Building a RAG System for PDF Analysis: Methodology and Reasoning

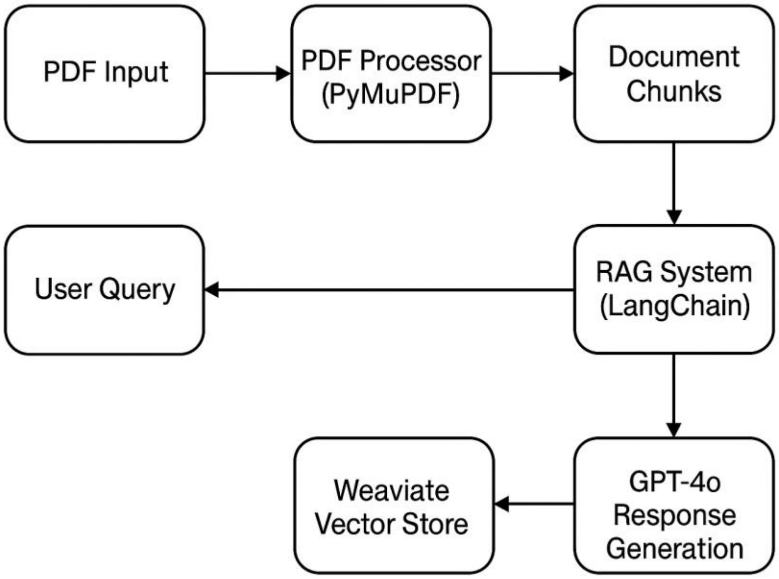
The Problem

I needed a system that could intelligently answer questions about complex PDF documents containing not just text, but also tables and figures. The challenge wasn't just extracting content - it was making sure the system could reason across different content types and provide accurate citations.

Architecture Overview

I kept the architecture straightforward:

```
PDF → Content Extraction → Vector Store (Weaviate) → RAG Pipeline → Response
```



Architecture

The key insight was treating each content type (text, tables, figures) as first-class citizens rather than trying to force everything into a single text representation.

Content Extraction

I tried several approaches before settling on PyMuPDF4LLM for text extraction because it preserves document structure maintaining semantic boundaries while ensuring optimal chunk sizes for embedding generation which is better than alternatives like PyPDF2.

For tables, I initially tried to convert everything to text descriptions, but that lost too much information. Instead, I extract tables into pandas DataFrames and serialize them as markdown for consistency- this preserves structure while keeping them LLM-friendly.

For figures, I extracted them with PyMuPDF, filtered out tiny decorative elements (anything under 100x100 pixels), and stored them as base64 strings in the metadata. This lets me do multimodal analysis when needed without complicating the basic text pipeline.

The Chunking Strategy

- Chunk by semantic boundaries (paragraphs for text, whole tables, individual figures)
- Keep chunks around 512 characters with 100-character overlap (Didn't use large for free-tier embedding model)
- Store rich metadata: page numbers, content type, bounding boxes. This content-type specific filtering and processing and preserves spatial relationships within documents

Each chunk looks like this:

```
{
    content: str                      # Actual content
    content_type: str                # text/table/figure
    page_number: int                 # Source page
    chunk_index: int                 # Position within page
    document_name: str               # Source document
    bbox: Optional[Tuple]            # Spatial coordinates
    metadata: Optional[Dict]         # Additional context
}
```

This metadata proved crucial for accurate citations later.

Vector Store Choice

I went with Weaviate after evaluating several options. The deciding factors: - Hybrid search (vector + metadata filtering) - Managed cloud offering (I didn’t want to manage infrastructure) - Good embedding integration - Flexible schema for my metadata needs

The schema is pretty simple - just the chunk content plus metadata fields. I serialize complex data like bounding boxes as JSON strings, which works fine for filtering.

The RAG Pipeline

I use LangChain for orchestration, mainly because it handles the plumbing well and integrates nicely with LangSmith for observability. Also, it's easy and straightforward and is a good community.

The retrieval process: 1. Embed the user query 2. Vector search for top 10 similar chunks 3. Apply any metadata filters (content type, document name) 4. Format chunks with clear markers: `[TEXT from page 5]`, `[TABLE from page 7]`, etc.

For the generation side, GPT-4o is used with a carefully crafted system prompt that emphasizes accuracy and proper citations. Temperature set to 0.1 to keep responses consistent. For multimodal, send the decoded base64 images to GPT4-o with a simple prompt asking for description. Only analyze images when they're actually retrieved, not for every query.

What I Learned

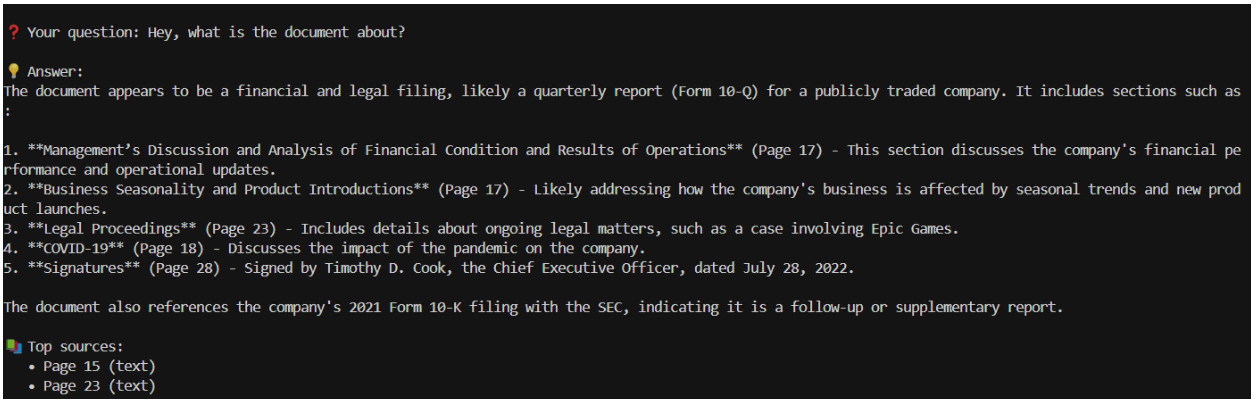
- **Content type markers** Initially, I tried to let the LLM figure out what type of content it was looking at. Adding explicit `[TEXT]`, `[TABLE]`, `[FIGURE]` markers dramatically improved response quality.
- **Metadata** Rich metadata enables so many useful features - filtering by content type, accurate page citations, spatial reasoning about document layout.
- **Keep it simple initially.** I started with just text, got that working well, then added tables and then figures.

Current Limitations

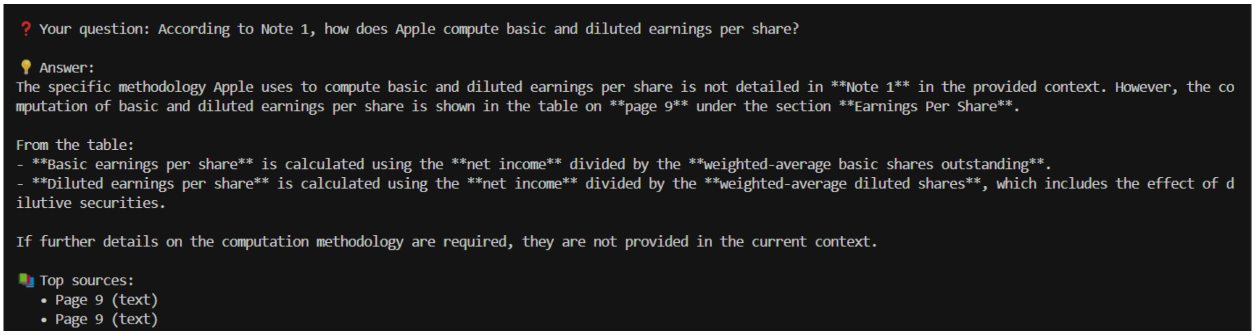
- Complex tables sometimes lose structure in markdown conversion
- Image analysis is relatively basic (could use more specialized vision models)
- No batch query processing
- Doesn't have a memory

Some examples

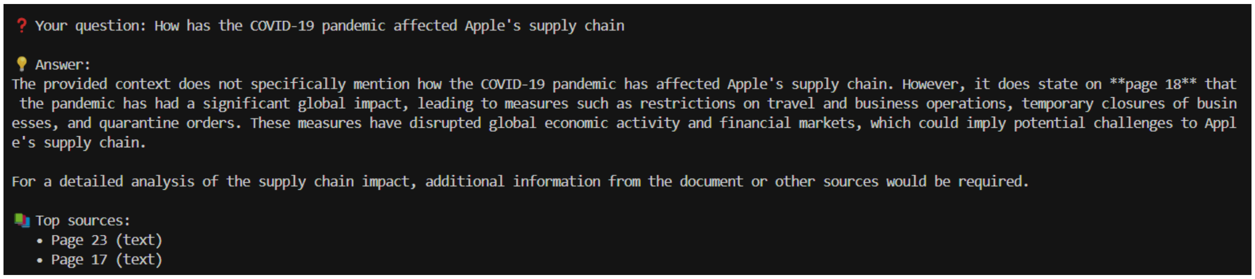
Basic information questions



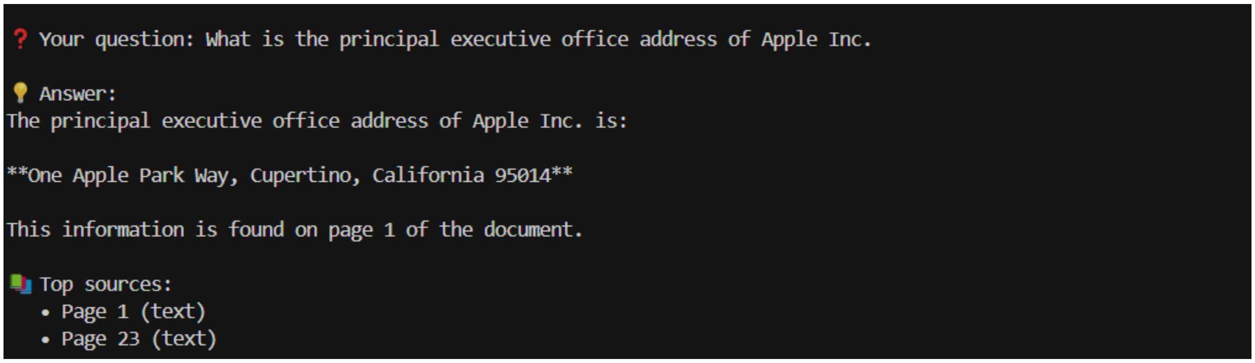
```
❓ Your question: Hey, what is the document about?

💡 Answer:
The document appears to be a financial and legal filing, likely a quarterly report (Form 10-Q) for a publicly traded company. It includes sections such as
:

1. **Management's Discussion and Analysis of Financial Condition and Results of Operations** (Page 17) - This section discusses the company's financial pe
rformance and operational updates.
2. **Business Seasonality and Product Introductions** (Page 17) - Likely addressing how the company's business is affected by seasonal trends and new prod
uct launches.
3. **Legal Proceedings** (Page 23) - Includes details about ongoing legal matters, such as a case involving Epic Games.
4. **COVID-19** (Page 18) - Discusses the impact of the pandemic on the company.
5. **Signatures** (Page 28) - Signed by Timothy D. Cook, the Chief Executive Officer, dated July 28, 2022.

The document also references the company's 2021 Form 10-K filing with the SEC, indicating it is a follow-up or supplementary report.

📚 Top sources:
  • Page 15 (text)
  • Page 23 (text)
```



```
❓ Your question: According to Note 1, how does Apple compute basic and diluted earnings per share?

💡 Answer:
The specific methodology Apple uses to compute basic and diluted earnings per share is not detailed in **Note 1** in the provided context. However, the co
mputation of basic and diluted earnings per share is shown in the table on **page 9** under the section **Earnings Per Share**.

From the table:
- **Basic earnings per share** is calculated using the **net income** divided by the **weighted-average basic shares outstanding**.
- **Diluted earnings per share** is calculated using the **net income** divided by the **weighted-average diluted shares**, which includes the effect of d
ilutive securities.

If further details on the computation methodology are required, they are not provided in the current context.

📚 Top sources:
  • Page 9 (text)
  • Page 9 (text)
```

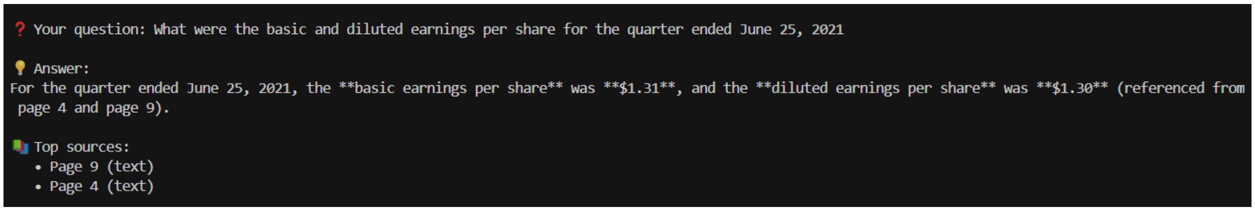


```
❓ Your question: How has the COVID-19 pandemic affected Apple's supply chain

💡 Answer:
The provided context does not specifically mention how the COVID-19 pandemic has affected Apple's supply chain. However, it does state on **page 18** that
 the pandemic has had a significant global impact, leading to measures such as restrictions on travel and business operations, temporary closures of busin
esses, and quarantine orders. These measures have disrupted global economic activity and financial markets, which could imply potential challenges to Appl
e's supply chain.

For a detailed analysis of the supply chain impact, additional information from the document or other sources would be required.

📚 Top sources:
  • Page 23 (text)
  • Page 17 (text)
```

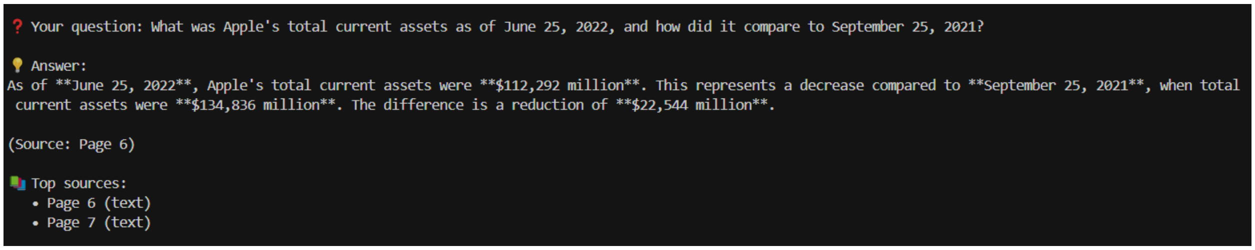


```
❓ Your question: What is the principal executive office address of Apple Inc.

💡 Answer:
The principal executive office address of Apple Inc. is:

**One Apple Park Way, Cupertino, California 95014**

This information is found on page 1 of the document.

📚 Top sources:
  • Page 1 (text)
  • Page 23 (text)
```

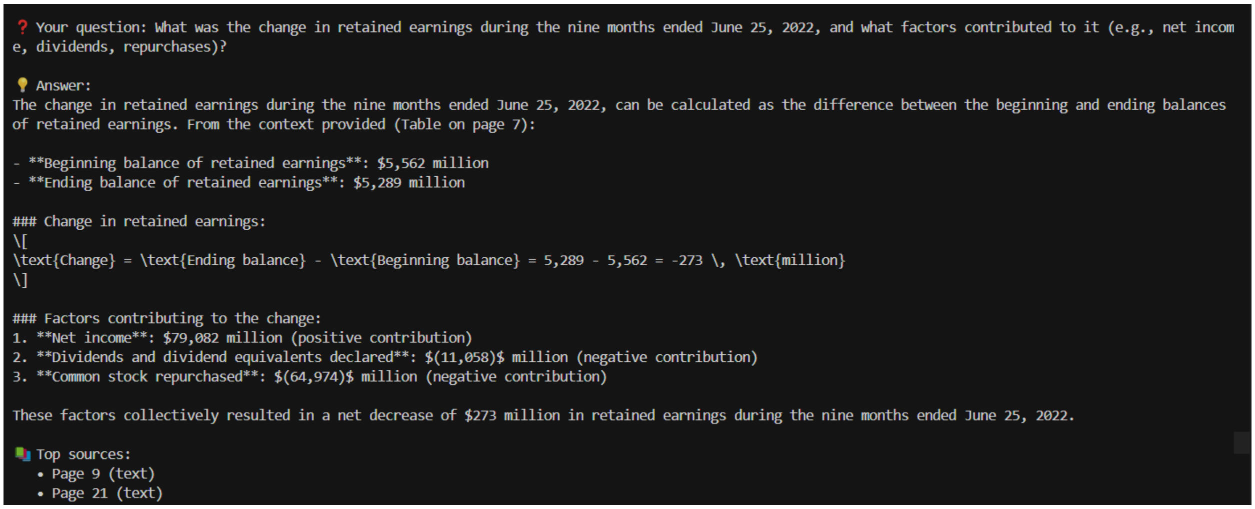
Finding stats



```
❓ Your question: What were the basic and diluted earnings per share for the quarter ended June 25, 2021

💡 Answer:
For the quarter ended June 25, 2021, the **basic earnings per share** was **$1.31**, and the **diluted earnings per share** was **$1.30** (referenced from
 page 4 and page 9).

📚 Top sources:
  • Page 9 (text)
  • Page 4 (text)
```

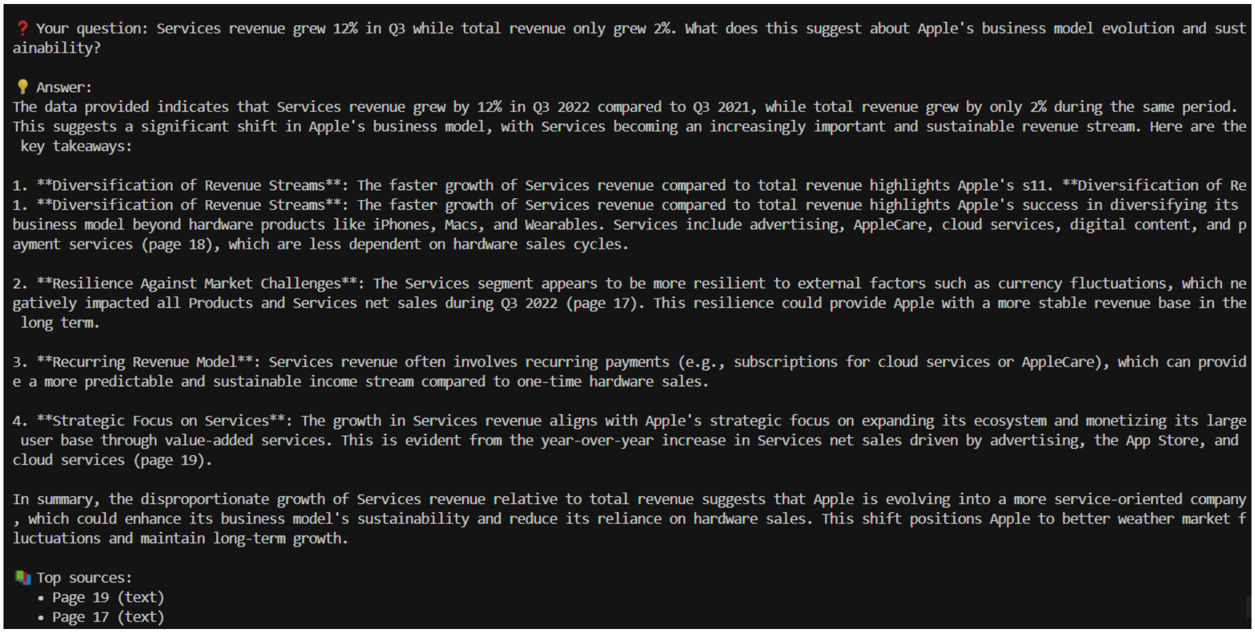


```
❓ Your question: What was Apple's total current assets as of June 25, 2022, and how did it compare to September 25, 2021?

💡 Answer:
As of **June 25, 2022**, Apple's total current assets were **$112,292 million**. This represents a decrease compared to **September 25, 2021**, when total
 current assets were **$134,836 million**. The difference is a reduction of **$22,544 million**.

(Source: Page 6)

📚 Top sources:
   • Page 6 (text)
   • Page 7 (text)
```



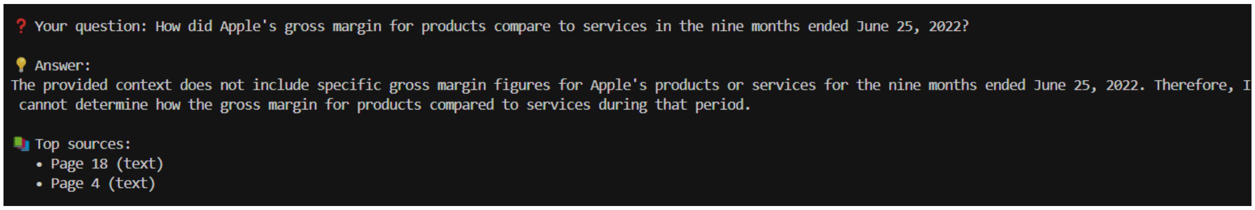
```
❓ Your question: What was the change in retained earnings during the nine months ended June 25, 2022, and what factors contributed to it (e.g., net incom
e, dividends, repurchases)?

💡 Answer:
The change in retained earnings during the nine months ended June 25, 2022, can be calculated as the difference between the beginning and ending balances
of retained earnings. From the context provided (Table on page 7):

- **Beginning balance of retained earnings**: $5,562 million
- **Ending balance of retained earnings**: $5,289 million

### Change in retained earnings:
\[
\text{Change} = \text{Ending balance} - \text{Beginning balance} = 5,289 - 5,562 = -273 \, \text{million}
\]

### Factors contributing to the change:
1. **Net income**: $79,082 million (positive contribution)
2. **Dividends and dividend equivalents declared**: $(11,058)$ million (negative contribution)
3. **Common stock repurchased**: $(64,974)$ million (negative contribution)

These factors collectively resulted in a net decrease of $273 million in retained earnings during the nine months ended June 25, 2022.

📚 Top sources:
   • Page 9 (text)
   • Page 21 (text)
```

Analysis



```
❓ Your question: Services revenue grew 12% in Q3 while total revenue only grew 2%. What does this suggest about Apple's business model evolution and sust
ainability?

💡 Answer:
The data provided indicates that Services revenue grew by 12% in Q3 2022 compared to Q3 2021, while total revenue grew by only 2% during the same period.
This suggests a significant shift in Apple's business model, with Services becoming an increasingly important and sustainable revenue stream. Here are the
 key takeaways:

1. **Diversification of Revenue Streams**: The faster growth of Services revenue compared to total revenue highlights Apple's s11. **Diversification of Re
1. **Diversification of Revenue Streams**: The faster growth of Services revenue compared to total revenue highlights Apple's success in diversifying its
business model beyond hardware products like iPhones, Macs, and Wearables. Services include advertising, AppleCare, cloud services, digital content, and p
ayment services (page 18), which are less dependent on hardware sales cycles.

2. **Resilience Against Market Challenges**: The Services segment appears to be more resilient to external factors such as currency fluctuations, which ne
gatively impacted all Products and Services net sales during Q3 2022 (page 17). This resilience could provide Apple with a more stable revenue base in the
 long term.

3. **Recurring Revenue Model**: Services revenue often involves recurring payments (e.g., subscriptions for cloud services or AppleCare), which can provid
e a more predictable and sustainable income stream compared to one-time hardware sales.

4. **Strategic Focus on Services**: The growth in Services revenue aligns with Apple's strategic focus on expanding its ecosystem and monetizing its large
 user base through value-added services. This is evident from the year-over-year increase in Services net sales driven by advertising, the App Store, and
cloud services (page 19).

In summary, the disproportionate growth of Services revenue relative to total revenue suggests that Apple is evolving into a more service-oriented company
, which could enhance its business model's sustainability and reduce its reliance on hardware sales. This shift positions Apple to better weather market f
luctuations and maintain long-term growth.

📚 Top sources:
   • Page 19 (text)
   • Page 17 (text)
```

No info



```
❓ Your question: How did Apple's gross margin for products compare to services in the nine months ended June 25, 2022?

💡 Answer:
The provided context does not include specific gross margin figures for Apple's products or services for the nine months ended June 25, 2022. Therefore, I
 cannot determine how the gross margin for products compared to services during that period.

📚 Top sources:
   • Page 18 (text)
   • Page 4 (text)
```

THANK YOU