Project Report
Document Comparison Chatbot for Regulatory Compliance

CMI Summer Internship with Alumni

Date: 30 July 2024

Interns:

Vishal Maurya (vishal.mds2023@cmi.ac.in)

Soumabha Roy (soumabha.ug2023@cmi.ac.in)

Mentors:

Shashi Satyam(shashisatyam7@gmail.com)

Sarvesh Bhandaokar (sarveshb.1999@gmail.com)

1. Introduction

The 'Document Comparison Chatbot for Regulatory Compliance' project focuses on automating the comparison of regulatory documents across different countries. This is particularly crucial for companies in the automotive industry who need to comply with varying regulations when entering new markets. The chatbot accepts two regulatory documents and a user query, then generates answers and provides a comparative analysis, highlighting similarities and differences.

2. Project Overview

Background and Motivation:

- Each country has its own rules and regulations related to vehicles. These rules and regulation must be followed for a legal sale of vehicles.
- Difficulty in comparing documents manually from different countries with varying regulations.

Objective: Develop a chatbot to compare legal documents from different countries based on user queries and generate accurate responses to support regulatory compliance.

Sub-Objectives: To generate precise and contextually accurate responses to user queries.

3. The Why

The main motivations for this project include reducing the time and cost associated with manual document comparison, ensuring accuracy in adherence to regulations, and overcoming language barriers. Manually comparing regulatory documents can take up to 60 hours per set of documents. Automating this process not only saves time but also improves accuracy by reducing human error.

The homologation department, responsible for ensuring products meet international regulations, faces a significant challenge in efficiently comparing regulatory documents across various languages- This project proposes developing a specialized chatbot for the homologation department with a scope for automated document comparison using LLM and translation of various languages in a single source language.

4. Approaches and Challenges

Retrieval-Augmented Generation (RAG)

- **Approach**: Initially, we implemented RAG to generate answers from specific domains or knowledge bases. This involved retrieving relevant text segments and then generating answers based on these segments.

- **Results**: The approach provided some accurate responses but struggled with incomplete answers and missing context.
- **Challenges**: The complex structure of the PDF documents with multiple nested sections and varying formats posed a challenge for the RAG model, making it difficult to maintain context and accuracy.

Page-wise Cumulative Summarization

- **Approach**: To tackle the context maintenance issue, we experimented with cumulative page-wise summarization. Each page summary was generated by considering the current page and the summary of the previous pages.
- **Results**: This method improved context continuity but was highly time-consuming, requiring one LLM call per page.

RAG with Parent Document Retriever

- **Approach**: We employed a parent document retriever approach that utilized smaller chunks for information retrieval and larger chunks for generating answers, balancing the strengths of both chunk sizes.
- **Results**: This technique improved retrieval and generation quality but occasionally broke sections, leading to fragmented context.

Section-wise Chunking

- **Approach**: To address the issue of fragmented context, we switched to section-wise chunking, treating each section as an independent chunk.
- **Results**: This method preserved section integrity but struggled with very large sections (e.g., 50-page sections), particularly when concise answers were required.

Custom Multi-Retriever RAG

- **Approach**: A custom multi-retriever RAG was developed, using two retrievers for handling smaller and larger chunks.
- **Results**: Achieved high accuracy by leveraging the advantages of both retriever types, optimizing the retrieval process.

DSPy for optimized prompts

- **Approach**: Leveraged DSPy to fine-tune and optimize prompts using carefully selected examples. This involved adjusting prompt structures and content to better guide the LLM's output based on specific retrieval contexts.
- **Results**: The optimized prompts led to accurate and contextually relevant responses, enhancing the overall generation process.

5. Major Challenges

Here are some major challenges we encountered while experimenting with different approaches:

- **Complex PDF Structure**: The complex structure of the PDFs, with 6-7 levels of nested sections, posed a significant challenge. Understanding the context of a particular section often depended on its ancestor sections, making it difficult to maintain coherence and relevance throughout the extraction process.
- **Optimal Chunk Size**: Determining the appropriate chunk size for different types of questions was problematic. Some questions required concise, one-line answers, while others necessitated detailed responses. Striking the right balance between chunk size and answer detail proved challenging.
- **Dispersed Information**: Answers were often spread across various parts of the PDF. This dispersion required the retriever to extract and compile information from multiple sections, complicating the retrieval process.

6. Evaluation

We have done evaluation manually. Following are the metrices and how we have evaluated them manually.

- **High Accuracy**: Ensured all facts in the answer are correct.
- **High Recall**: Answers are exhaustive, covering all relevant details.
- **High Precision**: No irrelevant facts are included; all information is directly related to the query.

7. Results

QUERY	RESPONSES
What are the requirements of the Off-road vehicle?	The requirements of an Off-road vehicle, specifically category N G, include: * The maximum height of the light may be increased to 1,500 mm (section 6.2.4.2). * For headlamps exceeding a height of 1,200 mm, the limits for the vertical inclination of the cut-off shall be between -1.5 per cent and -3.5 per cent, and the initial aim shall be set between -2 per cent and -2.5 per cent (section 6.2.6.1.2).
Tell me about Parking Lamp?	According to section 6.12, a parking lamp is an optional feature on motor vehicles not exceeding 6 meters in length and 2 meters in width (6.12.1). However, it is prohibited on all other vehicles (6.12.1). The number of lamps is determined by the arrangement, which can be either two lamps at the front and two lamps at the rear, or one lamp on each side (6.12.2 and 6.12.3). The position of the lamp is such that the point on the apparent surface in the direction of the reference axis, which is farthest from the vehicle's median longitudinal plane, shall not be more than 400 mm from the extreme outer edge of the vehicle (6.12.4.1). If there are two lamps, they shall be on the sides of the vehicle (6.12.4.1). In terms of height, there is no special requirement for M and N category vehicles, but for all other categories of vehicles, the lamp shall be above the ground, not less than 350 mm nor more than 1,500 mm (2,100 mm if the shape of the bodywork makes it impossible to keep within 1,500 mm) (6.12.4.2). There is no special

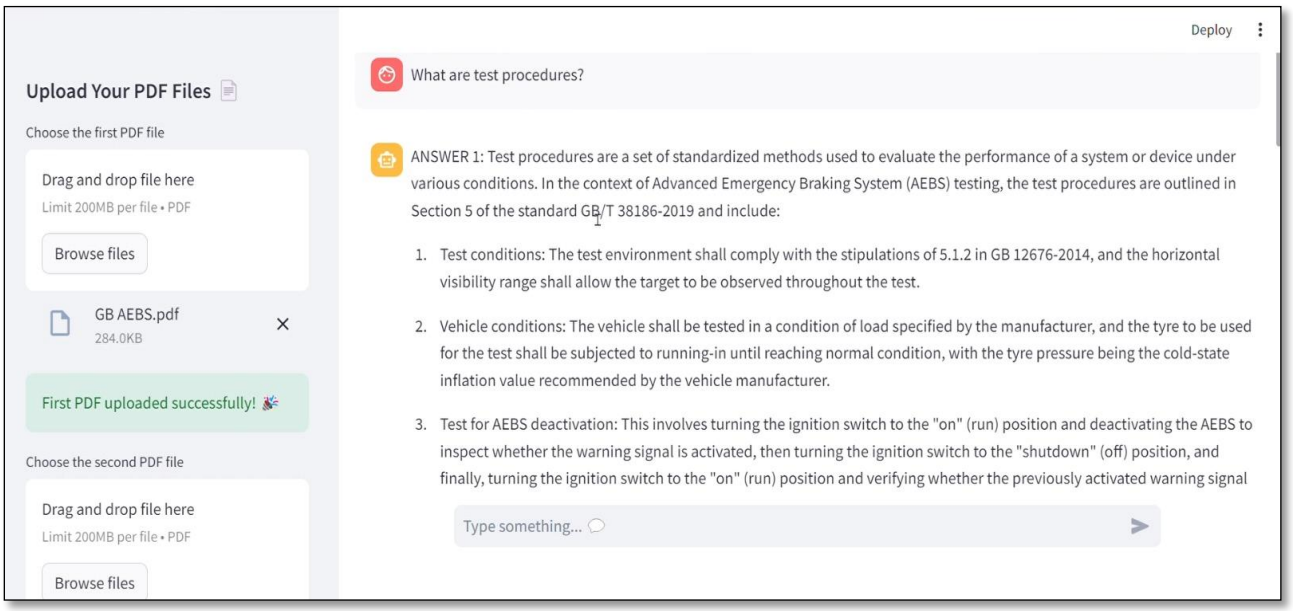
	requirement for the length of the lamp (6.12.4.3). The geometric visibility of the lamp is such that the horizontal angle is 45Â° outwards, forwards, and rearwards, and the vertical angle is 15Â° above and below the horizontal (6.12.5). The orientation of the lamp is such that it meets the requirements for visibility forwards and rearwards (6.12.6). The electrical connections of the lamp shall allow the parking lamp(s) on the same side of the vehicle to be lit independently of any other lamps (6.12.7). A circuit-closed tell-tale is optional, but if there is one, it shall not be possible to confuse it with the tell-tale for the front and rear position lamps (6.12.8). Finally, the functioning of this lamp may also be performed by simultaneously switching on the front and rear position lamps on the same side of the vehicle (6.12.9).
Is it possible to combine Main beam and Passing beam as a combined light?	Based on the provided context, it appears that combining Main beam and Passing beam as a single light is not explicitly allowed or mentioned. While there is no specific information that explicitly prohibits combining them, the context emphasizes the distinction between Main beam and Passing beam, highlighting their separate functions and switching operations. The provisions in sections 6.1.7.4, 6.1.7.5, 6.22.7.1.1, and 6.2.7.1, as well as the requirement to switch off all main-beam headlamps simultaneously when changing over to the dipped-beam, suggest that they are distinct and separate lighting functions. Additionally, the separate controls and electrical connections for main beam and passing beam, as mentioned in sections 6.22.7.2, 6.22.7.3, and 5.12, further support the notion that they are intended to operate independently. Therefore, without further information, it cannot be concluded that combining Main beam and Passing beam as a combined light is allowed.

Most results were accurate and complete, but there's always room for improvement. Further enhancements could be made by refining the DSPy implementation.

8. Application

Access the Application: [Document Comparison Chatbot]

Access the Code: [GitHub Repository]



Reason for high Latency in Response Generation:

- **Complex Workflow**: The process involved 4 vector stores and 7 LLM calls, contributing to delays.
- **Potential Solutions**: This application is deployed for our personal showcase, client could switch to cloud services like Claude from Bedrock, could significantly reduce response times.

9. Future Improvements

Here are some potential future improvements that could be explored further:

- Future work could enhance multi-language support and develop a robust knowledge database for better handling of diverse regulatory documents.
- Optimizing computational efficiency and response times using advanced techniques and cloud services like Claude from Bedrock is another area for improvement.
- Due to time constraints, we implemented a minimal version of DSPy. However, there is potential for further refinement and enhancement in future iterations.

10. Conclusion

The Document Comparison Chatbot project has made significant strides in automating the compliance verification process. Within the limited time available, we tested and refined various approaches, achieving notable improvements. However, there is always room for further enhancement. Future work could focus on improving efficiency and expanding capabilities to make the tool even more versatile and effective across diverse regulatory scenarios.

11. References

Articles: List of relevant articals consulted during the project

- DSPy Introduction
- Improving RAG Performance: 5 Key Techniques

Libraries and Tools:

- Libraries: langchain, pypdf, langchain_groq, sentence_transformers, pdfplumber, faiss-cpu, streamlit
- Tools: FAISS, RecursiveCharacterTextSplitter, HuggingFace BGE Embeddings, PromptTemplate, DSPy, RetrievalQA