📜 Finance Laws Retrieval-Augmented Generation (RAG) System

👩‍💻 Authors:

Ginosca Alejandro Dávila
Natanael Santiago Morales

🏫Ironhack Bootcamp:
Data Science and Machine Learning

📅 Date: February 7, 2025

Project Overview

Why do we need this system?

- Financial laws and regulations are **complex** and **voluminous** 📚.
- Traditional search methods are **inefficient** and **time-consuming** ⏳.
- A **Retrieval-Augmented Generation (RAG) System** integrates **AI-driven retrieval & LLM-based responses** to provide **accurate, contextualized answers**.

Objective

✔ Develop a **RAG system** specialized in finance law.
✔ Compare **embedding models** for legal text retrieval.
✔ Evaluate the system based on **retrieval accuracy & response relevance**.

Dataset Description

🔍 Dataset

- **Source:** Collection of finance laws and regulations (EU Directives & Regulations)
- **Format:** PDF files
- **Data Challenges:**
  - Legal documents have complex structures (preambles, annexes, references).
  - Extracting text while maintaining legal references requires careful preprocessing.

Exploratory Data Analysis (EDA)

Dataset Exploration

✔ **Checked folder structure** and file organization 📁.
✔ **Reviewed metadata**: Word count, author, creation date 📊.
✔ **Identified formatting inconsistencies** for preprocessing.

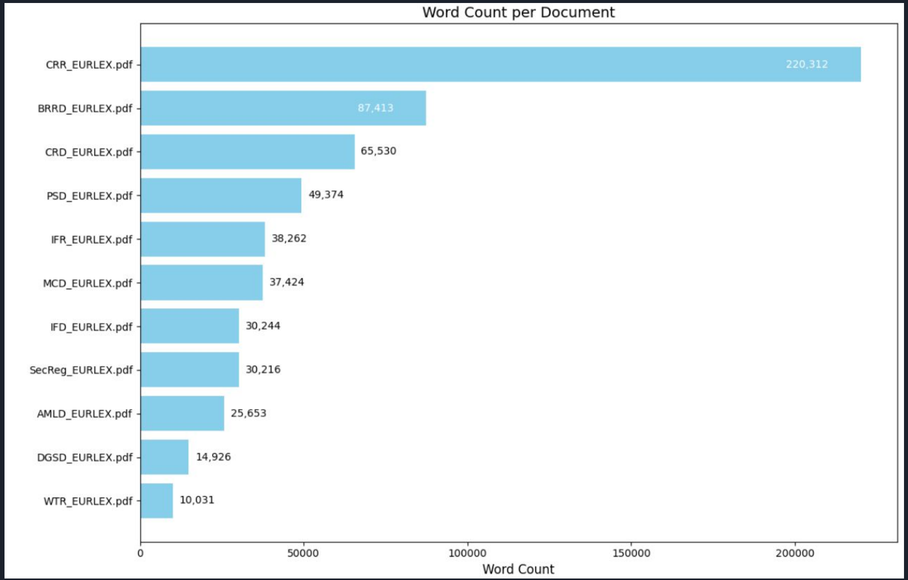
📌 **Key Findings:**

- Documents range from **10,000 to 220,000 words** 📏.
- Headers, footers, and legal references need special handling.
- Some documents have **missing titles**.

Word Count Distribution

📊 Understanding Our Legal Dataset

- Dataset contains **finance laws & regulations** with varying document lengths.
- **Largest document**: *CRR_EURLEX.pdf* (**220,312 words**) 📖
- **Smallest document**: *WTR_EURLEX.pdf* (**10,031 words**) 📝



Text Preprocessing for Legal Documents

📌 **Why do we need preprocessing?**

- Legal documents contain scattered formatting issues 📝.
- Text extraction caused **split words, missing spaces, and broken sentences**.
- Cleaning **improves retrieval accuracy & LLM performance**.

✔ **Steps Taken:**
1️⃣ **Removed unnecessary line breaks**, replacing them with spaces.
2️⃣ **Merged split words** (e.g., "financia l system" → "financial system").
3️⃣ **Applied spell-checking** to correct fragmented text.
4️⃣ **Standardized punctuation & spacing** for readability.
5️⃣ **Ensured legal references remain intact** (e.g., "Article 5, Directive 2005/60/EC").

📌 **Before vs. After Cleaning Example:**
"THE EUR OPEAN P ARLIAMENT AND THE COUNC IL" →
"THE EUROPEAN PARLIAMENT AND THE COUNCIL"

💡 **Impact:** Cleaner text **improves embeddings & retrieval accuracy**.

Text Chunking Strategy

✔ **Why do we need chunking?**

- Legal texts **vary in length & structure**.
- Large texts **must be split** for **efficient retrieval**.

📌 **Chosen Approach: `CharacterTextSplitter`**
✔ **Preserves context** with **200-character overlap**.
✔ Prevents **fragmentation**, keeping **meaningful text chunks**.
✔ Ensures **consistent structure** for embeddings & retrieval.

Chucking: Implementation and Results

- 🔹 **Processing Steps:**

1️⃣ **Loaded PDF files** & extracted text.
2️⃣ **Split pages into manageable text chunks**.
3️⃣ **Preserved metadata** (document name, page numbers).
4️⃣ **Stored structured text chuck** for embedding & retrieval in ChromaDB.

📊 **Result**: **1,561** document chunks ready for embedding!

Embedding Models for Retrieval

📌 **Why do we need embeddings?**

- Converts text into **numerical vectors** for **efficient similarity search**.

✔ **Models Used:**
1️⃣ **OpenAI's `text-embedding-3-large`** → Commercial, high-accuracy 🔥
2️⃣ **Hugging Face `all-MiniLM-L6-v2`** → Open-source, efficient ⚡

📊 **Comparison Strategy**:
✔ Stored embeddings in **ChromaDB** for fast retrieval.
✔ Evaluated **retrieval accuracy & response quality**.

Model Training & Embedding Process

📌 **Steps for creating embeddings:**

1. Preprocessed legal text chunks → Converted into embeddings.
2. Used ChromaDB as vector storage for fast retrieval.
3. Tested similarity search with different embedding models.

🛠️ **Key Challenges:** Managing computational cost vs. retrieval accuracy.
🔍 **Optimization:** Reduced embedding size to improve efficiency.

Legal Query Retrieval & Testing

📌 **How do we retrieve relevant legal documents?**

1️⃣**User query** → Convert to an embedding.

2️⃣**Compare against stored document embeddings.**

3️⃣**Retrieve top 5 most relevant chunks.**

4️⃣**Generate a response** using GPT-4 based on retrieved context.

📝 **Example Query:**

"What are the key points of financial legislation in Europe?"

Sample Question and Responses

In what type of assests shoud DGSs be invested? (Q4)

“(35) DGSs should invest in low-risk assets.” DGSD EURLEX, Page 5

Types of Assets for DGS Investments

Deposit Guarantee Schemes (DGSs) should primarily invest in cash, deposits, payment commitments, and low-risk assets that can be liquidated within a short period of time. It is necessary that these assets are of low risk to ensure adequate financial means are available for fulfilling their obligations (DGSD, Article 10).

Additionally, DGSs must have adequate alternative funding arrangements in place to obtain short-term funding to meet claims made against them (DGSD, Article 10).

Source:

- From `DGSD_EURLEX.pdf`, page 5 of the European Union Finance Laws, from the Official Journal of the European Union.
- From `DGSD_EURLEX.pdf`, page 19 of the European Union Finance Laws, from the Official Journal of the European Union.

Hugging Face Embeddings

Investment Assets for DGSs

Deposit Guarantee Schemes (DGSs) should primarily focus on assets that enhance their financial stability and ability to meet their obligations. The regulations indicate that DGSs are expected to be financially sound and transparent in their operations and governance. While the specific types of assets for investment are not directly enumerated in the provided context, it is implied that DGSs should use available financial means in a manner that ensures they can repay depositors and meet claims effectively.

DGSs are allowed to use their financial means primarily for repaying depositors as mandated under Article 11. Additionally, they may finance the resolution of credit institutions under specific regulatory frameworks. The investment strategy should be aligned with maintaining adequate liquidity and managing risks effectively (source: DGSD_EURLEX.pdf, page 11, 18).

Furthermore, DGSs must ensure that their available financial means are sufficient to cover potential liabilities, as outlined in Article 10. This necessitates investing in assets that are stable and liquid, allowing for quick access when claims arise.

Source:

- From `DGSD_EURLEX.pdf`, page `11` of the European Union Finance Laws, from the Official Journal of the European Union
- From `DGSD_EURLEX.pdf`, page `18` of the European Union Finance Laws, from the Official Journal of the European Union

OpenAI Embeddings

Evaluation

In what type of assests shoud DGSs be invested? (Q4)

	Hugging Face	OpenAI
Correctness	- Accurate. - Specifies the required investment in low-risk assets - Includes the necessity for alternative funding arrangements.	- Correct in its general assessment - Doesn't mention specific types of low-risk assets
Completeness	- Complete. - Mentions the key assets DGSs should invest in.	- Broader view of DGSs' obligations - Omits the specific types of assets
Citation	- References page 5.	- References page 11 and page 18, which are different from the expected page (5)
Strengths	- More specific, mentioning the exact types of assets	- Broader context of DGSs' financial responsibilities
Weaknesses		- More general and did not provide specific types of low-risk assets

Evaluation

Hugging Face

- Outperformed the OpenAI response in both accuracy and specificity.
- Included the specific asset types but also provided a more relevant citation.
- Scored 0.9225

OpenAI

- Correct in general
- Lacked the precise details required to fully answer the question.
- Scores 0.735

Retrieval Performance Metrics

📌 **Evaluation Criteria for Legal Text Retrieval**

✔ **Correctness (40%)** – Does the response accurately reflect the legal document? 📖
✔ **Completeness (30%)** – Does it fully answer the legal question? ✅
✔ **Conciseness (10%)** – Is the response clear and to the point? ✍
✔ **Relevance (10%)** – Does it stay on topic? 🎯
✔ **Citation Accuracy (5%)** – Are the correct pages cited? 📑
✔ **Language Quality (5%)** – Is the legal language professional and clear? 💡

💡 **Why This Matters?**

🔹 Ensures responses are legally reliable.
🔹 Helps determine the best embedding model for finance laws.

Performance Comparison: Hugging Face vs. OpenAI Across Legal Queries

Metric	Weight	Q1 Hugging Face	Q1 OpenAI	Q2 Hugging Face	Q2 OpenAI	Q3 Hugging Face	Q3 OpenAI	Q4 Hugging Face	Q4 OpenAI
Correctness	0.4	0.9	0.8	0.9	0.9	0.85	0.9	0.95	0.75
Completeness	0.3	0.95	0.85	0.85	0.95	0.9	0.85	0.9	0.7
Conciseness	0.1	0.9	0.7	0.9	0.75	0.8	0.75	0.9	0.75
Relevance	0.1	0.9	0.75	0.9	0.85	0.9	0.85	0.9	0.75
Citation Accuracy	0.05	0.2	0.1	0.9	0.95	0.6	0.85	0.95	0.6
Language Quality	0.05	0.95	0.9	0.95	0.9	0.95	0.9	0.9	0.9
Weighted Score	-	0.88	0.78	0.8875	0.8975	0.865	0.86	0.9225	0.735

Key Observations and Insights

🔹 **Hugging Face consistently outperforms OpenAI** in retrieval accuracy and citation precision across all four legal queries.

🔹 **OpenAI embeddings tend to generalize more,** resulting in lower retrieval accuracy and citation precision. While its responses are fluent, they sometimes introduce unrelated legal details.

🔹 **Weighted scores indicate that Hugging Face is the preferred option** for legal queries requiring **high precision and correct citations**.

🔹 **OpenAI performed best in Q2**, offering a more detailed response. However, **Hugging Face still provided more accurate references** to the source material.

💡 **Takeaway:**
For **general finance-related Q&A**, OpenAI may be useful due to its faster response time. However, for **high-stakes legal document retrieval**, Hugging Face embeddings provide superior accuracy and citation reliability. 🚀

Comparative Analysis: Hugging Face vs. OpenAI Retrieval Performance

Metric	**Hugging Face**	**OpenAI**
Detail and Relevance	More specific & retrieved correct pages 📑	More generalized & missed some details 📄
Response Speed	Slightly slower	Faster ⚡
Citation Accuracy	Higher citation precision	Occasionally retrieved incorrect sections ❗
Use Case	Best for legal research requiring precise references	Best for general finance Q&A

💡 **Key Takeaway:** Hugging Face provided **better recall for detailed queries**, while OpenAI **worked faster but was less precise in legal citations.**

Conclusion & Future Enhancements

📌 **Key Takeaways**

✔ **Data Cleaning Improved Retrieval Quality** 🏆
✔ **Chunking Strategy Prevented Context Loss** 📖
✔ **Hugging Face Performed Better for Legal-Specific Queries** 🎯
✔ **OpenAI Was Faster but Less Precise in Citations** ⏳

📌 **Future Work**

🔹 **Fine-tune citation accuracy by adjusting retrieval ranking** 📜
🔹 **Experiment with chunk size variations for better recall** 🏗
🔹 **Test additional embedding models for increased legal precision** 🎯
🔹 **Deploy the system as a web app for real-world testing** 🚀

Future Work

📌 **Final Observations:**
✔ **Data Cleaning Enhanced Retrieval Quality** 🏆.
✔ **Chunking Strategy Prevented Loss of Context** 📖.
✔ **Hugging Face performed better** for **legal research queries** 🎯.
✔ **Future Work**:

- Improve citation accuracy 📜.
- Fine-tune retrieval ranking for **better user experience**.

THANK YOU!

IRON HACK