Research Report

Varrsan D

Task 1: Research & Tool Selection

1. **Problem Statement**

The task requires extracting an "Approved Makes and Manufacturer" table from 100+ PDFs that may have:

- Different layouts and formats.
- Multi-page tables.
- Scanned or handwritten content requiring OCR.

The extracted data should be structured in JSON format, ensuring accuracy and scalability.

2. **Problem Breakdown**

Digital PDFs: Extract tables directly from structured PDFs.
Scanned PDFs: Convert images to text using OCR.
Handwritten PDFs: Use advanced OCR tools and LLM-based structuring.
Table Structure Recognition: Identifying the location of tables in PDFs.
Text Extraction: Extracting the actual content, whether it's digital text or scanned text requiring OCR.
Multi-Page Handling: Detect and merge table data across pages.
Error Handling: Return errors for unrecognized table formats rather than incorrect data.

3**. List of Methods for Extraction**

A. OCR-Based Approaches (For scanned & handwritten documents)

- **Tesseract OCR:** Open-source OCR, good for printed text but struggles with handwriting.
- **AWS Textract:** Cloud-based OCR with structured data extraction capabilities.
- **Google Vision OCR:** Good for extracting structured text from images but requires API usage.

B. Digital PDF Parsing Methods (For well-formatted tables)

- **Camelot & PDFPlumber:** Best for structured PDFs where tables have clear layouts.
- **Tabula:** Works for simple table extraction but struggles with complex layouts.

C. LLM-Based Approaches (For complex unstructured tables)

- **GPT-4, Claude 3:** Can infer structure in noisy data but computationally expensive.

4. Comparison of Methods

Method	Accuracy	Speed	Cost	Complexity	Best Use Case
Tesseract OCR	Medium	Fast	Free	Medium	Simple scanned PDFs
AWS Textract	High	Medium	Paid	Low	High-quality OCR extraction
Google Vision OCR	High	Medium	Paid	Low	Extracting text from images
Camelot/ PDFPlumber	High	Fast	Free	Low	Well-structured PDFs
Tabula	Medium	Fast	Free	Low	Simple tables in PDFs
LLM (GPT-4, Claude 3)	Very High	Slow	Expensive	High	Handling complex, unstructured data

5. What Will I Try and Why?

Final Approach: Hybrid Pipeline

1. Use Camelot/PDFPlumber for extracting structured tables from PDFs.
2. If the table is missing or unstructured, apply OCR (Tesseract/AWS Textract).
3. If OCR output is unclear, use LLM-based post-processing for structuring.
4. Implement robust error handling – return an error message instead of incorrect data if confidence is low.