RAG-Powered Information Extraction: Tables & Text

Introduction

This project is about building a smart system that can search and retrieve information from PDF documents. Since PDFs often contain a mix of text, tables, and figures, it can be tricky to find exactly what you're looking for. The system uses a method called Retrieval-Augmented Generation (RAG) to understand and pull out information from both the written content and the tables. This way, whether your question is about the text or data in a table, the system can give you clear and helpful answers.

Methodology

PDF Table and Figure Extraction (Using Docling)

The PDF content was converted into a structured tabular format by exploring several extraction tools, including **PyMuPDF**, **PyMuPDF-LLM**, and **pdfplumber**. These tools, however, failed to accurately preserve the tables when converting to markdown, often losing key content or altering the original layout.

Subsequently, **Docling** was used, which successfully extracted tables and figures while maintaining the original format and content integrity. Docling's extraction provides a faithful representation of the PDF's data, enabling more accurate retrieval and querying.

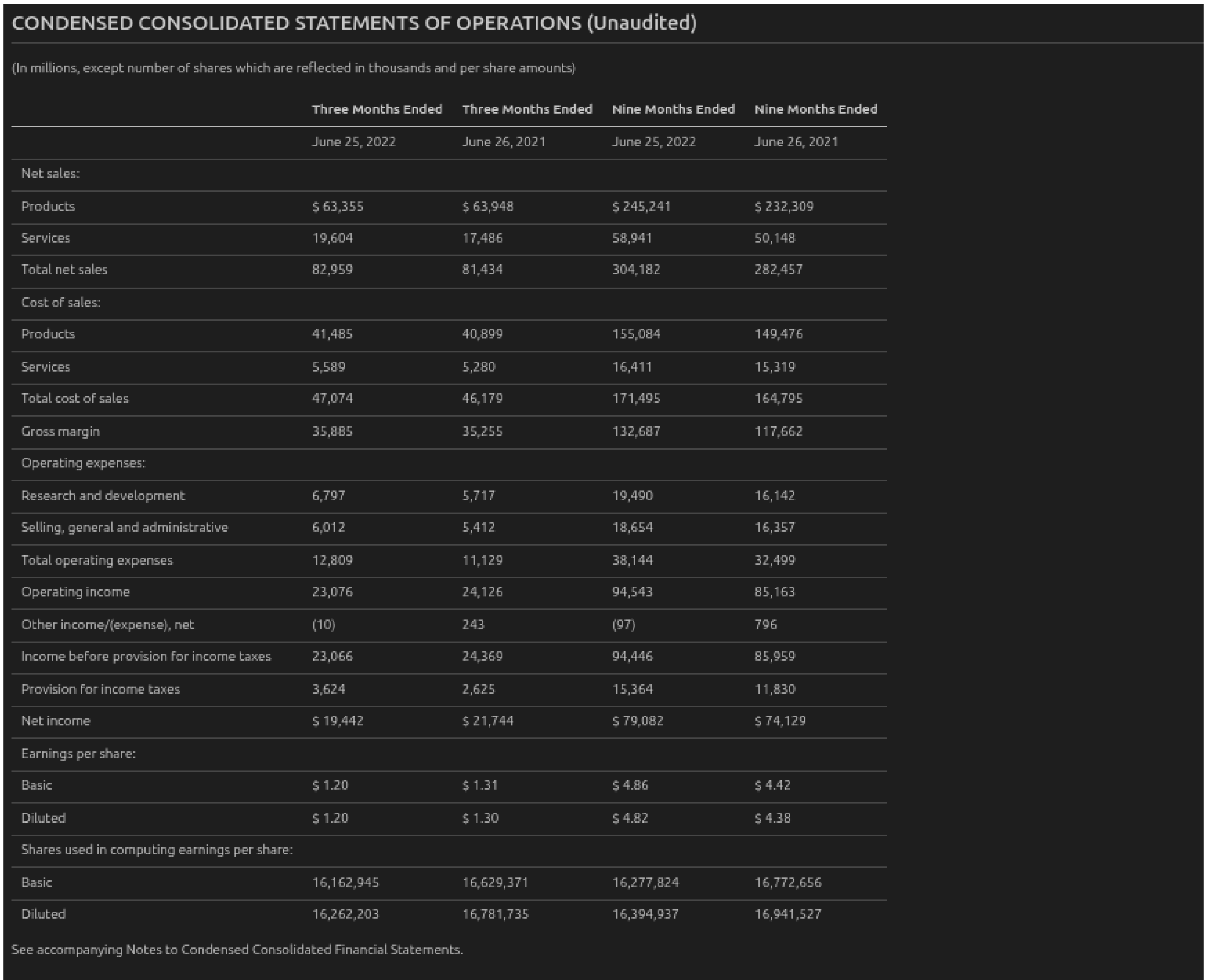
Below are samples of the tables and markdown extracted using Docling, demonstrating its effectiveness in preserving the document's structure.

Apple Inc.

CONDENSED CONSOLIDATED STATEMENTS OF OPERATIONS (Unaudited)
(In millions, except number of shares which are reflected in thousands and per share amounts)

	Three Months Ended		Nine Months Ended	
	June 25, 2022	June 26, 2021	June 25, 2022	June 26, 2021
Net sales:				
Products	$ 63,355	$ 63,948	$ 245,241	$ 232,309
Services	19,604	17,486	58,941	50,148
Total net sales	82,959	81,434	304,182	282,457
Cost of sales:				
Products	41,485	40,899	155,084	149,476
Services	5,589	5,280	16,411	15,319
Total cost of sales	47,074	46,179	171,495	164,795
Gross margin	35,885	35,255	132,687	117,662
Operating expenses:				
Research and development	6,797	5,717	19,490	16,142
Selling, general and administrative	6,012	5,412	18,654	16,357
Total operating expenses	12,809	11,129	38,144	32,499
Operating income	23,076	24,126	94,543	85,163
Other income/(expense), net	(10)	243	(97)	796
Income before provision for income taxes	23,066	24,369	94,446	85,959
Provision for income taxes	3,624	2,625	15,364	11,830
Net income	$ 19,442	$ 21,744	$ 79,082	$ 74,129
Earnings per share:				
Basic	$ 1.20	$ 1.31	$ 4.86	$ 4.42
Diluted	$ 1.20	$ 1.30	$ 4.82	$ 4.38
Shares used in computing earnings per share:				
Basic	16,162,945	16,629,371	16,277,824	16,772,656
Diluted	16,262,203	16,781,735	16,394,937	16,941,527

See accompanying Notes to Condensed Consolidated Financial Statements.

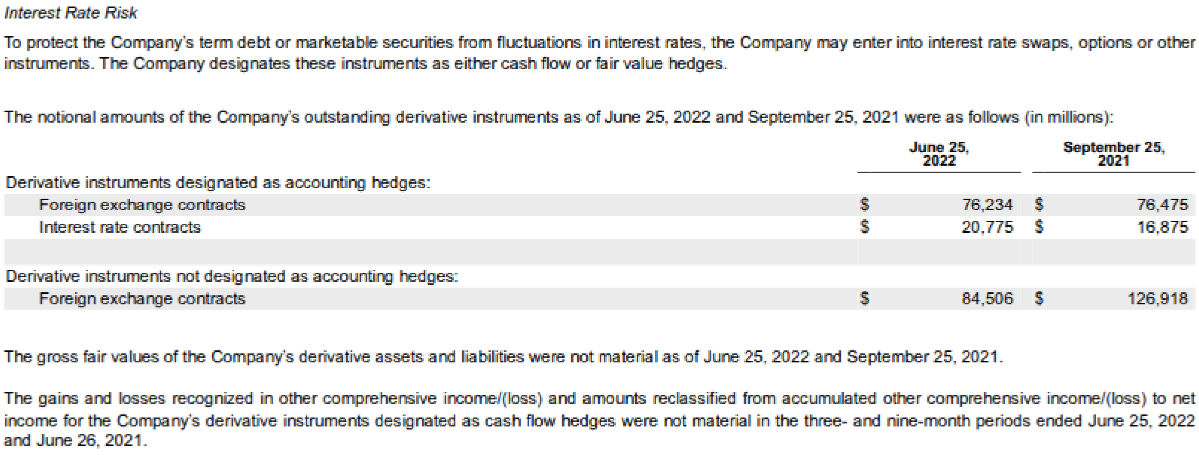


CONDENSED CONSOLIDATED STATEMENTS OF OPERATIONS (Unaudited)

(In millions, except number of shares which are reflected in thousands and per share amounts)

	Three Months Ended	Three Months Ended	Nine Months Ended	Nine Months Ended
	June 25, 2022	June 26, 2021	June 25, 2022	June 26, 2021
Net sales:				
Products	$ 63,355	$ 63,948	$ 245,241	$ 232,309
Services	19,604	17,486	58,941	50,148
Total net sales	82,959	81,434	304,182	282,457
Cost of sales:				
Products	41,485	40,899	155,084	149,476
Services	5,589	5,280	16,411	15,319
Total cost of sales	47,074	46,179	171,495	164,795
Gross margin	35,885	35,255	132,687	117,662
Operating expenses:				
Research and development	6,797	5,717	19,490	16,142
Selling, general and administrative	6,012	5,412	18,654	16,357
Total operating expenses	12,809	11,129	38,144	32,499
Operating income	23,076	24,126	94,543	85,163
Other income/(expense), net	(10)	243	(97)	796
Income before provision for income taxes	23,066	24,369	94,446	85,959
Provision for income taxes	3,624	2,625	15,364	11,830
Net income	$ 19,442	$ 21,744	$ 79,082	$ 74,129
Earnings per share:				
Basic	$ 1.20	$ 1.31	$ 4.86	$ 4.42
Diluted	$ 1.20	$ 1.30	$ 4.82	$ 4.38
Shares used in computing earnings per share:				
Basic	16,162,945	16,629,371	16,277,824	16,772,656
Diluted	16,262,203	16,781,735	16,394,937	16,941,527

See accompanying Notes to Condensed Consolidated Financial Statements.



Interest Rate Risk

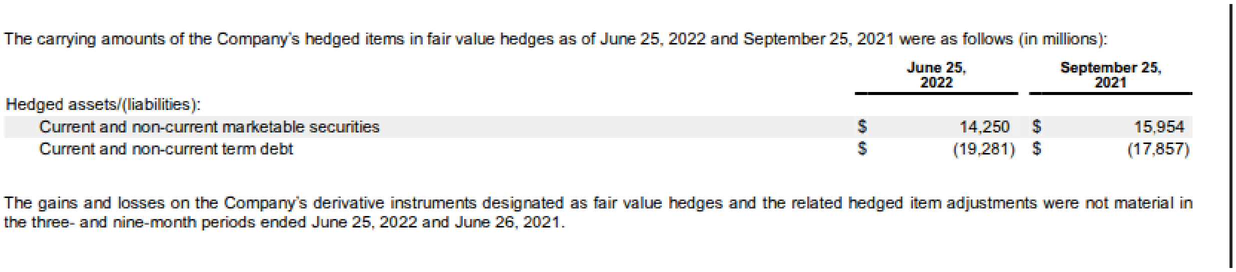
To protect the Company's term debt or marketable securities from fluctuations in interest rates, the Company may enter into interest rate swaps, options or other instruments. The Company designates these instruments as either cash flow or fair value hedges.

The notional amounts of the Company's outstanding derivative instruments as of June 25, 2022 and September 25, 2021 were as follows (in millions):

	June 25, 2022	September 25, 2021
Derivative instruments designated as accounting hedges:		
Foreign exchange contracts	$ 76,234	$ 76,475
Interest rate contracts	$ 20,775	$ 16,875
Derivative instruments not designated as accounting hedges:		
Foreign exchange contracts	$ 84,506	$ 126,918

The gross fair values of the Company's derivative assets and liabilities were not material as of June 25, 2022 and September 25, 2021.

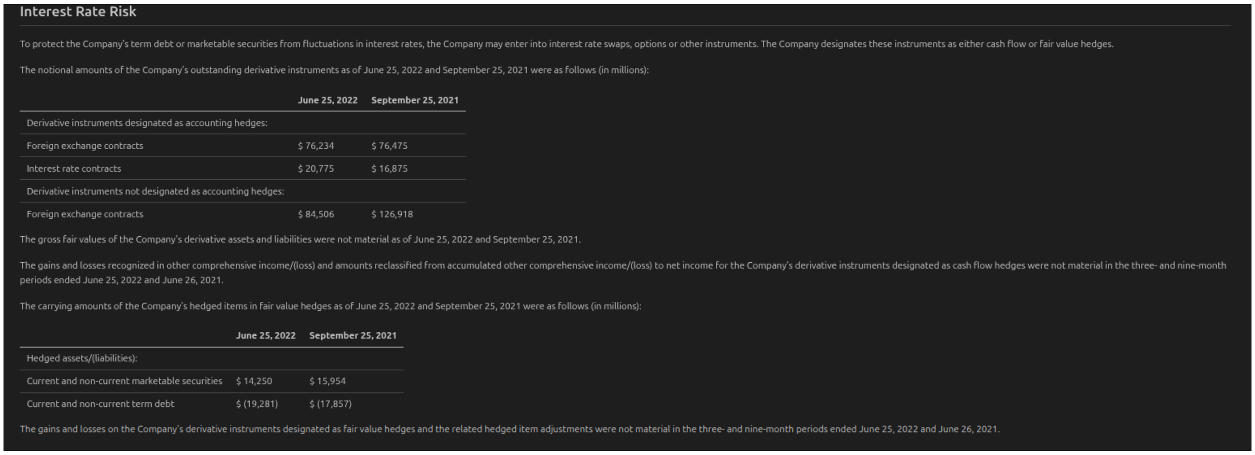
The gains and losses recognized in other comprehensive income/(loss) and amounts reclassified from accumulated other comprehensive income/(loss) to net income for the Company's derivative instruments designated as cash flow hedges were not material in the three- and nine-month periods ended June 25, 2022 and June 26, 2021.



The carrying amounts of the Company's hedged items in fair value hedges as of June 25, 2022 and September 25, 2021 were as follows (in millions):

	June 25, 2022	September 25, 2021
Hedged assets/(liabilities):		
Current and non-current marketable securities	$ 14,250	$ 15,954
Current and non-current term debt	$ (19,281)	$ (17,857)

The gains and losses on the Company's derivative instruments designated as fair value hedges and the related hedged item adjustments were not material in the three- and nine-month periods ended June 25, 2022 and June 26, 2021.



Interest Rate Risk

To protect the Company's term debt or marketable securities from fluctuations in interest rates, the Company may enter into interest rate swaps, options or other instruments. The Company designates these instruments as either cash flow or fair value hedges.

The notional amounts of the Company's outstanding derivative instruments as of June 25, 2022 and September 25, 2021 were as follows (in millions):

	June 25, 2022	September 25, 2021
Derivative instruments designated as accounting hedges:		
Foreign exchange contracts	$ 76,234	$ 76,475
Interest rate contracts	$ 20,775	$ 16,875
Derivative instruments not designated as accounting hedges:		
Foreign exchange contracts	$ 84,506	$ 126,918

The gross fair values of the Company's derivative assets and liabilities were not material as of June 25, 2022 and September 25, 2021.

The gains and losses recognized in other comprehensive income/(loss) and amounts reclassified from accumulated other comprehensive income/(loss) to net income for the Company's derivative instruments designated as cash flow hedges were not material in the three- and nine-month periods ended June 25, 2022 and June 26, 2021.

The carrying amounts of the Company's hedged items in fair value hedges as of June 25, 2022 and September 25, 2021 were as follows (in millions):

	June 25, 2022	September 25, 2021
Hedged assets/(liabilities):		
Current and non-current marketable securities	$ 14,250	$ 15,954
Current and non-current term debt	$ (19,281)	$ (17,857)

The gains and losses on the Company's derivative instruments designated as fair value hedges and the related hedged item adjustments were not material in the three- and nine-month periods ended June 25, 2022 and June 26, 2021.

Preserving Table Structure During Chunking

The extracted document was chunked using a Markdown header splitter, which divides the content at each level 2 header (##). This process ensures that tables are not split into multiple chunks, preserving their entire structure and content within a single chunk.

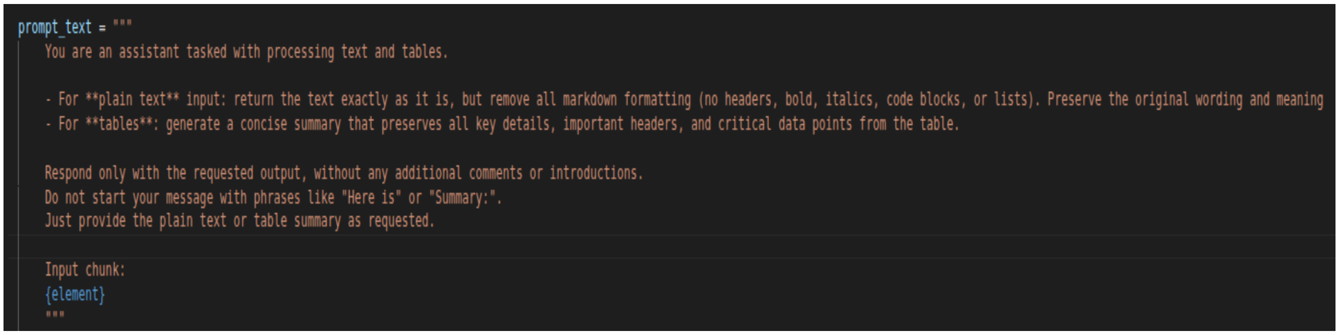
Maintaining tables intact is important because splitting a table across chunks can lead to incomplete or fragmented data, which would reduce the accuracy and usefulness of any retrieval or analysis performed later. By keeping each table whole, the system can better understand and respond to queries involving complex tabular data.

After chunking, a Document object from LangChain was created for each chunk, enabling efficient indexing and retrieval of both textual and tabular content.

Using LLM for Table and Text Summarization

The LLM model **“llama-3.3-70b-versatile”** was used with carefully designed prompts to summarize the content of each chunked document, especially focusing on the tables. These summaries help create embeddings for the retrieval model.

The prompts were crafted to ensure that key points and important details from both tables and text are preserved during summarization. This approach guarantees that no essential information is lost when generating embeddings for each chunk, improving the overall quality and accuracy of the retrieval system.



```
prompt_text = """
    You are an assistant tasked with processing text and tables.

    - For **plain text** input: return the text exactly as it is, but remove all markdown formatting (no headers, bold, italics, code blocks, or lists). Preserve the original wording and meaning
    - For **tables**: generate a concise summary that preserves all key details, important headers, and critical data points from the table.

    Respond only with the requested output, without any additional comments or introductions.
    Do not start your message with phrases like "Here is" or "Summary:".
    Just provide the plain text or table summary as requested.

    Input chunk:
    {element}
    """
```

Vector Storage Using Qdrant

Qdrant serves as the vector store for efficient semantic search, indexing embedding vectors generated from document summaries. The embedding model **models/gemini-embedding-001** creates dense vector representations of each summarized chunk, capturing the essential meaning of the content.

The embedding vectors are stored in Qdrant’s vector space, while the **actual chunked document data (the real content)** is kept in the payload metadata. This design ensures that each vector corresponds to a meaningful summary embedding, while the full chunk data is readily available for retrieval and context.

This structure enables fast and accurate retrieval of relevant document sections based on vector similarity, preserving both the semantic summary and the original content.

Retrieval and Response Generation in RAG Framework

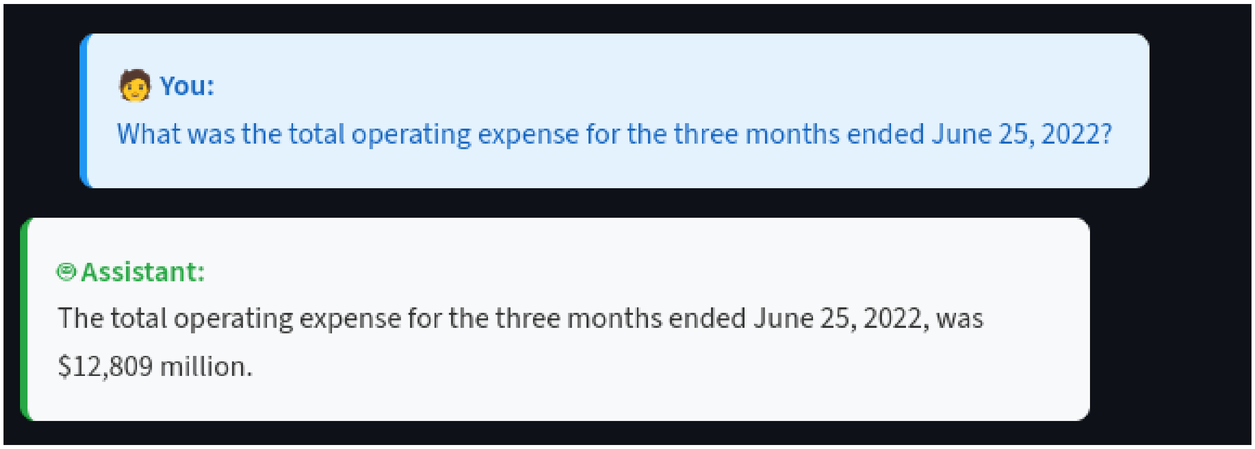
In the retrieval phase, the vector store (Qdrant) performs a nearest neighbor search using the embedding of the user query to identify the most relevant document chunks. These retrieved chunks serve as external knowledge context for the LLM model.

The LLM, specifically **“llama-3.3-70b-versatile”**, receives this contextual information as part of its prompt to generate accurate, context-aware responses. This retrieval-augmented generation approach leverages the semantic embeddings stored in Qdrant to ground the model’s outputs in the actual document content, improving response relevance and factual accuracy.

The entire retrieval and generation pipeline is orchestrated via a **FastAPI** backend, The frontend, built with **Streamlit**, streams results to the user.

Results

Question:



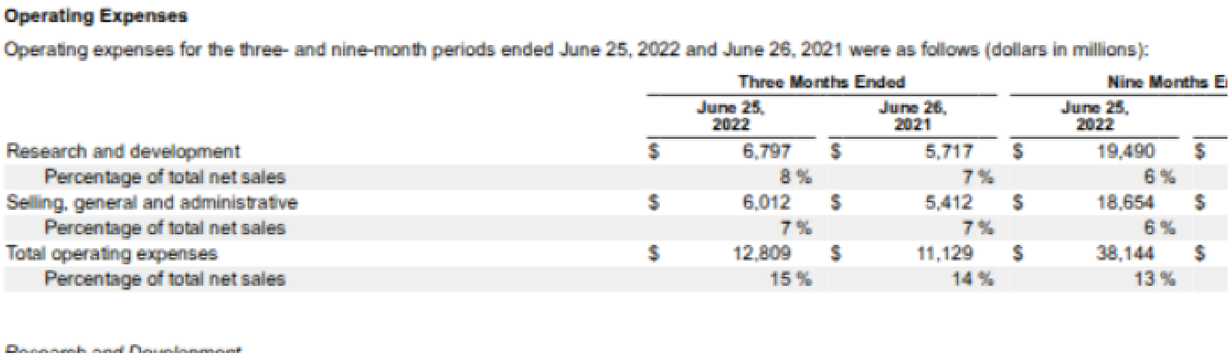
🧑 **You:**

What was the total operating expense for the three months ended June 25, 2022?

Assistant:

The total operating expense for the three months ended June 25, 2022, was $12,809 million.

Real Context :



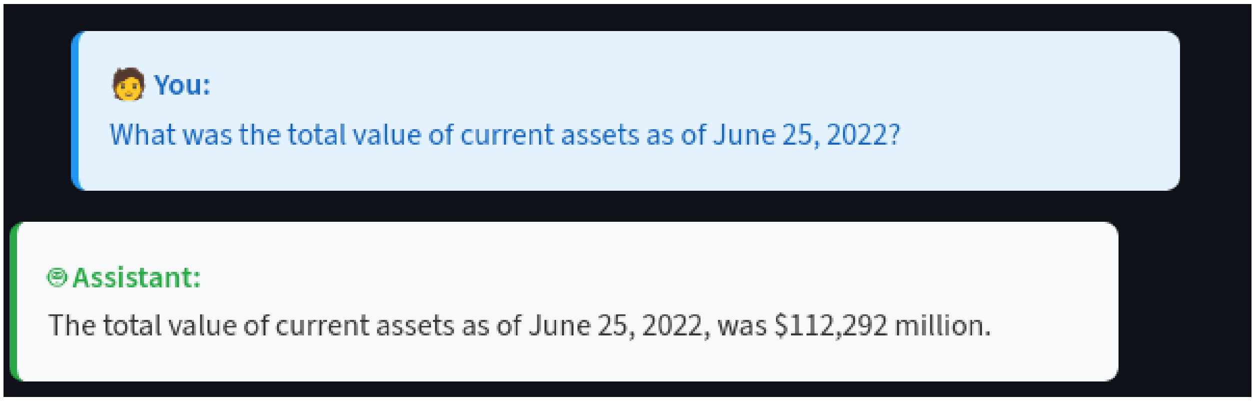
Operating Expenses

Operating expenses for the three- and nine-month periods ended June 25, 2022 and June 26, 2021 were as follows (dollars in millions):

	Three Months Ended		Nine Months E	
	June 25, 2022	June 26, 2021	June 25, 2022	
Research and development	$ 6,797	$ 5,717	$ 19,490	$
Percentage of total net sales	8 %	7 %	6 %	
Selling, general and administrative	$ 6,012	$ 5,412	$ 18,654	$
Percentage of total net sales	7 %	7 %	6 %	
Total operating expenses	$ 12,809	$ 11,129	$ 38,144	$
Percentage of total net sales	15 %	14 %	13 %	

Research and Development

Question :

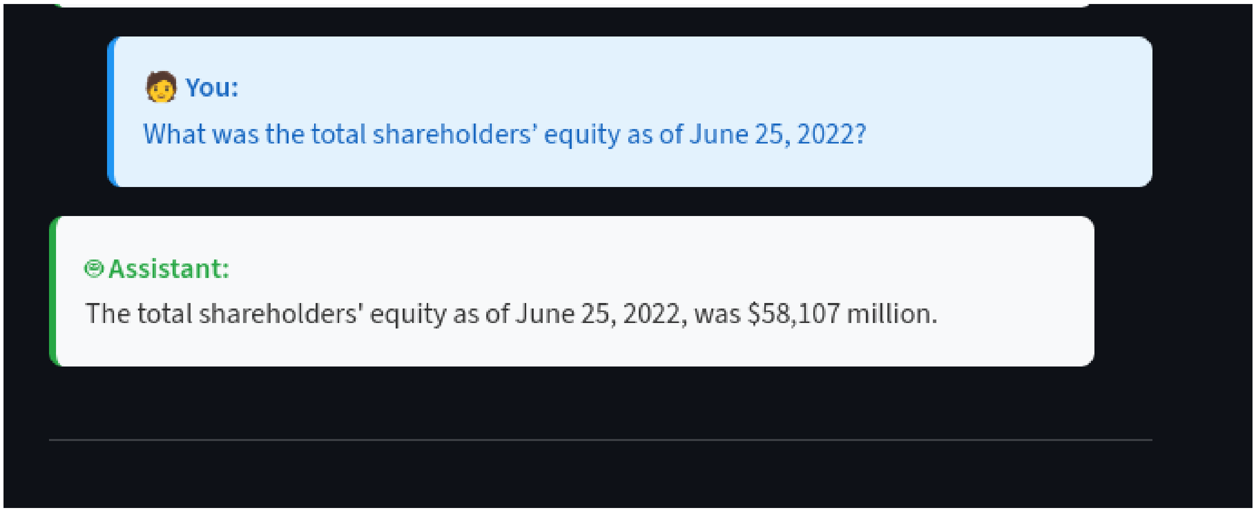


🧑 **You:**

What was the total value of current assets as of June 25, 2022?

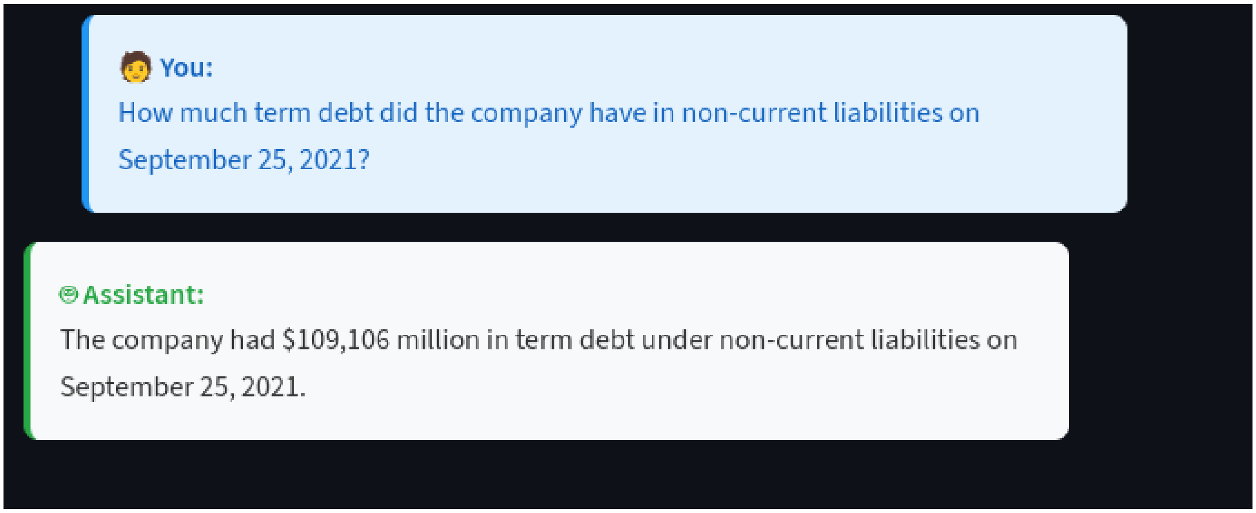
Assistant:

The total value of current assets as of June 25, 2022, was $112,292 million.



🧑 You:
What was the total shareholders’ equity as of June 25, 2022?

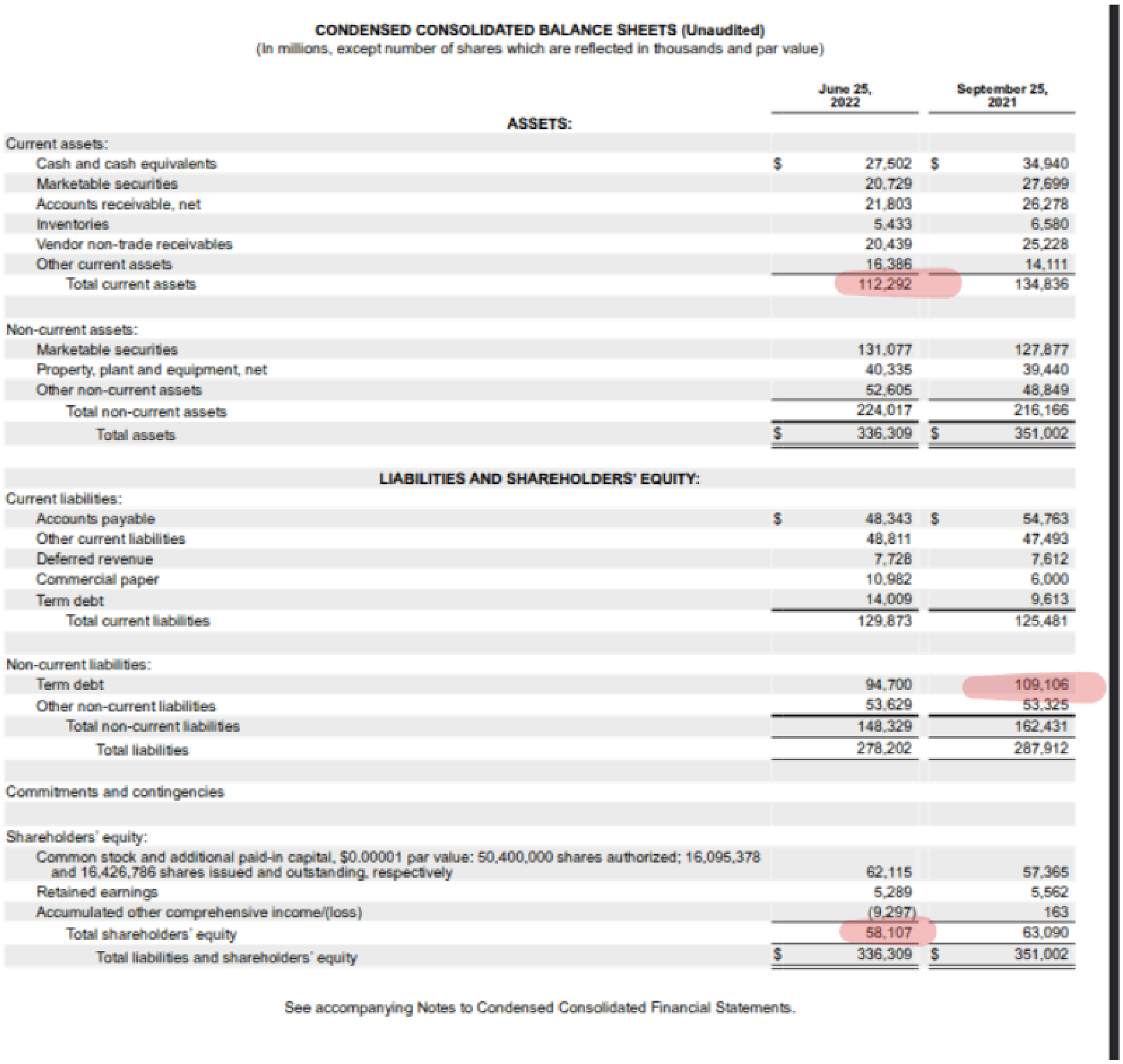
🤖 Assistant:
The total shareholders' equity as of June 25, 2022, was $58,107 million.



🧑 You:
How much term debt did the company have in non-current liabilities on September 25, 2021?

🤖 Assistant:
The company had $109,106 million in term debt under non-current liabilities on September 25, 2021.

Real Context :



CONDENSED CONSOLIDATED BALANCE SHEETS (Unaudited)

(In millions, except number of shares which are reflected in thousands and par value)

	June 25, 2022	September 25, 2021
ASSETS:		
Current assets:		
Cash and cash equivalents	$ 27,502	$ 34,940
Marketable securities	20,729	27,699
Accounts receivable, net	21,803	26,278
Inventories	5,433	6,580
Vendor non-trade receivables	20,439	25,228
Other current assets	16,386	14,111
Total current assets	112,292	134,836
Non-current assets:		
Marketable securities	131,077	127,877
Property, plant and equipment, net	40,335	39,440
Other non-current assets	52,605	48,849
Total non-current assets	224,017	216,166
Total assets	$ 336,309	$ 351,002
LIABILITIES AND SHAREHOLDERS' EQUITY:		
Current liabilities:		
Accounts payable	$ 48,343	$ 54,763
Other current liabilities	48,811	47,493
Deferred revenue	7,728	7,612
Commercial paper	10,982	6,000
Term debt	14,009	9,613
Total current liabilities	129,873	125,481
Non-current liabilities:		
Term debt	94,700	109,106
Other non-current liabilities	53,629	53,325
Total non-current liabilities	148,329	162,431
Total liabilities	278,202	287,912
Commitments and contingencies		
Shareholders' equity:		
Common stock and additional paid-in capital, $0.00001 par value: 50,400,000 shares authorized; 16,095,378 and 16,426,786 shares issued and outstanding, respectively	62,115	57,365
Retained earnings	5,289	5,562
Accumulated other comprehensive income/(loss)	(9,297)	163
Total shareholders' equity	58,107	63,090
Total liabilities and shareholders' equity	$ 336,309	$ 351,002

See accompanying Notes to Condensed Consolidated Financial Statements.

Conclusion

The PDF was successfully extracted into markdown format without any errors, preserving both textual and tabular content accurately. The retrieval process produced effective and relevant results, thanks largely to the high-quality conversion by Docling.

While most components of the system operated with low latency and high efficiency, the Docling PDF-to-markdown conversion step was noticeably slower, impacting overall response time. Nonetheless, this approach remains robust and reliable for processing complex documents containing tables and figures.