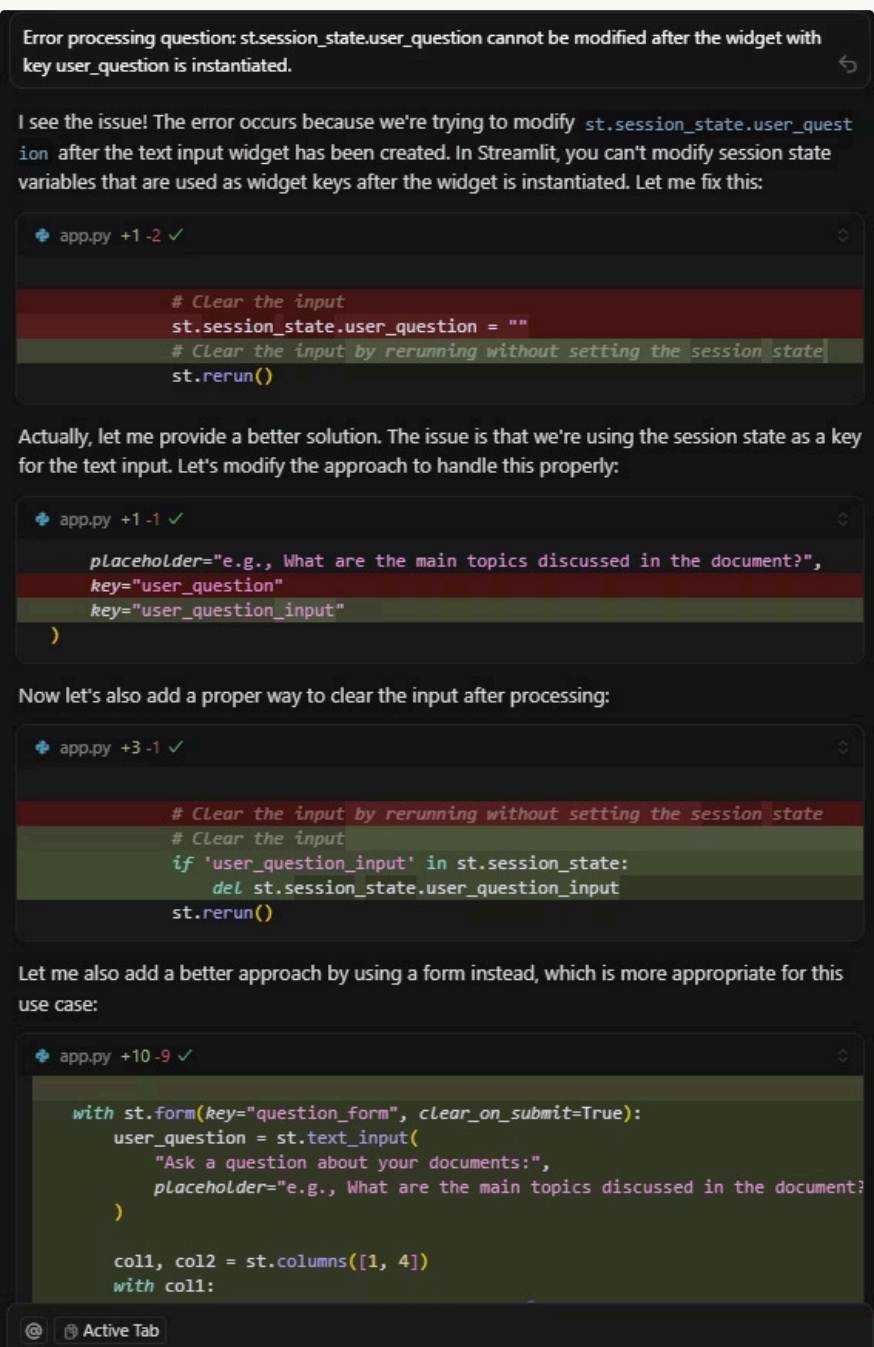
## Prompt for Top 3 Collateral Recommendations



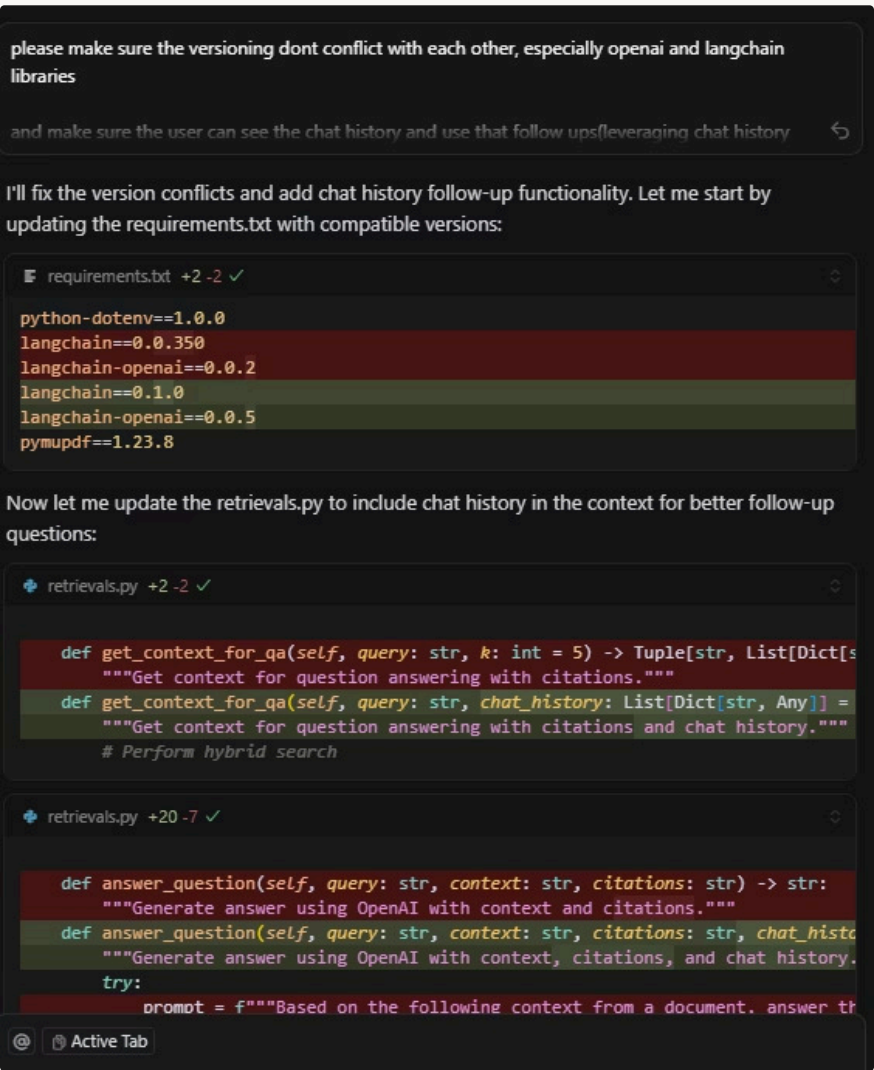
1)To Update the UI having a similar color palate to google Gemini.



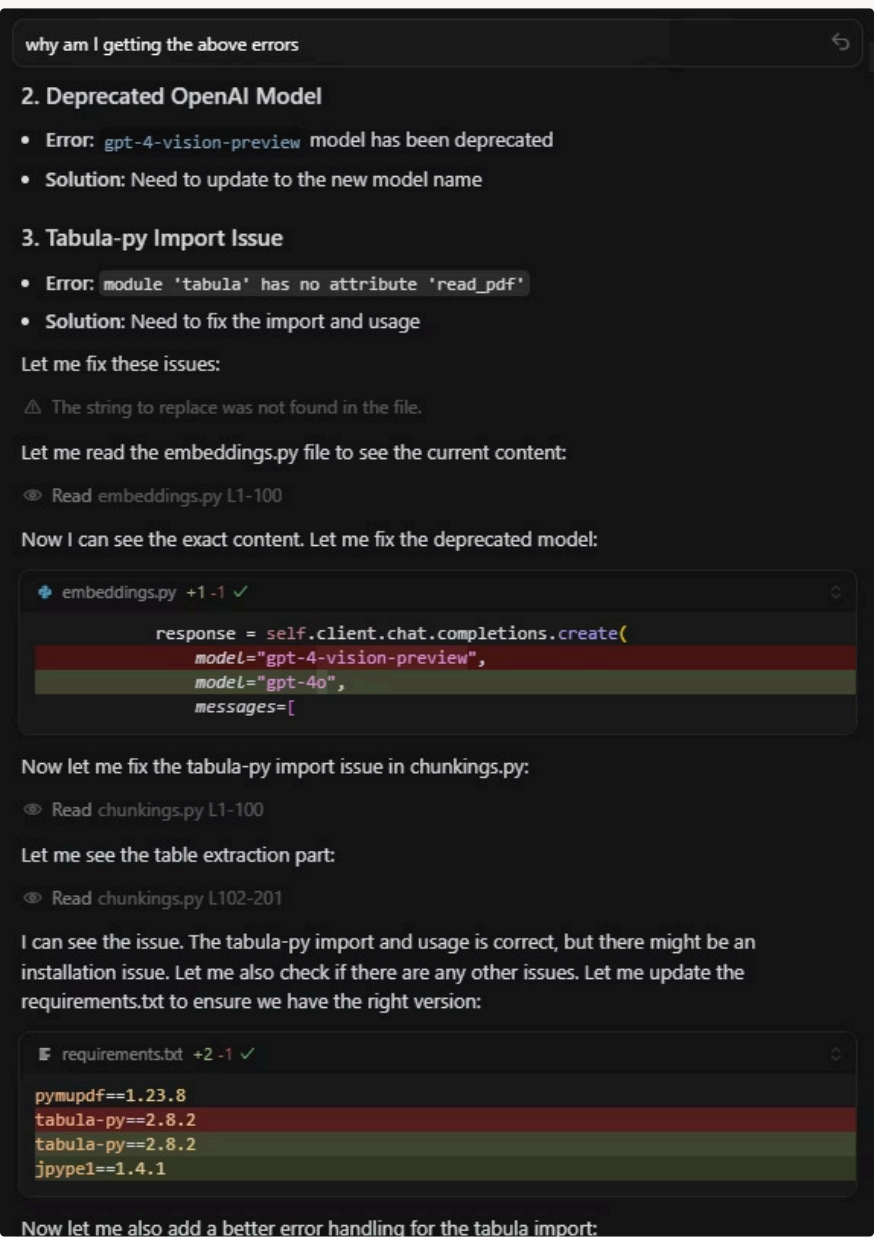
2)removed options where users input their OpenAI api key.



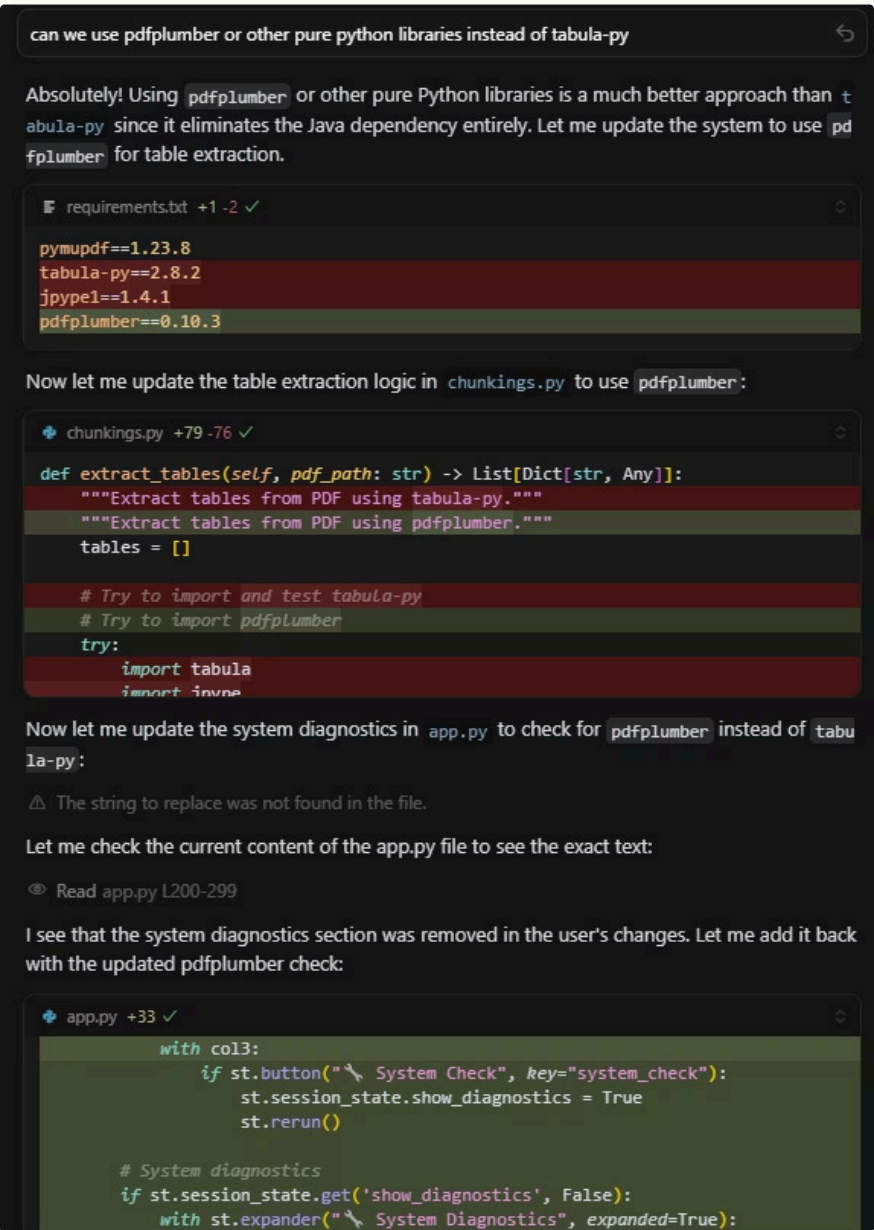
3) Error Fixes where the logs have been attached.



4) Version conflicts of libraries.



5) Error fixes, where the logs have been attached.




6) Resolving tabula-py error, and finding other alternatives.



### 3. Chats – RAG Application

Question 1

 You:


whats the total products sales in 2023

Assistant:

The total product sales in 2023 amounted to \$298,085 million. This figure is derived from the total net sales reported for the twelve months ended September 30, 2023, specifically for products.  
Citations: [1] Page 1, Table 0; [2] Page 1, Lines 204-269.

	Three Months Ended		Twelve Months Ended	
	September 28, 2024	September 30, 2023	September 28, 2024	September 30, 2023
Net sales:				
Products	\$ 69,958	\$ 67,184	\$ 294,866	\$ 298,085
Services	24,972	22,314	96,169	85,200
Total net sales <sup>(1)</sup>	94,930	89,498	391,035	383,285

Question 2

 You:

whats the total sales of ipad in 2023


Assistant:

The total sales of iPad in 2023 amounted to \$28,300 million. This figure is derived from the net sales reported for the twelve months ended September 30, 2023, where iPad sales were specifically listed.  
Citations: [1] Table: Products; [3] Net sales by category.

<sup>(1)</sup> Net sales by category:

iPhone	\$ 46,222	\$ 43,805	\$ 201,183	\$ 200,583
Mac	7,744	7,614	29,984	29,357
iPad	6,950	6,443	26,694	28,300
Wearables, Home and Accessories	9,042	9,322	37,005	39,845
Services	24,972	22,314	96,169	85,200
Total net sales	<u>\$ 94,930</u>	<u>\$ 89,498</u>	<u>\$ 391,035</u>	<u>\$ 383,285</u>

Question 3

 You:

What is the difference in iPad sales between 2024 and 2023?


Assistant:

The total sales of iPad in 2024 amounted to \$26,694 million, while in 2023, it was \$28,300 million.  
To find the difference in iPad sales between 2024 and 2023, we calculate:  
  
26,694*million*(2024) − 28,300 million (2023) = -\$1,606 million.  
  
Thus, the difference in iPad sales between 2024 and 2023 is a decrease of \$1,606 million.  
  
Citations: [1], [2], [3].

<sup>(1)</sup> Net sales by category:

iPhone	\$ 46,222	\$ 43,805	\$ 201,183	\$ 200,583
Mac	7,744	7,614	29,984	29,357
iPad	6,950	6,443	26,694	28,300
Wearables, Home and Accessories	9,042	9,322	37,005	39,845
Services	24,972	22,314	96,169	85,200
Total net sales	<u>\$ 94,930</u>	<u>\$ 89,498</u>	<u>\$ 391,035</u>	<u>\$ 383,285</u>

Question 4

 You:


whats the tax paid in 2023

Assistant:

The tax paid in 2023 was \$18,679 million. This figure is derived from the supplemental cash flow disclosure, which states "Cash paid for income taxes, net" for the year ended September 30, 2023.  
Citations: [4].

Other	(301)	(381)
Cash used in financing activities	(121,983)	(108,488)
Increase/(Decrease) in cash, cash equivalents, and restricted cash and cash equivalents	(794)	5,760
Cash, cash equivalents, and restricted cash and cash equivalents, ending balances	<u>\$ 29,943</u>	<u>\$ 30,737</u>
Supplemental cash flow disclosure:		
Cash paid for income taxes, net	\$ 26,102	\$ 18,679

Question 5

 You:

How much of sales was generated from America compared to Asia, how big is the market in Asia

Assistant:

From the provided context, the sales generated from America (Americas) amounted to \$41,664 million, while the sales from Asia (which includes Greater China, Japan, and the Rest of Asia Pacific) totaled \$28,342 million (15,033 million from Greater China + 5,926 million from Japan + 7,383 million from Rest of Asia Pacific).  
To compare:

- Sales from America: \$41,664 million
- Sales from Asia: \$28,342 million

In terms of the market size in Asia:


- Greater China: \$15,033 million
- Japan: \$5,926 million
- Rest of Asia Pacific: \$7,383 million
- Total market size in Asia: \$28,342 million

Thus, the sales from America were higher than those from Asia, with America generating 41,664*million compared to Asia's* 28,342 million.  
  
Citations: [4].

<sup>(1)</sup> Net sales by reportable segment:

Americas	\$ 41,664	\$ 40,115	\$ 167,045	\$ 162,560
Europe	24,924	22,463	101,328	94,294
Greater China	15,033	15,084	66,952	72,559
Japan	5,926	5,505	25,052	24,257
Rest of Asia Pacific	7,383	6,331	30,658	29,615
Total net sales	<u>\$ 94,930</u>	<u>\$ 89,498</u>	<u>\$ 391,035</u>	<u>\$ 383,285</u>

Question 6

 You:

whats the total debt in 2023 and 2024, compare it with the net income in both these years

Assistant:

The provided context does not contain specific information regarding the total debt for the years 2023 and 2024, nor does it provide the net income for these years. Therefore, I cannot compare the total debt with the net income for 2023 and 2024 based on the available information. If you have access to additional data or specific figures regarding total debt and net income, please provide that, and I would be happy to help analyze it.

	Twelve Months Ended	
	September 28, 2024	September 30, 2023
Cash, cash equivalents, and restricted cash and cash equivalents, beginning balances	\$ 30,737	\$ 24,977
Operating activities:		
Net income	93,736	96,995