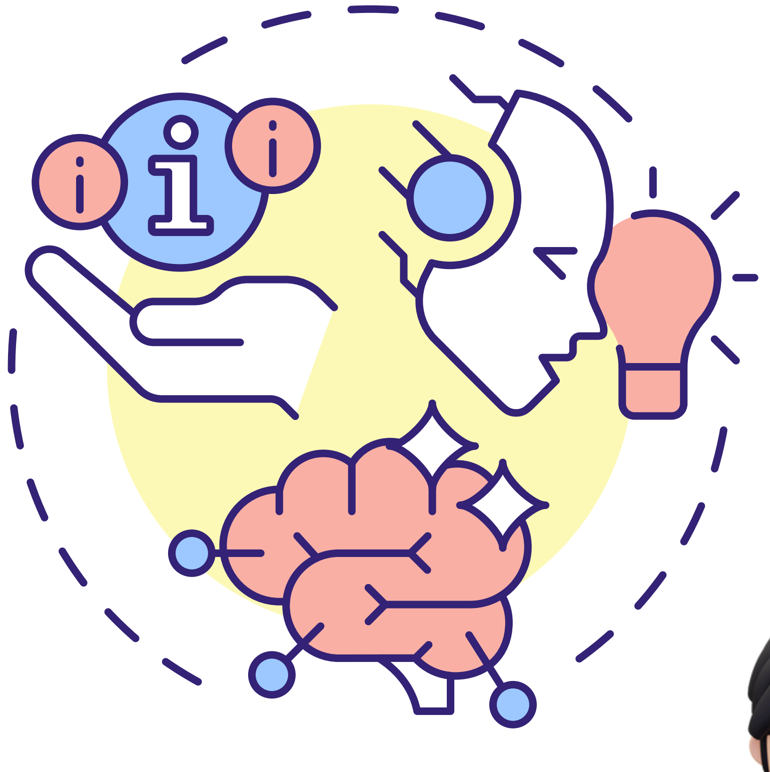


# AUTOMATED FINANCIAL DATA EXTRACTION FROM REGULATORY FILINGS using RAG

TEAM NAME

CHENNAI DATA FOLKS

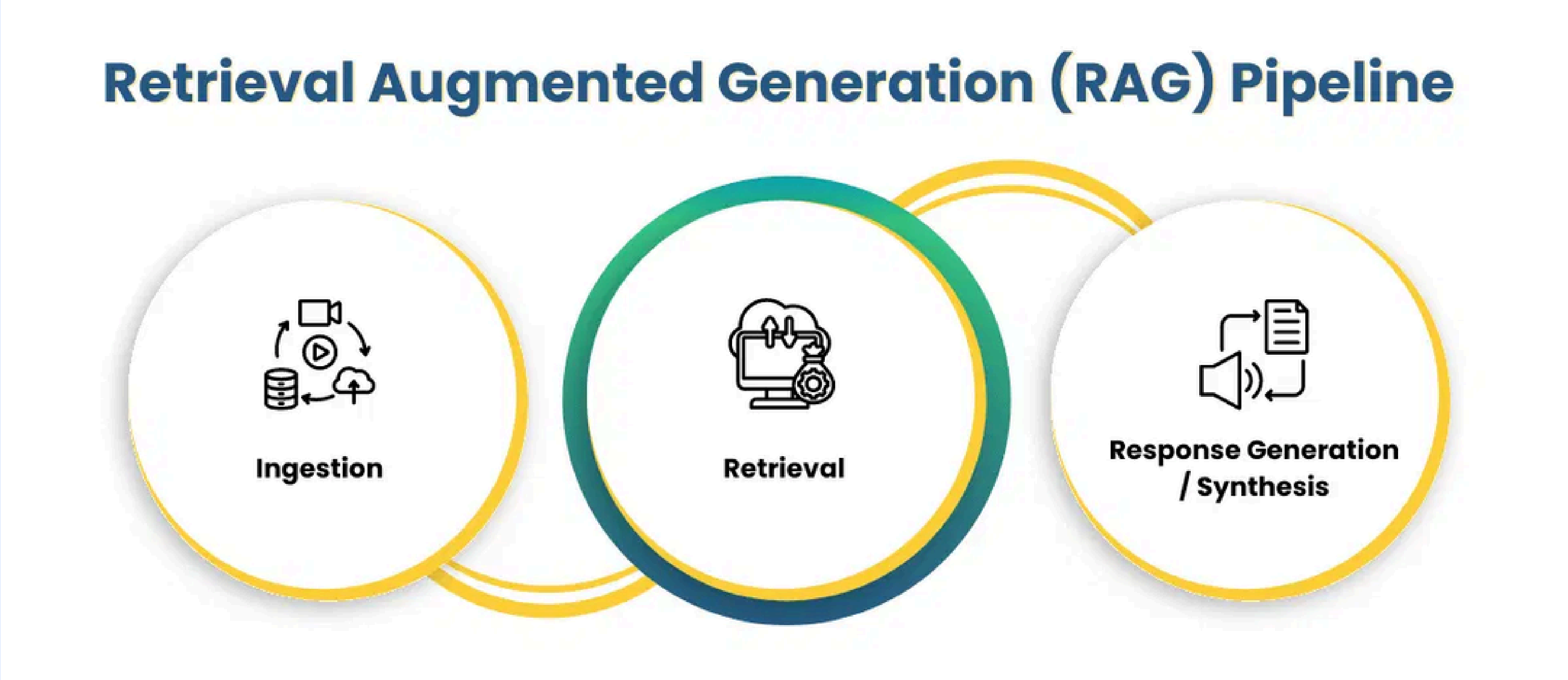
# THE CHALLENGE

- Time-consuming and labor-intensive.
- Prone to human error.
- Difficult to scale across multiple companies and filings.
- Lack of transparency and auditability.

# INTELLIGENT AUTOMATION WITH RETRIEVAL AUGMENTED GENERATION

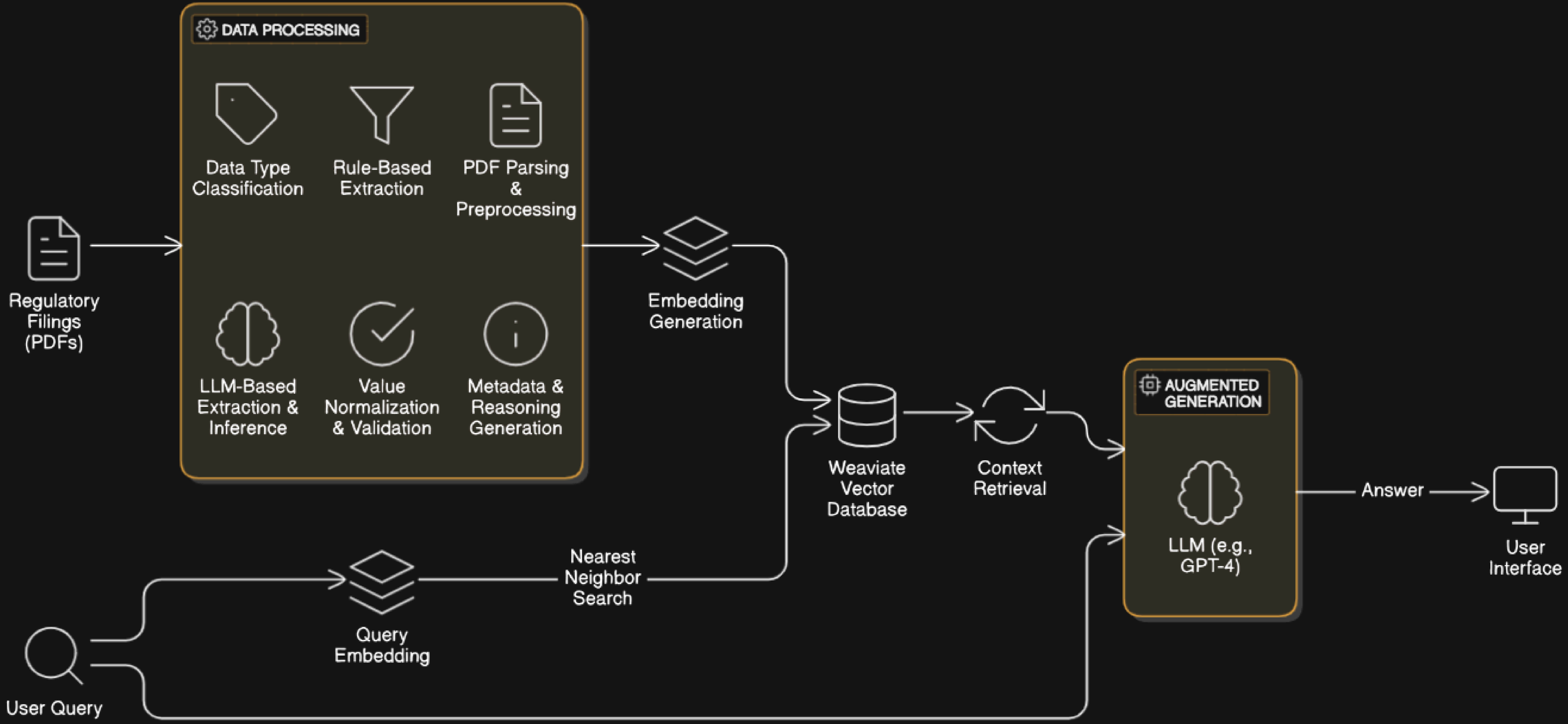
**Combining the power of information retrieval with the intelligence of LLMs.**

**Retrieval:**
Finding the most relevant information from the filings.


Retrieval Augmented Generation (RAG) Pipeline
Ingestion
Retrieval
Response Generation / Synthesis

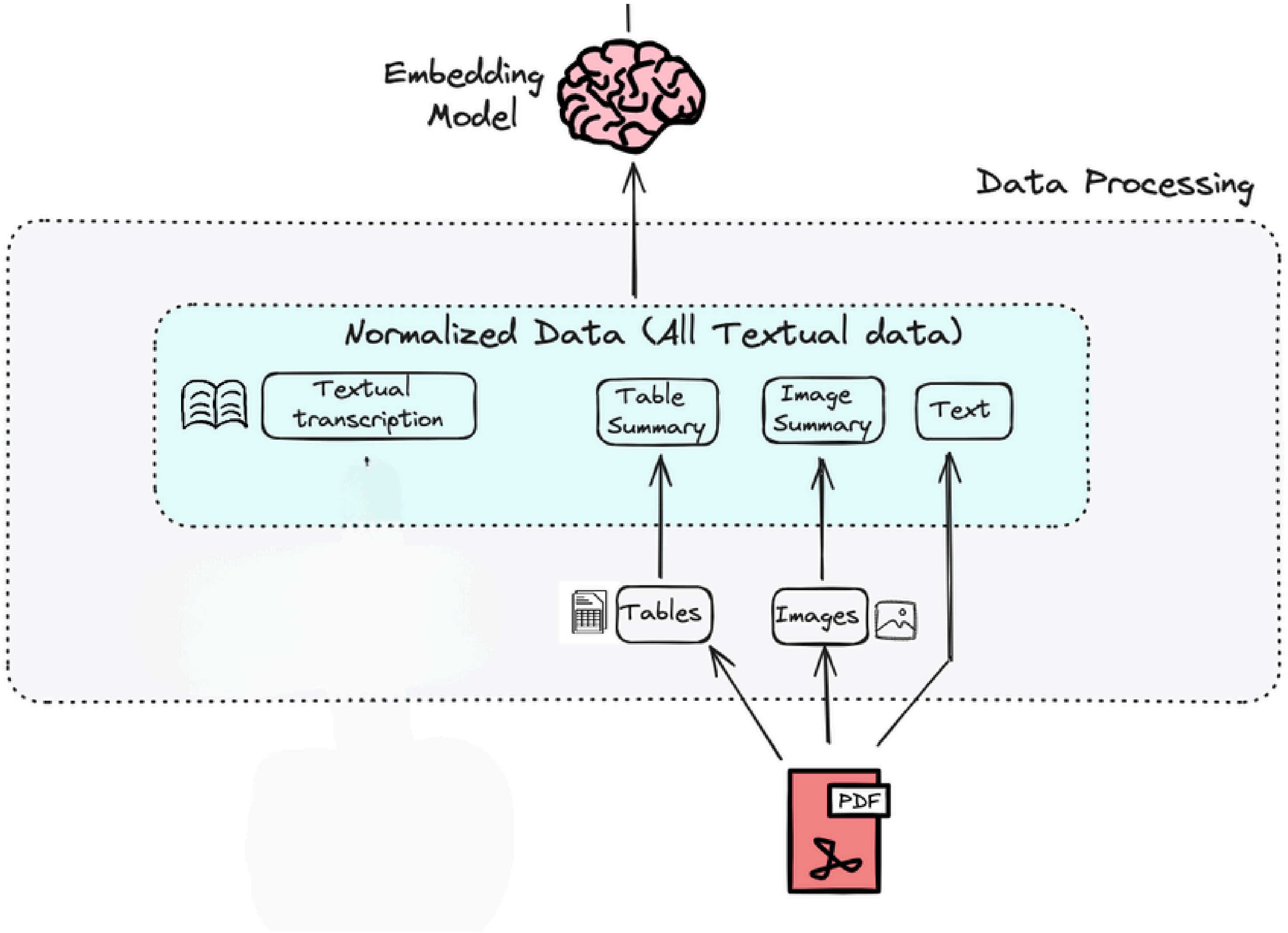
**Generation:**
Using this information to generate accurate and insightful results.

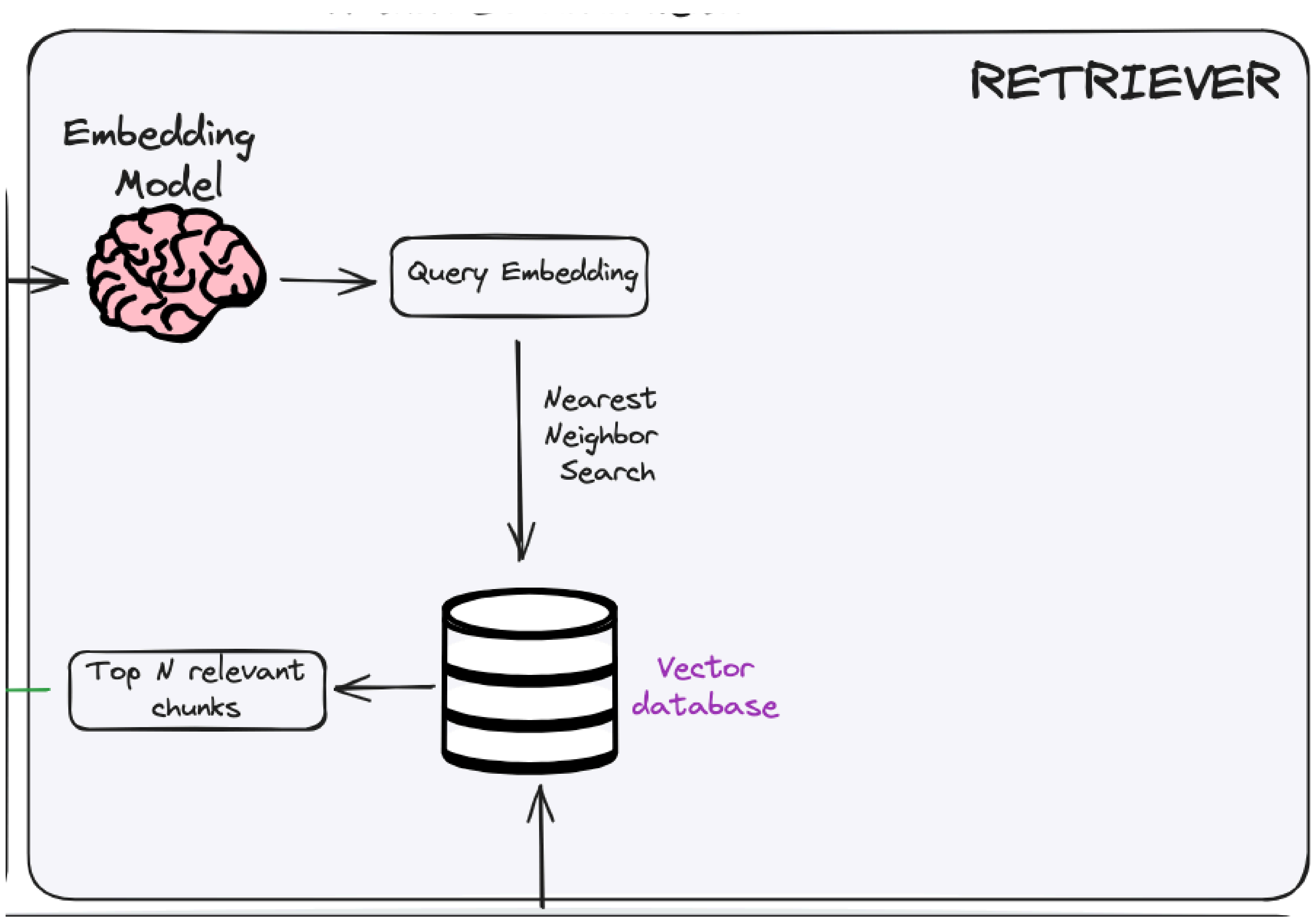
# Retrieval Augmented Generation (RAG) System for Financial Data Extraction



# BUILDING THE KNOWLEDGE BASE: FROM PDFS TO VECTORS

- PDFs are parsed, and relevant data is extracted.
- LLMs generate summaries and perform calculations.
- Metadata (snippets, coordinates, page numbers) is captured.
- Embeddings are generated for text and summaries.
- All this information is stored in Weaviate.



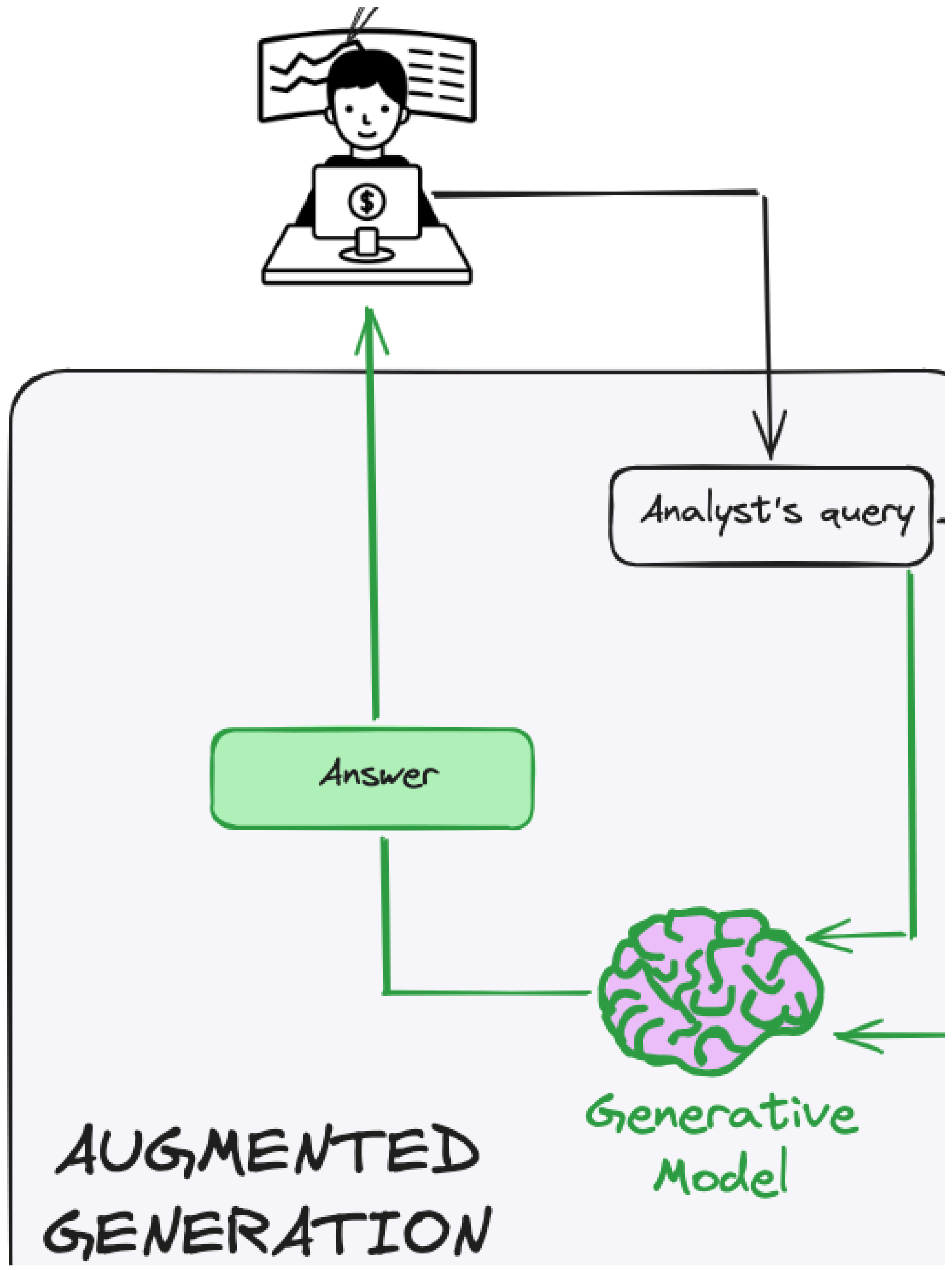


# FINDING THE RIGHT INFORMATION: EFFICIENT AND ACCURATE RETRIEVAL

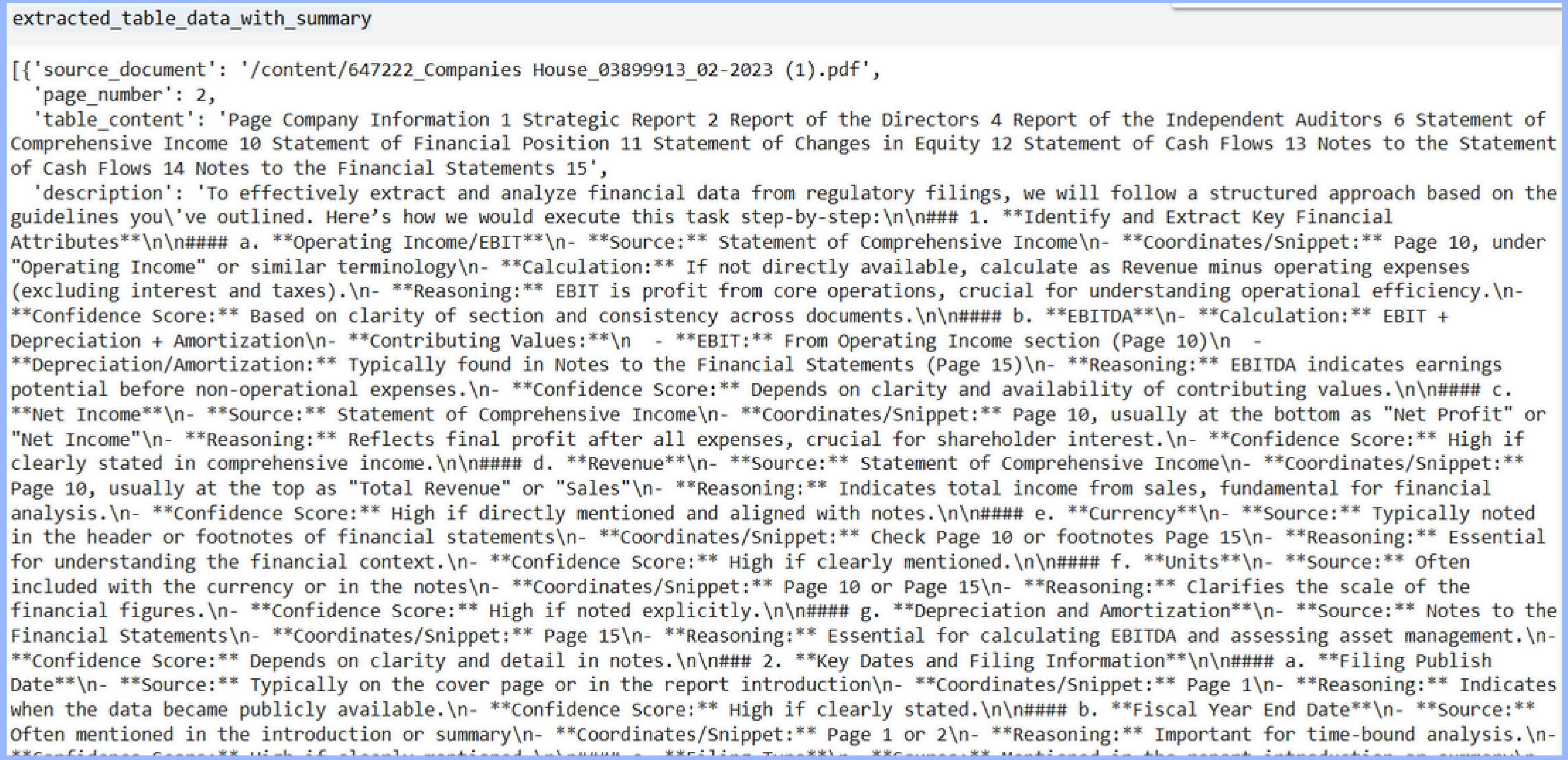
- User query is embedded.
- Weaviate performs a fast similarity search.
- Relevant data (context) is retrieved.

# GENERATING INTELLIGENT RESPONSES: LEVERAGING LLMS WITH CONTEXT

- The context is included in the prompt to the LLM.
- The LLM generates a concise and accurate response, grounded in the retrieved information.
- Reasoning and source attribution are provided.



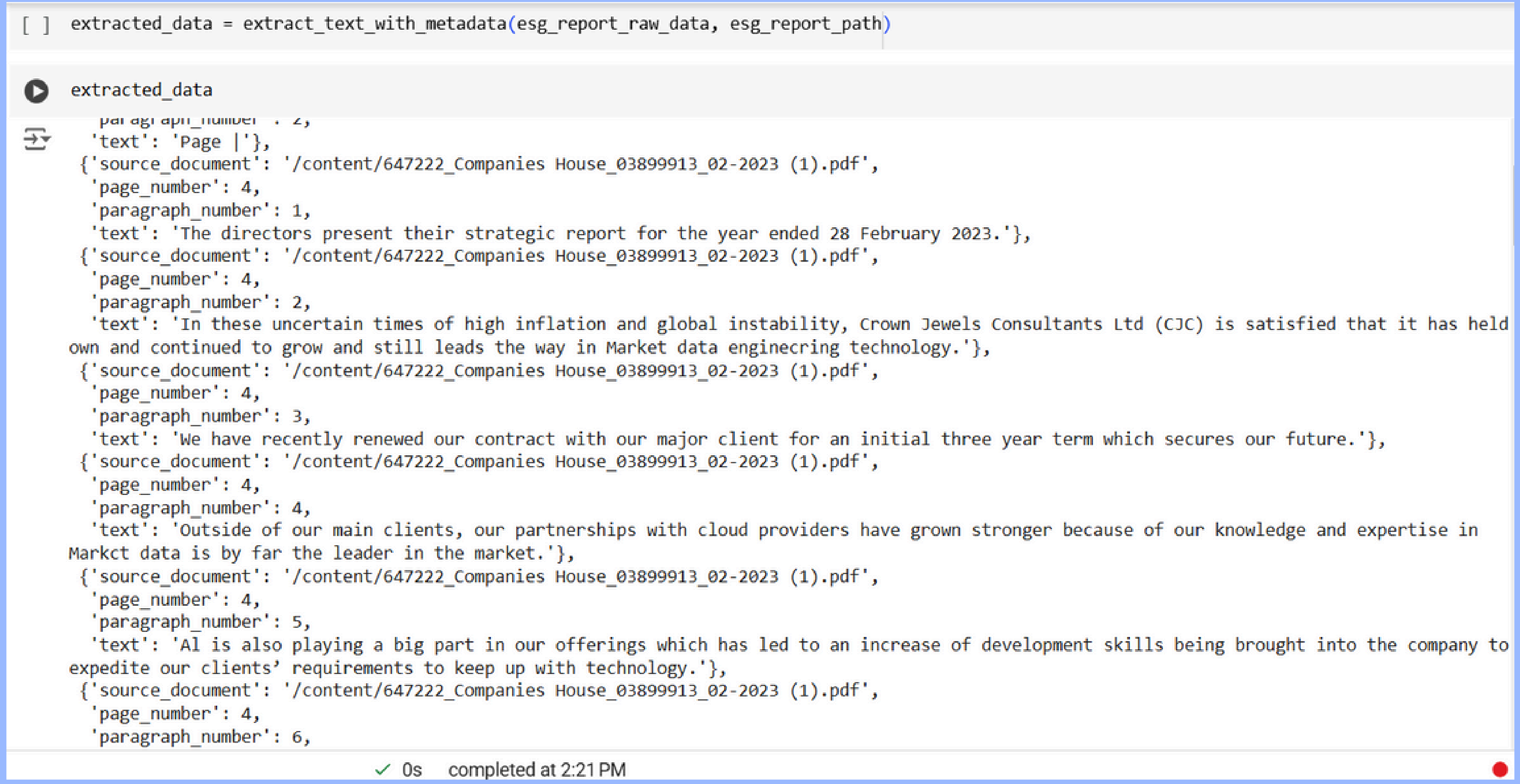
# SOME SCREENSHOTS



```
extracted_table_data_with_summary
```

```
[{'source_document': '/content/647222_Companies House_03899913_02-2023 (1).pdf',
  'page_number': 2,
  'table_content': 'Page Company Information 1 Strategic Report 2 Report of the Directors 4 Report of the Independent Auditors 6 Statement of
Comprehensive Income 10 Statement of Financial Position 11 Statement of Changes in Equity 12 Statement of Cash Flows 13 Notes to the Statement
of Cash Flows 14 Notes to the Financial Statements 15',
  'description': 'To effectively extract and analyze financial data from regulatory filings, we will follow a structured approach based on the
guidelines you\'ve outlined. Here’s how we would execute this task step-by-step:\n\n### 1. **Identify and Extract Key Financial
Attributes**\n\n#### a. **Operating Income/EBIT**\n- **Source:** Statement of Comprehensive Income\n- **Coordinates/Snippet:** Page 10, under
"Operating Income" or similar terminology\n- **Calculation:** If not directly available, calculate as Revenue minus operating expenses
(excluding interest and taxes).\n- **Reasoning:** EBIT is profit from core operations, crucial for understanding operational efficiency.\n-
**Confidence Score:** Based on clarity of section and consistency across documents.\n\n#### b. **EBITDA**\n- **Calculation:** EBIT +
Depreciation + Amortization\n- **Contributing Values:**\n  - **EBIT:** From Operating Income section (Page 10)\n  -
**Depreciation/Amortization:** Typically found in Notes to the Financial Statements (Page 15)\n- **Reasoning:** EBITDA indicates earnings
potential before non-operational expenses.\n- **Confidence Score:** Depends on clarity and availability of contributing values.\n\n#### c.
**Net Income**\n- **Source:** Statement of Comprehensive Income\n- **Coordinates/Snippet:** Page 10, usually at the bottom as "Net Profit" or
"Net Income"\n- **Reasoning:** Reflects final profit after all expenses, crucial for shareholder interest.\n- **Confidence Score:** High if
clearly stated in comprehensive income.\n\n#### d. **Revenue**\n- **Source:** Statement of Comprehensive Income\n- **Coordinates/Snippet:**
Page 10, usually at the top as "Total Revenue" or "Sales"\n- **Reasoning:** Indicates total income from sales, fundamental for financial
analysis.\n- **Confidence Score:** High if directly mentioned and aligned with notes.\n\n#### e. **Currency**\n- **Source:** Typically noted
in the header or footnotes of financial statements\n- **Coordinates/Snippet:** Check Page 10 or footnotes Page 15\n- **Reasoning:** Essential
for understanding the financial context.\n- **Confidence Score:** High if clearly mentioned.\n\n#### f. **Units**\n- **Source:** Often
included with the currency or in the notes\n- **Coordinates/Snippet:** Page 10 or Page 15\n- **Reasoning:** Clarifies the scale of the
financial figures.\n- **Confidence Score:** High if noted explicitly.\n\n#### g. **Depreciation and Amortization**\n- **Source:** Notes to the
Financial Statements\n- **Coordinates/Snippet:** Page 15\n- **Reasoning:** Essential for calculating EBITDA and assessing asset management.\n-
**Confidence Score:** Depends on clarity and detail in notes.\n\n### 2. **Key Dates and Filing Information**\n\n#### a. **Filing Publish
Date**\n- **Source:** Typically on the cover page or in the report introduction\n- **Coordinates/Snippet:** Page 1\n- **Reasoning:** Indicates
when the data became publicly available.\n- **Confidence Score:** High if clearly stated.\n\n#### b. **Fiscal Year End Date**\n- **Source:**
Often mentioned in the introduction or summary\n- **Coordinates/Snippet:** Page 1 or 2\n- **Reasoning:** Important for time-bound analysis.\n-
```

EXTRACTED SUMMARY
OF TABLE DATA



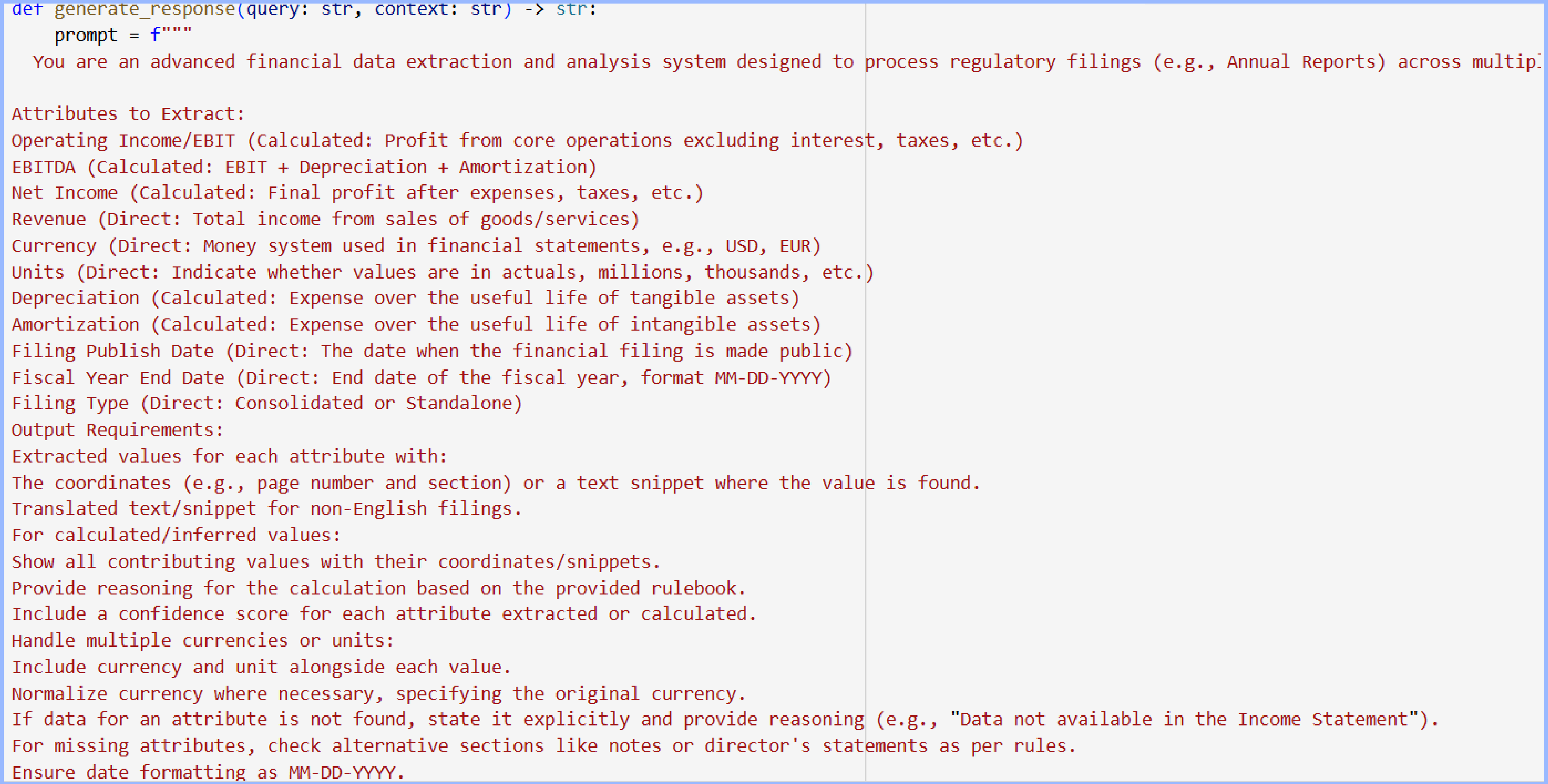
```
[ ]  extracted_data = extract_text_with_metadata(esg_report_raw_data, esg_report_path)
```

```
extracted_data
```

```
    'paragraph_number': 2,
    'text': 'Page |'},
   {'source_document': '/content/647222_Companies House_03899913_02-2023 (1).pdf',
    'page_number': 4,
    'paragraph_number': 1,
    'text': 'The directors present their strategic report for the year ended 28 February 2023.'},
   {'source_document': '/content/647222_Companies House_03899913_02-2023 (1).pdf',
    'page_number': 4,
    'paragraph_number': 2,
    'text': 'In these uncertain times of high inflation and global instability, Crown Jewels Consultants Ltd (CJC) is satisfied that it has held
  own and continued to grow and still leads the way in Market data engincering technology.'},
   {'source_document': '/content/647222_Companies House_03899913_02-2023 (1).pdf',
    'page_number': 4,
    'paragraph_number': 3,
    'text': 'We have recently renewed our contract with our major client for an initial three year term which secures our future.'},
   {'source_document': '/content/647222_Companies House_03899913_02-2023 (1).pdf',
    'page_number': 4,
    'paragraph_number': 4,
    'text': 'Outside of our main clients, our partnerships with cloud providers have grown stronger because of our knowledge and expertise in
  Markct data is by far the leader in the market.'},
   {'source_document': '/content/647222_Companies House_03899913_02-2023 (1).pdf',
    'page_number': 4,
    'paragraph_number': 5,
    'text': 'Al is also playing a big part in our offerings which has led to an increase of development skills being brought into the company to
  expedite our clients’ requirements to keep up with technology.'},
   {'source_document': '/content/647222_Companies House_03899913_02-2023 (1).pdf',
    'page_number': 4,
    'paragraph_number': 6,
```

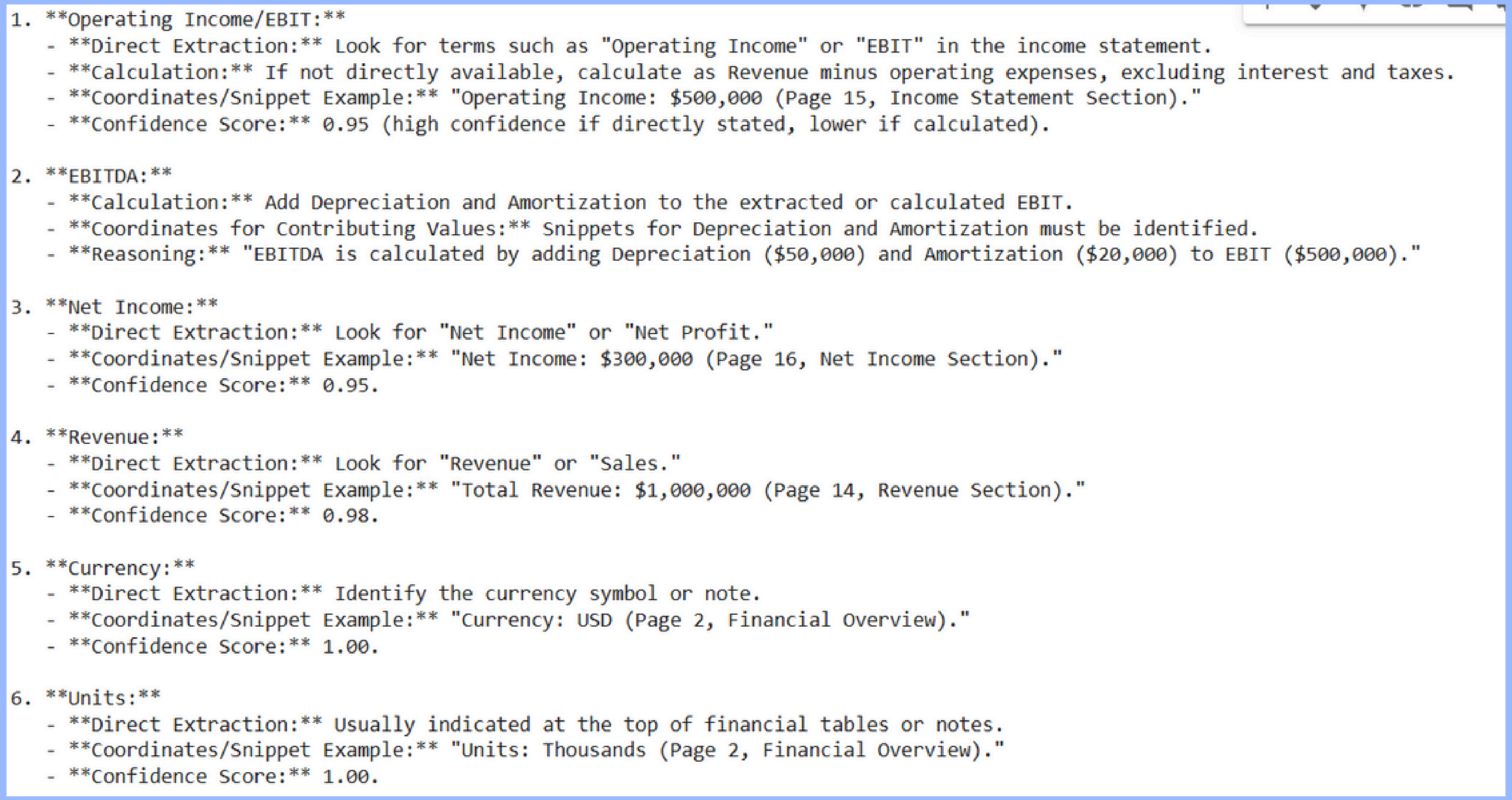
EXTRACTED SUMMARY
OF TEXT DATA

# PROMPT TO ACQUIRE THE OUTPUT

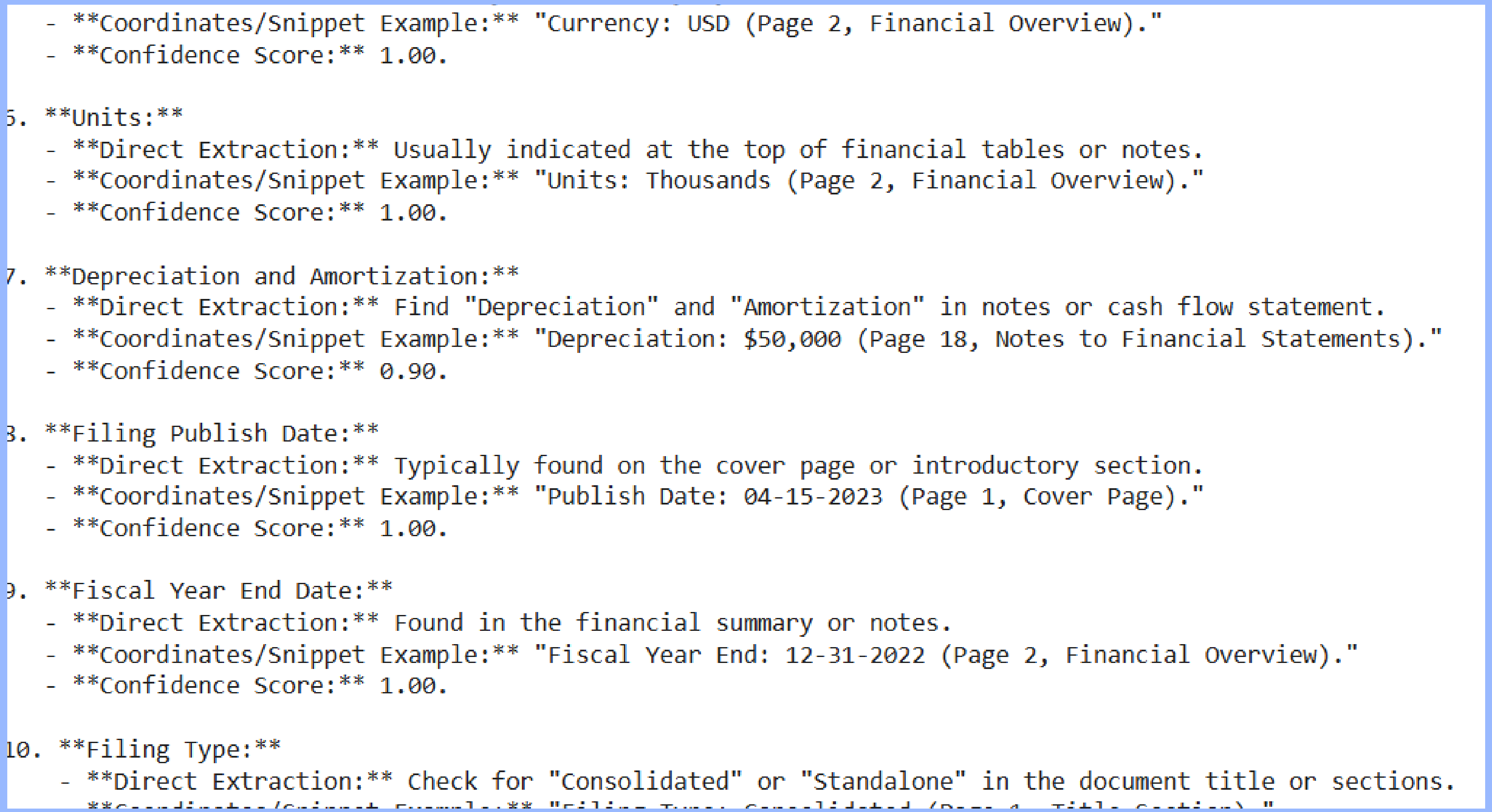


```
def generate_response(query: str, context: str) -> str:
    prompt = f"""
  You are an advanced financial data extraction and analysis system designed to process regulatory filings (e.g., Annual Reports) across multip

Attributes to Extract:
Operating Income/EBIT (Calculated: Profit from core operations excluding interest, taxes, etc.)
EBITDA (Calculated: EBIT + Depreciation + Amortization)
Net Income (Calculated: Final profit after expenses, taxes, etc.)
Revenue (Direct: Total income from sales of goods/services)
Currency (Direct: Money system used in financial statements, e.g., USD, EUR)
Units (Direct: Indicate whether values are in actuals, millions, thousands, etc.)
Depreciation (Calculated: Expense over the useful life of tangible assets)
Amortization (Calculated: Expense over the useful life of intangible assets)
Filing Publish Date (Direct: The date when the financial filing is made public)
Fiscal Year End Date (Direct: End date of the fiscal year, format MM-DD-YYYY)
Filing Type (Direct: Consolidated or Standalone)
Output Requirements:
Extracted values for each attribute with:
The coordinates (e.g., page number and section) or a text snippet where the value is found.
Translated text/snippet for non-English filings.
For calculated/inferred values:
Show all contributing values with their coordinates/snippets.
Provide reasoning for the calculation based on the provided rulebook.
Include a confidence score for each attribute extracted or calculated.
Handle multiple currencies or units:
Include currency and unit alongside each value.
Normalize currency where necessary, specifying the original currency.
If data for an attribute is not found, state it explicitly and provide reasoning (e.g., "Data not available in the Income Statement").
For missing attributes, check alternative sections like notes or director's statements as per rules.
Ensure date formatting as MM-DD-YYYY.
```

# RETREIVAL OUTPUT FROM PDF



```
1. **Operating Income/EBIT:**
   - **Direct Extraction:** Look for terms such as "Operating Income" or "EBIT" in the income statement.
   - **Calculation:** If not directly available, calculate as Revenue minus operating expenses, excluding interest and taxes.
   - **Coordinates/Snippet Example:** "Operating Income: $500,000 (Page 15, Income Statement Section)."
   - **Confidence Score:** 0.95 (high confidence if directly stated, lower if calculated).

2. **EBITDA:**
   - **Calculation:** Add Depreciation and Amortization to the extracted or calculated EBIT.
   - **Coordinates for Contributing Values:** Snippets for Depreciation and Amortization must be identified.
   - **Reasoning:** "EBITDA is calculated by adding Depreciation ($50,000) and Amortization ($20,000) to EBIT ($500,000)."

3. **Net Income:**
   - **Direct Extraction:** Look for "Net Income" or "Net Profit."
   - **Coordinates/Snippet Example:** "Net Income: $300,000 (Page 16, Net Income Section)."
   - **Confidence Score:** 0.95.

4. **Revenue:**
   - **Direct Extraction:** Look for "Revenue" or "Sales."
   - **Coordinates/Snippet Example:** "Total Revenue: $1,000,000 (Page 14, Revenue Section)."
   - **Confidence Score:** 0.98.

5. **Currency:**
   - **Direct Extraction:** Identify the currency symbol or note.
   - **Coordinates/Snippet Example:** "Currency: USD (Page 2, Financial Overview)."
   - **Confidence Score:** 1.00.

6. **Units:**
   - **Direct Extraction:** Usually indicated at the top of financial tables or notes.
   - **Coordinates/Snippet Example:** "Units: Thousands (Page 2, Financial Overview)."
   - **Confidence Score:** 1.00.
```



```
   - **Coordinates/Snippet Example:** "Currency: USD (Page 2, Financial Overview)."
   - **Confidence Score:** 1.00.

6. **Units:**
   - **Direct Extraction:** Usually indicated at the top of financial tables or notes.
   - **Coordinates/Snippet Example:** "Units: Thousands (Page 2, Financial Overview)."
   - **Confidence Score:** 1.00.

7. **Depreciation and Amortization:**
   - **Direct Extraction:** Find "Depreciation" and "Amortization" in notes or cash flow statement.
   - **Coordinates/Snippet Example:** "Depreciation: $50,000 (Page 18, Notes to Financial Statements)."
   - **Confidence Score:** 0.90.

8. **Filing Publish Date:**
   - **Direct Extraction:** Typically found on the cover page or introductory section.
   - **Coordinates/Snippet Example:** "Publish Date: 04-15-2023 (Page 1, Cover Page)."
   - **Confidence Score:** 1.00.

9. **Fiscal Year End Date:**
   - **Direct Extraction:** Found in the financial summary or notes.
   - **Coordinates/Snippet Example:** "Fiscal Year End: 12-31-2022 (Page 2, Financial Overview)."
   - **Confidence Score:** 1.00.

10. **Filing Type:**
    - **Direct Extraction:** Check for "Consolidated" or "Standalone" in the document title or sections.
    - **Coordinates/Snippet Example:** "Filing Type: Consolidated (Page 1, Title Section)."
```

# CONCLUSION AND NEXT STEPS

- Transforming Financial Analysis: Our RAG-powered system offers a significant advancement in financial data analysis by automating a traditionally manual and error-prone process.
- Key Benefits:Reduced Costs: By automating data extraction, companies can significantly reduce labor costs.
- Improved Decision-Making: Accurate and readily available financial data enables better informed business decisions.
- Enhanced Compliance: Transparent and auditable data extraction ensures compliance with regulatory requirements.
- Next Steps:Pilot program with select financial institutions to gather real-world feedback and refine the system.
- Development of a user-friendly API for seamless integration with other applications.
- Research into advanced NLP techniques for even more accurate and nuanced data extraction.