Assignment: Vehicle Specification Extraction

Objective

Build a basic system that extracts **vehicle specifications** (e.g., torque values, fluid capacities, part numbers) from an **automotive service manual PDF** using **LLMs and retrieval**.
Focus on **text-based extraction** only (ignore images or diagrams).

Task Details

1. **PDF Text Extraction**
   Parse and extract text from the service manual using tools such as PyMuPDF, pdfminer, or pypdf.
2. **Chunking & Embedding**
   Split text into logical sections and create embeddings (e.g., OpenAI, HuggingFace).
3. **Retrieval-Augmented Query**
   For a given query like "Torque for brake caliper bolts," retrieve the most relevant chunks and use an LLM to extract structured data.
4. **Output Format**
   Return structured results in JSON or CSV, e.g.:

```
[
        {
          "component": "Brake Caliper Bolt",
          "spec_type": "Torque",
          "value": "35",
          "unit": "Nm"
        }
]
```

Deliverables

1. A code notebook or Python repo implementing the pipeline.
2. A README explaining the design, tools used, and ideas for improvement.

Evaluation Criteria

Criteria	Description
Concept Understanding	Application of LLM and retrieval fundamentals
Code Clarity	Readable, modular, and documented code
Pipeline Design	Logical approach to text cleaning, chunking, and retrieval
Output Quality	Accuracy and clarity of extracted specifications
Bonus	Creativity such as UI or basic OCR integration

Suggested Tools

1. **PDF Parsing:** PyMuPDF, pdfminer
2. **Embeddings:** OpenAI, Sentence-Transformers
3. **LLMs:** GPT-3.5, GPT-4, Mistral, Llama-3
4. **Vector Store:** FAISS, Chroma
5. **Framework (optional):** LangChain

Notes

You will be given a sample service manual along with this assignment. Use the provided service manual for working on the assignment.

For any questions, please reach out to abhishek.kumar@predii.com