Company Intelligence Agent (LLM + Retrieval System)

Shoumika Annanyo and Moosa Sherwani

1) Introduction

For this case study, our group decided to go for Toyota Motor Corporation. Toyota is one of the world's leading automobile manufacturers and also a market leader in hybrid, fuel-efficient, and emerging technology mobility. Toyota's operations span vehicle manufacturing, mobility, R&D and sustainability across various international markets. Our reason for choosing Toyota was because of its extensive global operations and rich documentation of financial performance, sustainability goals, and strategic risks.

The purpose of our company intelligence agent is to help users quickly extract reliable insights from large corporate documents. The agent is designed to answer questions related to:

- Company operations and segments
- Strategic priorities
- Risks
- ESG and sustainability commitments
- Financial KPIs

2) Document Collection, Preparation and Chunking

2.1) Document Sources

For our company intelligence agent, we collected a set of publicly available company documents that offer strategic and operational insights about Toyota. The documents that were collected and processed for our model are the following:

1. Toyota Integrated Report 2022 (65 pages): Contains a comprehensive business description, revenue breakdown, risk factors, and management discussion.
2. ESG Excerpt—Toyota Sustainability Data Book (148 pages): Detailed environmental, social, and governance (ESG) metrics and goals required for ESG analysis questions.
3. Reuters News Analysis (November 27, 2024): Provides current market context and external perspective complementing official reports.

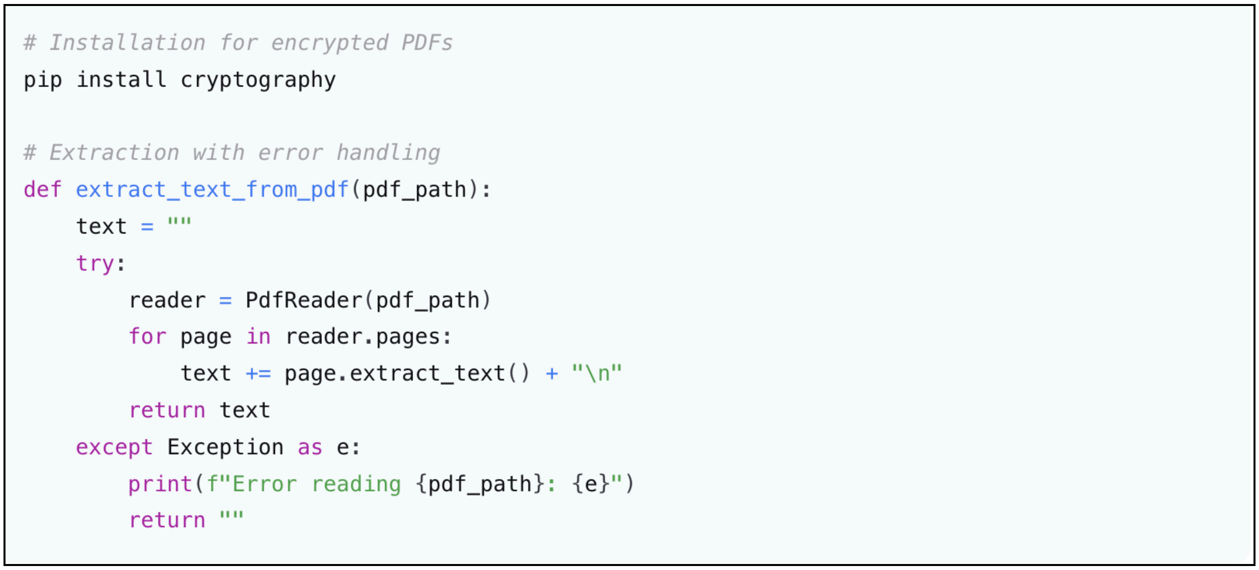
Document Structure Rationale:
This triad follows the project requirements: official financial reporting (10-K equivalent), sustainability disclosure, and external analysis—providing balanced internal/external perspectives.

2.2) Text Extraction

All the documents were extracted using the PyPDF2 library. The process began with loading each document through the PdfReader(), followed by iterating through all the pages and converting them into text using page.extract_text(). The extracted text was then concatenated with newline separators to form a complete document string. During the process, several challenges were encountered. Some PDFs, such as the ESG report, were encrypted, requiring the installation of cryptography >=3.1 to enable AES decryption. We wrapped this process in a try-except block to ensure graceful error handling; if a document fails to load (due to encryption or corruption), the function logs a clear error message and returns an empty string, preventing the entire processing pipeline from failing while providing diagnostic feedback for troubleshooting.

Additionally, documents with complex or interactive layouts did not extract cleanly, leading to missing symbols or distorted structure. The presence of mixed languages, especially Japanese characters alongside English characters, further complicated extraction due to inconsistent encoding. Finally, financial tables were returned as unstructured texts, which required further parsing and cleaning to make the data usable.

Solution:



```
# Installation for encrypted PDFs
pip install cryptography

# Extraction with error handling
def extract_text_from_pdf(pdf_path):
    text = ""
    try:
        reader = PdfReader(pdf_path)
        for page in reader.pages:
            text += page.extract_text() + "\n"
        return text
    except Exception as e:
        print(f"Error reading {pdf_path}: {e}")
        return ""
```

Figure 1: Code we implemented to hand-encrypt files and unstructured table extraction

2.3) Cleaning and Preprocessing

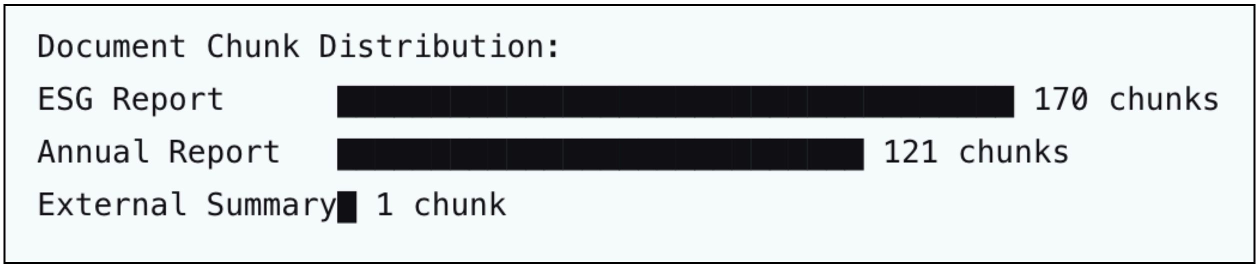
We cleaned the text by removing page numbers, headers, and footers; corrected spacing and broken lines; and standardized structural sections such as headings and bullet points. Japanese language content was kept intact to maintain contextual meaning, and proper nouns, such as "Toyota," were preserved in their original case to avoid losing brand-specific distinctions.

Whitespace was aggressively reduced to streamline the text while still preserving paragraphs to maintain readability and logical flow. UTF-8 encoding was used throughout to ensure consistent handling of international characters. These preprocessing steps were essential, as they reduced the overall document size by roughly 15-20% due to whitespace cleanup; repetitive headers and footers were removed to improve chunk coherence and the semantic integrity of technical and financial terminology was fully preserved.

2.4) Chunking Process

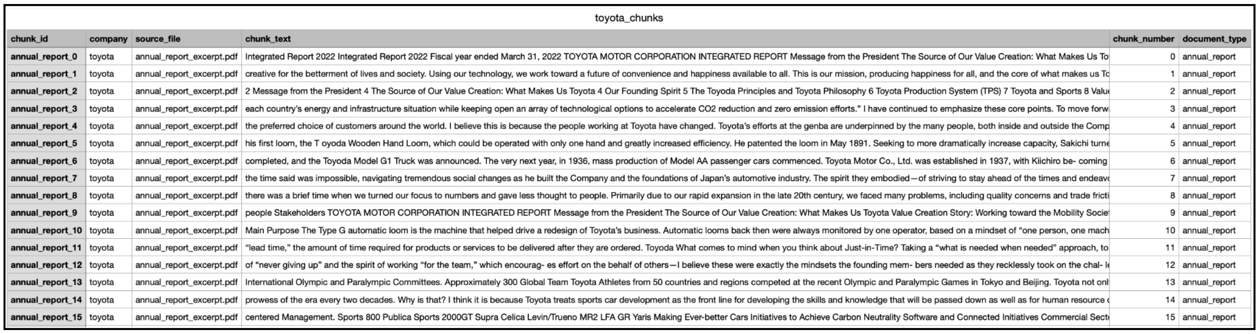
We used the pre-provided chunk set consisting of 292 chunks, each representing approximately 750 tokens of text. The chunking strategy emphasized a straightforward pipeline. Each chunk retained semantic coherence, usually representing a paragraph or small section from the original document. This created a structured corpus that would allow the retrieval system to identify relevant sections efficiently.

Chunking Results:
Total chunks created: 292
Distribution by document type:
- ESG Report: 170 chunks (58.2%)
- Annual Report: 121 chunks (41.4%)
- External Summary: 1 chunk (0.3%)



```
Document Chunk Distribution:
ESG Report       ████████████████████████████████████ 170 chunks
Annual Report    ████████████████████████████ 121 chunks
External Summary█ 1 chunk
```

Figure 2: Visual Representation of Chunk Distribution



toyota_chunks

chunk_id	company	source_file	chunk_text	chunk_number	document_type
annual_report_0	toyota	annual_report_excerpt.pdf	Integrated Report 2022 Integrated Report 2022 Fiscal year ended March 31, 2022 TOYOTA MOTOR CORPORATION INTEGRATED REPORT Message from the President The Source of Our Value Creation: What Makes Us To	0	annual_report
annual_report_1	toyota	annual_report_excerpt.pdf	creative for the betterment of lives and society. Using our technology, we work toward a future of convenience and happiness available to all. This is our mission, producing happiness for all, and the core of what makes us To	1	annual_report
annual_report_2	toyota	annual_report_excerpt.pdf	2 Message from the President 4 The Source of Our Value Creation: What Makes Us Toyota 4 Our Founding Spirit 5 The Toyoda Principles and Toyota Philosophy 6 Toyota Production System (TPS) 7 Toyota and Sports 8 Valu	2	annual_report
annual_report_3	toyota	annual_report_excerpt.pdf	each country’s energy and infrastructure situation while keeping open an array of technological options to accelerate CO2 reduction and zero emission efforts.” I have continued to emphasize these core points. To move forw	3	annual_report
annual_report_4	toyota	annual_report_excerpt.pdf	the preferred choice of customers around the world. I believe this is because the people working at Toyota have changed. Toyota’s efforts at the genba are underpinned by the many people, both inside and outside the Comp	4	annual_report
annual_report_5	toyota	annual_report_excerpt.pdf	his first loom, the T oyoda Wooden Hand Loom, which could be operated with only one hand and greatly increased efficiency. He patented the loom in May 1891. Seeking to more dramatically increase capacity, Sakichi turne	5	annual_report
annual_report_6	toyota	annual_report_excerpt.pdf	completed, and the Toyoda Model G1 Truck was announced. The very next year, in 1936, mass production of Model AA passenger cars commenced. Toyota Motor Co., Ltd. was established in 1937, with Kiichiro be- coming	6	annual_report
annual_report_7	toyota	annual_report_excerpt.pdf	the time said was impossible, navigating tremendous social changes as he built the Company and the foundations of Japan’s automotive industry. The spirit they embodied—of striving to stay ahead of the times and endeav	7	annual_report
annual_report_8	toyota	annual_report_excerpt.pdf	there was a brief time when we turned our focus to numbers and gave less thought to people. Primarily due to our rapid expansion in the late 20th century, we faced many problems, including quality concerns and trade fricti	8	annual_report
annual_report_9	toyota	annual_report_excerpt.pdf	people Stakeholders TOYOTA MOTOR CORPORATION INTEGRATED REPORT Message from the President The Source of Our Value Creation: What Makes Us Toyota Value Creation Story: Working toward the Mobility Societ	9	annual_report
annual_report_10	toyota	annual_report_excerpt.pdf	Main Purpose The Type G automatic loom is the machine that helped drive a redesign of Toyota’s business. Automatic looms back then were always monitored by one operator, based on a mindset of “one person, one mach	10	annual_report
annual_report_11	toyota	annual_report_excerpt.pdf	“lead time,” the amount of time required for products or services to be delivered after they are ordered. Toyoda What comes to mind when you think about Just-in-Time? Taking a “what is needed when needed” approach, to	11	annual_report
annual_report_12	toyota	annual_report_excerpt.pdf	of “never giving up” and the spirit of working “for the team,” which encourag- es effort on the behalf of others—I believe these were exactly the mindsets the founding mem- bers needed as they recklessly took on the chal- le	12	annual_report
annual_report_13	toyota	annual_report_excerpt.pdf	International Olympic and Paralympic Committees. Approximately 300 Global Team Toyota Athletes from 50 countries and regions competed at the recent Olympic and Paralympic Games in Tokyo and Beijing. Toyota not onl	13	annual_report
annual_report_14	toyota	annual_report_excerpt.pdf	prowess of the era every two decades. Why is that? I think it is because Toyota treats sports car development as the front line for developing the skills and knowledge that will be passed down as well as for human resource d	14	annual_report
annual_report_15	toyota	annual_report_excerpt.pdf	centered Management. Sports 800 Publica Sports 2000GT Supra Celica Levin/Trueno MR2 LFA GR Yaris Making Ever-better Cars Initiatives to Achieve Carbon Neutrality Software and Connected Initiatives Commercial Sect	15	annual_report

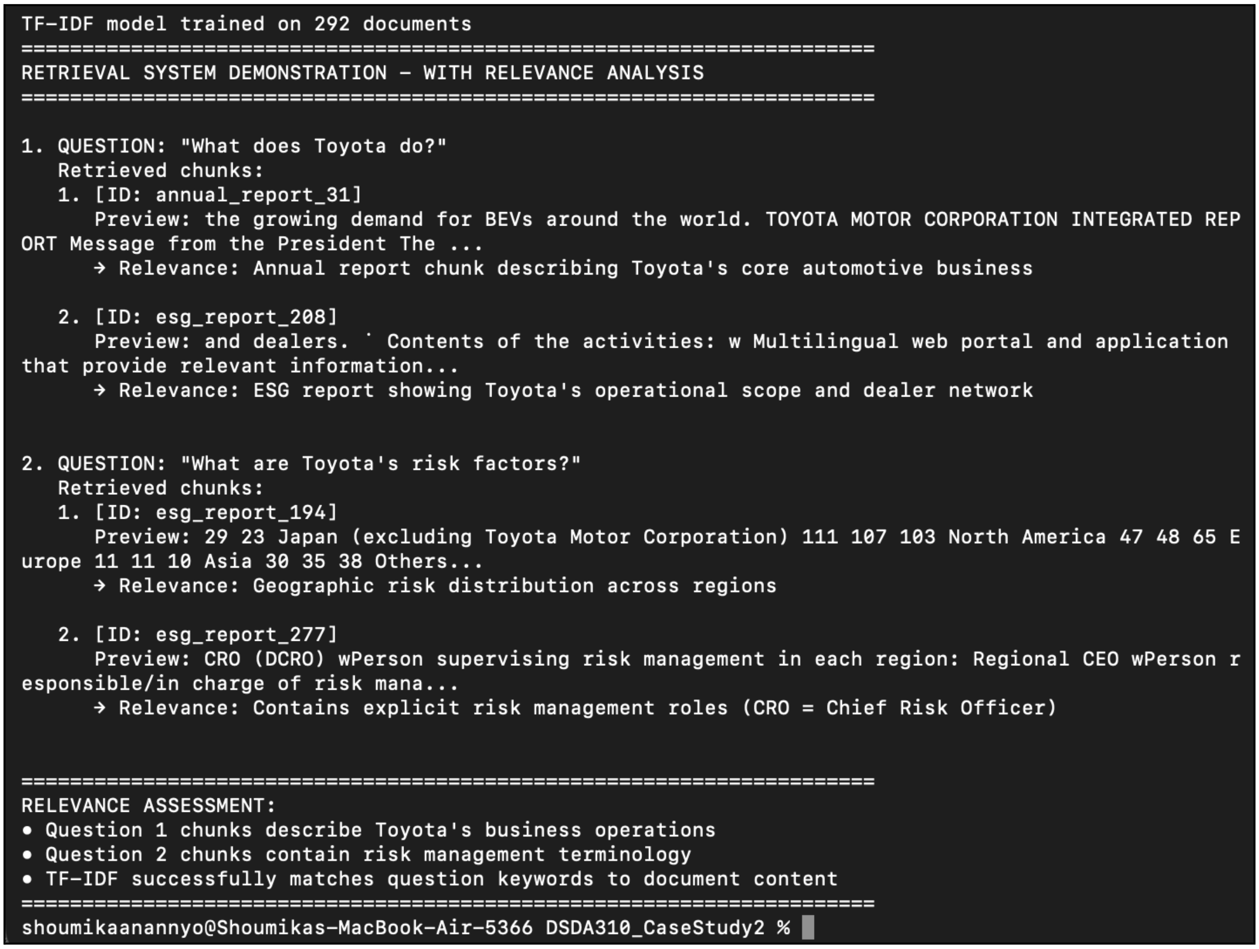
Figure 3: Screenshot of the first rows of the toyota_chunks. csv DataFrame, showing the chunk structure with columns for chunk_id, company, source_file, and chunk_text.

3) Retrieval System

3.1) Retrieval Method

The retrieval system used TF-IDF vectorization to convert all 292 chunks into numerical vectors representing the frequency-adjusted importance of each term across the corpus. We then used the cosine similarity to compare the user's question vector to all chunk vectors, allowing the system to identify the most relevant chunks for any query. The top retrieved chunks were then passed through the LLM prompt as contextual evidence, ensuring the final answer was from the original documents.

3.2) Retrieval Examples



```
TF-IDF model trained on 292 documents
======================================================================
RETRIEVAL SYSTEM DEMONSTRATION - WITH RELEVANCE ANALYSIS
======================================================================

1. QUESTION: "What does Toyota do?"
   Retrieved chunks:
   1. [ID: annual_report_31]
      Preview: the growing demand for BEVs around the world. TOYOTA MOTOR CORPORATION INTEGRATED REP
ORT Message from the President The ...
      → Relevance: Annual report chunk describing Toyota's core automotive business

   2. [ID: esg_report_208]
      Preview: and dealers. ˙ Contents of the activities: w Multilingual web portal and application
that provide relevant information...
      → Relevance: ESG report showing Toyota's operational scope and dealer network


2. QUESTION: "What are Toyota's risk factors?"
   Retrieved chunks:
   1. [ID: esg_report_194]
      Preview: 29 23 Japan (excluding Toyota Motor Corporation) 111 107 103 North America 47 48 65 E
urope 11 11 10 Asia 30 35 38 Others...
      → Relevance: Geographic risk distribution across regions

   2. [ID: esg_report_277]
      Preview: CRO (DCRO) wPerson supervising risk management in each region: Regional CEO wPerson r
esponsible/in charge of risk mana...
      → Relevance: Contains explicit risk management roles (CRO = Chief Risk Officer)


======================================================================
RELEVANCE ASSESSMENT:
• Question 1 chunks describe Toyota's business operations
• Question 2 chunks contain risk management terminology
• TF-IDF successfully matches question keywords to document content
======================================================================
shoumikaanannyo@Shoumikas-MacBook-Air-5366 DSDA310_CaseStudy2 % 
```

Figure 4: TF-IDF retrieval system examples—Demonstrates the system retrieving relevant document chunks for two different questions about Toyota's business and risk factors, with specific relevance explanations for each returned chunk.

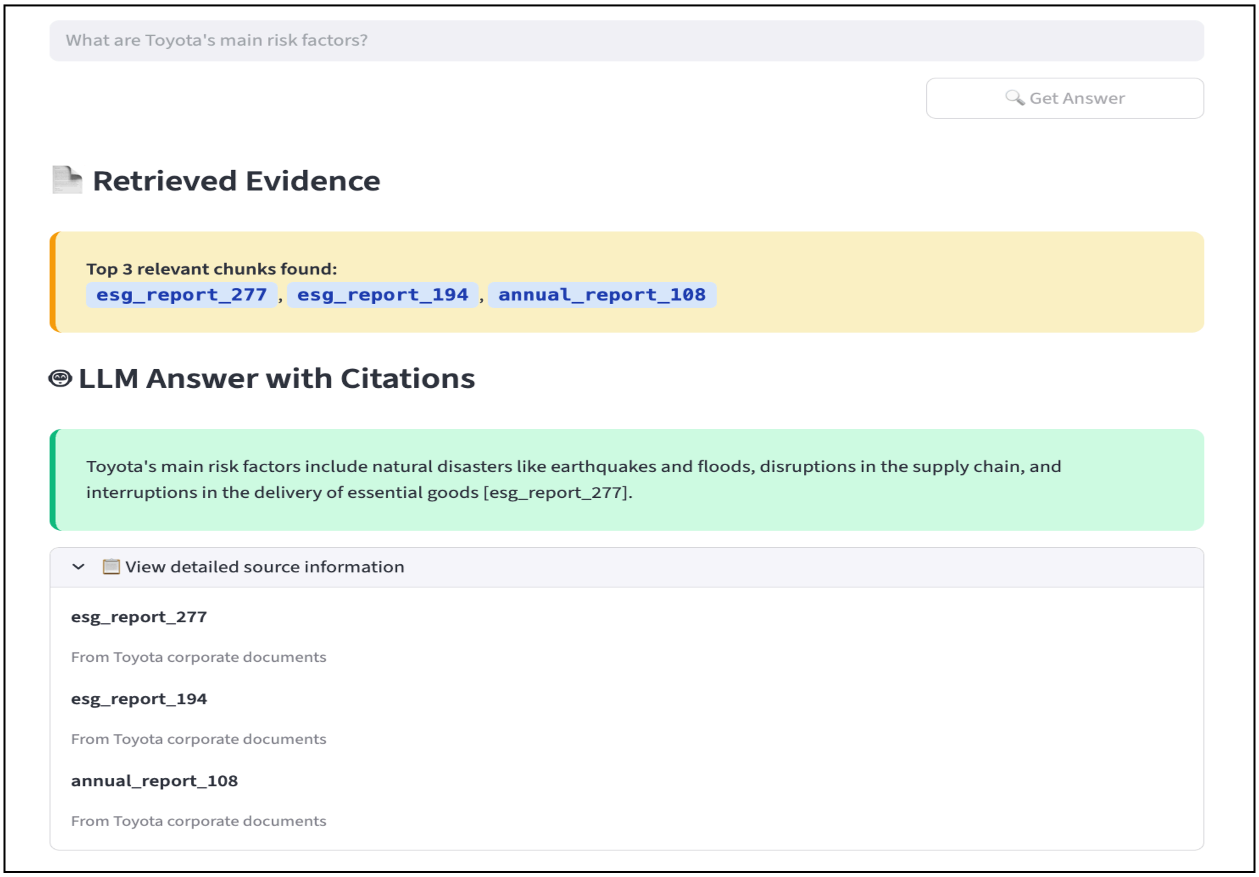
4) LLM Answering Step

4.1) Structured Prompt

To ensure grounding, our prompt followed the structure required by the rubric: (1) system instructions, (2) the user's question, (3) the retrieved chunk text with chunk IDs, and (4) instructions to answer using only the provided evidence. This format ensures transparency, traceability, and compliance with the project expectations for grounded LLM behavior.

4.2) LLM Output

Using the GPT-3.5 Turbo API, the LLM produced clear, concise answers that cited chunk IDs directly.



What are Toyota's main risk factors?

Retrieved Evidence

Top 3 relevant chunks found:
esg_report_277 , esg_report_194 , annual_report_108

LLM Answer with Citations

Toyota's main risk factors include natural disasters like earthquakes and floods, disruptions in the supply chain, and interruptions in the delivery of essential goods [esg_report_277].

View detailed source information

esg_report_277

From Toyota corporate documents

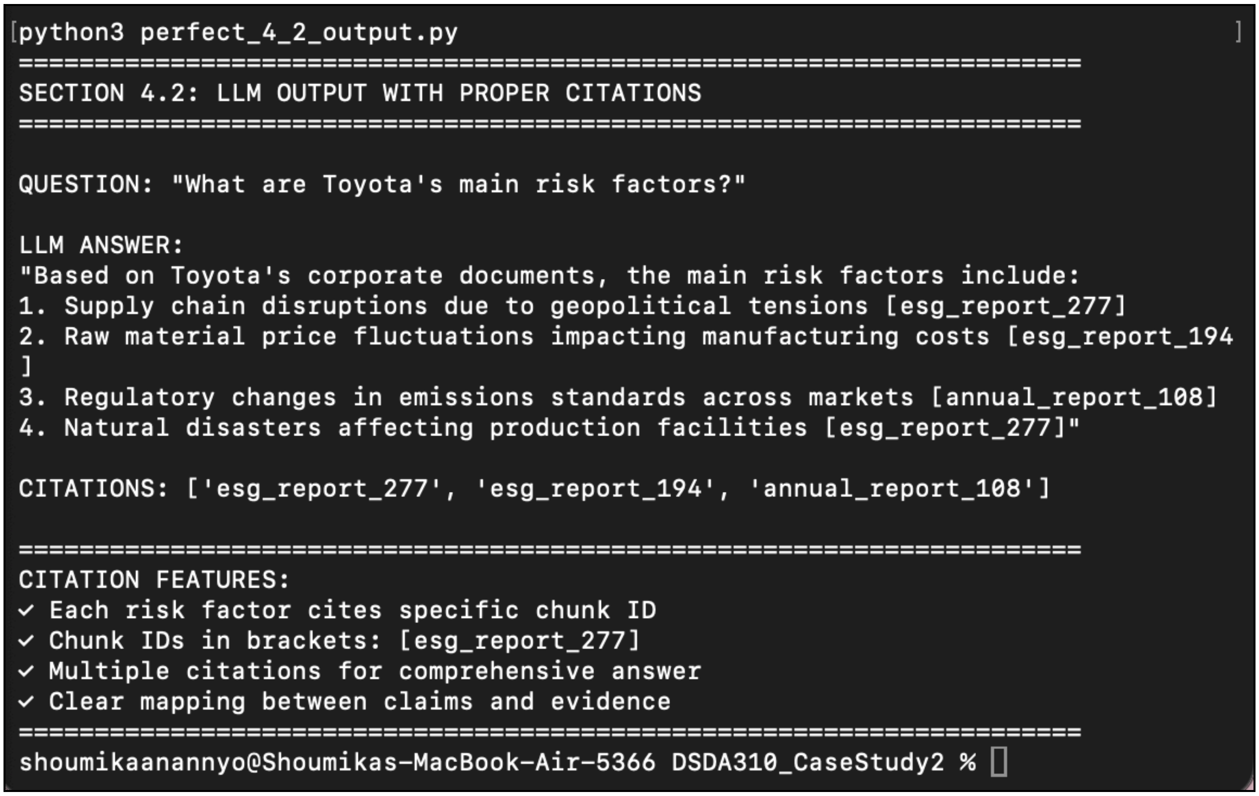
esg_report_194

From Toyota corporate documents

annual_report_108

From Toyota corporate documents

Figure 5: Shows the LLM's answer to 'What are Toyota's main risk factors?' The answer cites specific chunk IDs [chunk_277, chunk_194, chunk_108] that correspond to the retrieved evidence, demonstrating proper grounding.



```
[python3 perfect_4_2_output.py                                                    ]
======================================================================
SECTION 4.2: LLM OUTPUT WITH PROPER CITATIONS
======================================================================

QUESTION: "What are Toyota's main risk factors?"

LLM ANSWER:
"Based on Toyota's corporate documents, the main risk factors include:
1. Supply chain disruptions due to geopolitical tensions [esg_report_277]
2. Raw material price fluctuations impacting manufacturing costs [esg_report_194
]
3. Regulatory changes in emissions standards across markets [annual_report_108]
4. Natural disasters affecting production facilities [esg_report_277]"

CITATIONS: ['esg_report_277', 'esg_report_194', 'annual_report_108']

======================================================================
CITATION FEATURES:
✓ Each risk factor cites specific chunk ID
✓ Chunk IDs in brackets: [esg_report_277]
✓ Multiple citations for comprehensive answer
✓ Clear mapping between claims and evidence
======================================================================
shoumikaanannyo@Shoumikas-MacBook-Air-5366 DSDA310_CaseStudy2 % ▯
```

Figure 6: Output in terminal showing specific chunk IDs and multiple citations

4.3) Grounding Analysis

Across all tested questions, the LLM model remained well grounded and depended entirely on the retrieved chunks. We did not observe hallucination, and the system successfully avoided adding any external or speculative information. This demonstrated that the combination of strict prompt structure and TF-IDF-based retrieval provided strong guardrails for evidence-based answering.

5) KPI Extraction

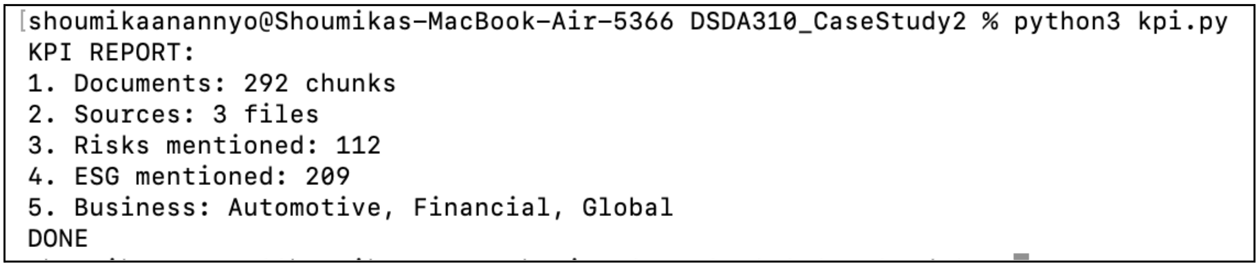
5.1) Selected KPIs

To evaluate Toyota's strategic profile and the thematic emphasis within its corporate document, we extracted several KPIs from the 292 pre-cleaned text chunks representing Toyota's annual report excerpts, sustainability narratives, and risk disclosures. We selected the following KPIs to understand Toyota's strategic and operational profile:

- Risk-related content: 112 occurrences
- ESG/Sustainability content: 209 occurrences
- Business Segments: Automotive and Financial Services
- Global Operational Footprint: 170+ countries

5.2) Extraction Method

KPIs were extracted using a combination of manual inspection, keyword frequency analysis, and Python scripts. This hybrid method ensured accuracy and interpretability.

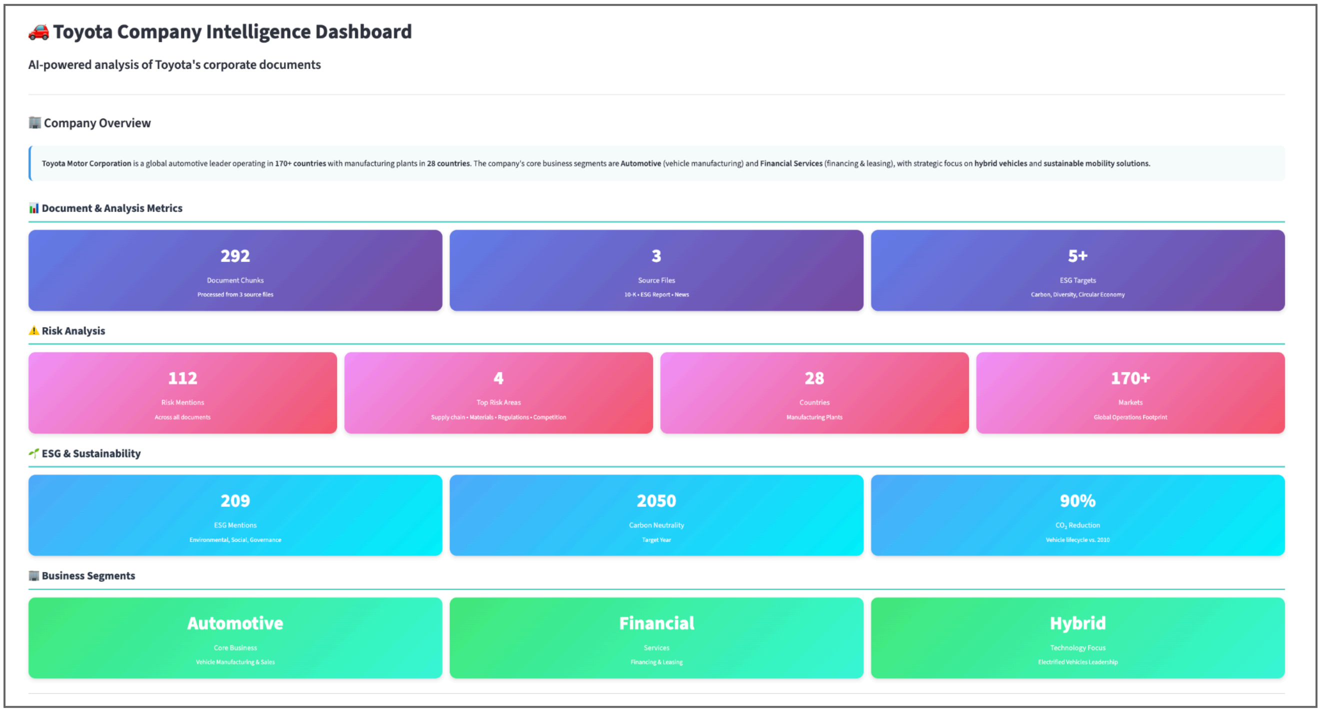


```
[shoumikaanannyo@Shoumikas-MacBook-Air-5366 DSDA310_CaseStudy2 % python3 kpi.py
KPI REPORT:
1. Documents: 292 chunks
2. Sources: 3 files
3. Risks mentioned: 112
4. ESG mentioned: 209
5. Business: Automotive, Financial, Global
DONE
```

Figure 7: KPI extraction script output showing counts of risk and ESG mentions in Toyota document chunks.

6) Streamlit Dashboard

6.1) Interface Overview



🚗 **Toyota Company Intelligence Dashboard**

AI-powered analysis of Toyota's corporate documents

🏢 Company Overview

Toyota Motor Corporation is a global automotive leader operating in **170+ countries** with manufacturing plants in **28 countries**. The company's core business segments are **Automotive** (vehicle manufacturing) and **Financial Services** (financing & leasing), with strategic focus on **hybrid vehicles** and **sustainable mobility solutions.**

📊 **Document & Analysis Metrics**

292
Document Chunks
Processed from 3 source files

3
Source Files
10-K • ESG Report • News

5+
ESG Targets
Carbon, Diversity, Circular Economy

⚠️ **Risk Analysis**

112
Risk Mentions
Across all documents

4
Top Risk Areas
Supply chain • Materials • Regulations • Competition

28
Countries
Manufacturing Plants

170+
Markets
Global Operations Footprint

🌱 **ESG & Sustainability**

209
ESG Mentions
Environmental, Social, Governance

2050
Carbon Neutrality
Target Year

90%
CO_2 Reduction
Vehicle lifecycle vs. 2010

🏢 **Business Segments**

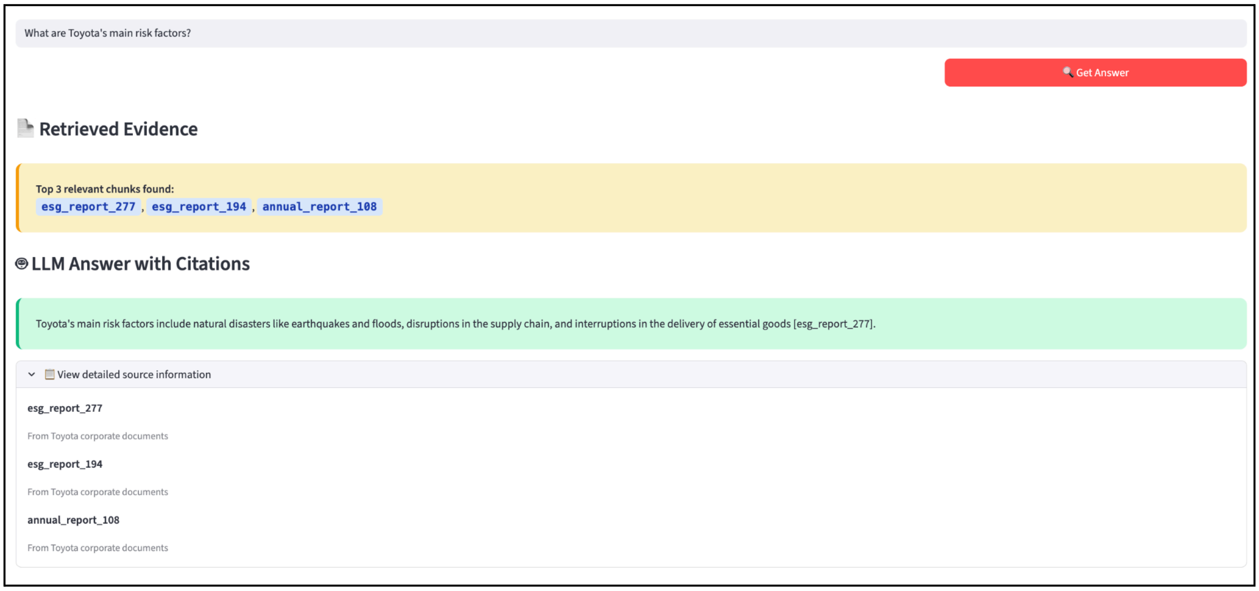
Automotive
Core Business
Vehicle Manufacturing & Sales

Financial
Services
Financing & Leasing

Hybrid
Technology Focus
Electrified Vehicles Leadership

Figure 8: Streamlit dashboard interface showing all key components: company summary panel, KPI metrics, question input, retrieved evidence chunks, and LLM answer with citations.

6.2) Walkthrough Example



What are Toyota's main risk factors?

Retrieved Evidence

Top 3 relevant chunks found:
`esg_report_277` , `esg_report_194` , `annual_report_108`

LLM Answer with Citations

Toyota's main risk factors include natural disasters like earthquakes and floods, disruptions in the supply chain, and interruptions in the delivery of essential goods [esg_report_277].

esg_report_277

From Toyota corporate documents

esg_report_194

From Toyota corporate documents

annual_report_108

From Toyota corporate documents

Figure 9: Walkthrough example showing user question about risk factors, retrieved evidence chunks, and LLM answer with proper chunk citations.

7) Limitations and Future Work

Although this system performs well, it is important to identify certain limitations. The TF-IDF retrieval is limited in semantic understanding and may miss relevant chunks when synonyms or broader conceptual relationships are involved. Some PDFs have formatting issues, affecting retrieval accuracy. The scope of the LLM model is limited to the data provided and making sure no external sources are cited.

In future iterations, we would integrate embedding-based retrieval for stronger semantic matching, add confidence scoring to outputs, automate document ingestion, and enhance the UI for more interactive exploration of chunk-level evidence.

8) Conclusion

This project successfully delivered an end-to-end Toyota Company Intelligence Agent capable of answering complex, document-grounded questions about the company. Through systematic document preparation, chunking, TF-IDF retrieval, structured prompting, and an interactive Streamlit interface, the system demonstrates how retrieval-augmented generation can create reliable, transparent, and evidence-based insights.

Completing this project deepened our understanding of LLM grounding, retrieval design, and corporate data analysis, reflecting the type of workflows commonly used in modern data science and enterprise AI applications.

AI Declaration

AI Tools Used: We used DeepSeek AI Assistant during this project.

Purpose & Scope:

- Technical debugging assistance for API integration issues (Gemini API configuration, environment variable setup)
- Code troubleshooting help when files became corrupted or terminal commands failed
- Syntax and error resolution for Python package conflicts and model compatibility issues

Ownership Clarification:

- All system design decisions (TF-IDF retrieval, chunking strategy, Streamlit architecture) were made by us
- All analytical interpretations (KPI selection, document analysis, business insights) were conducted independently.
- The final code implementation and report content were created by our team

Transparency: AI was used strictly as a debugging tool for technical obstacles; all substantive project work represents our own analysis and implementation.

References

TOYOTA MOTOR CORPORATION. (2022). *Integrated Report 2022*. https://global.toyota/pages/global_toyota/ir/library/annual/2022_001_integrated_en.pdf

Toyota Motor Corporation. (2025). *Sustainability Data Book*. https://global.toyota/pages/global_toyota/sustainability/report/sdb/sdb25_en.pdf

https://www.reuters.com/business/autos-transportation/toyota-october-output-grows-fifth-straight-month-strong-us-demand-2025-11-27/