OCR Data Extraction Performance Report

Overview

This report presents the performance results of an OCR-based data extraction system. The system is designed to handle a variety of file types including PDFs, images, documents, and spreadsheets. Performance is evaluated in terms of text length extracted and processing time (in seconds).

Supported File Types & Results

File Type	Text Length	Time Taken (sec)
JPG / JPEG / PNG	8,068	10.91
DOCX	53,798	0.03
XLSX	4,631	0.01
TXT	1,563	0.00

Detailed PDF Breakdown

PDF Type Description	Pages	Text Length	Time (sec)
Mixed content (Text + Image)	1	10,056	10.39
Mixed content (Text + Image)	2	14,695	10.74
Complex: Text, Images, QR, Signatures	4	42,568	50.91
Extensive complex data	30	235,432	225.34
Handwritten content	100	85,259	348.64
Simple text-only PDF	1	4,047	0.06

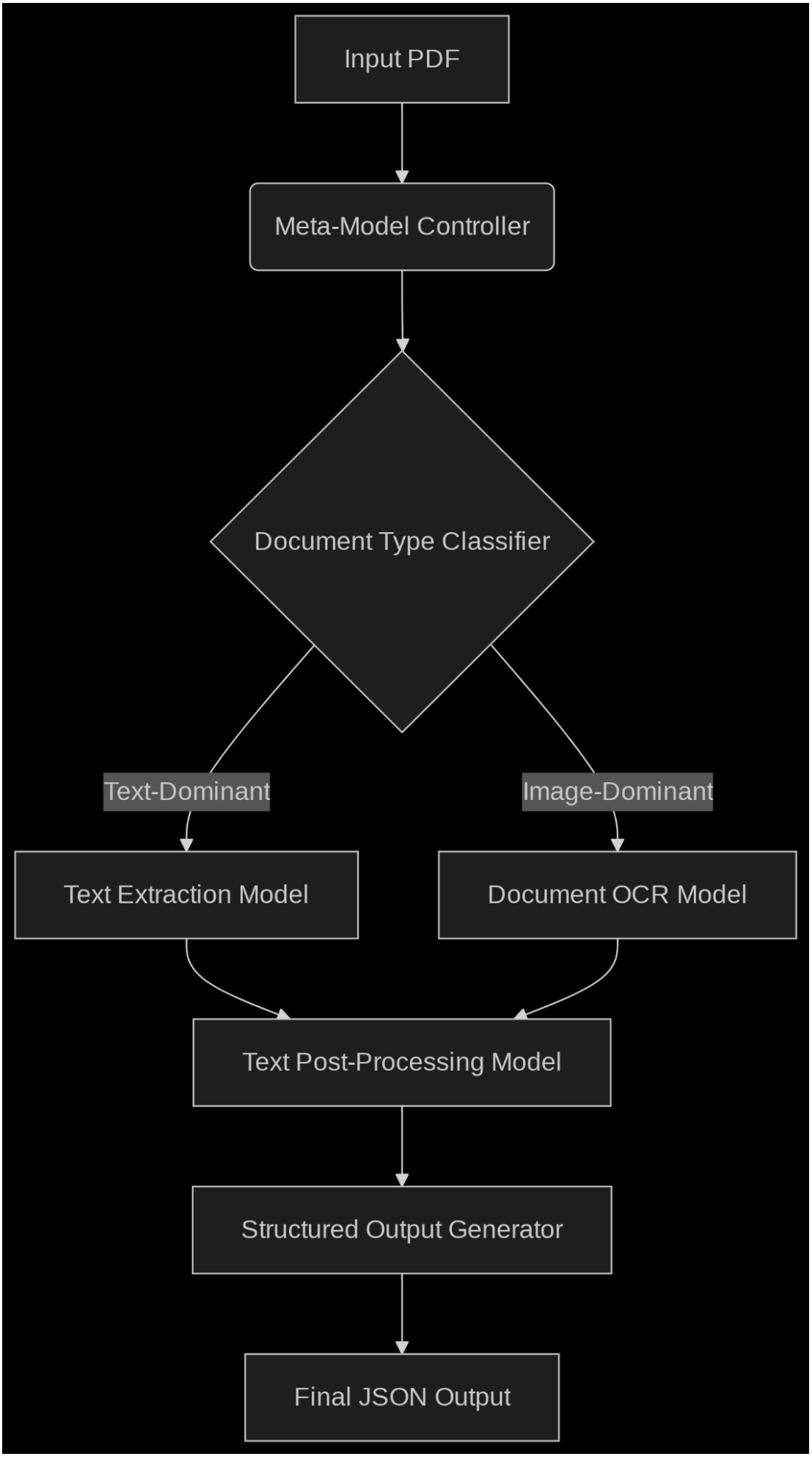
Summary

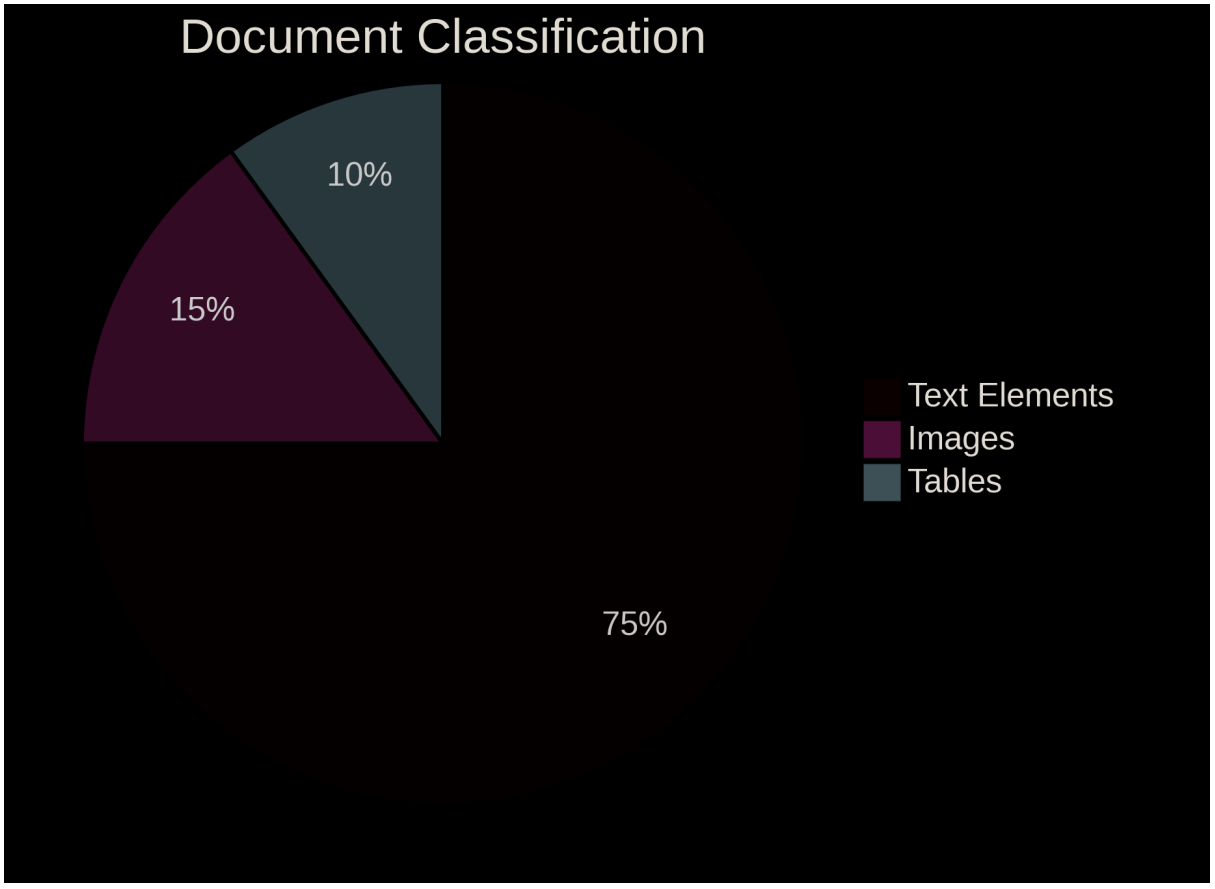
- The system performs exceptionally fast on structured digital files like DOCX, XLSX, and TXT.

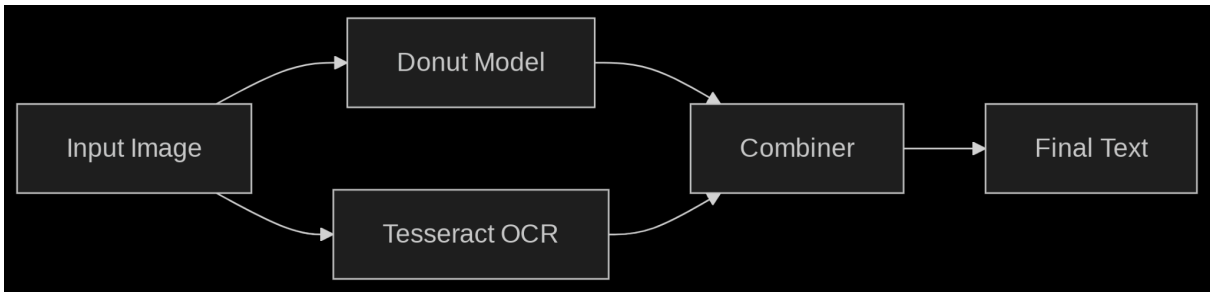
- Image-based files and handwritten PDFs require significantly more time due to the OCR and preprocessing steps involved.

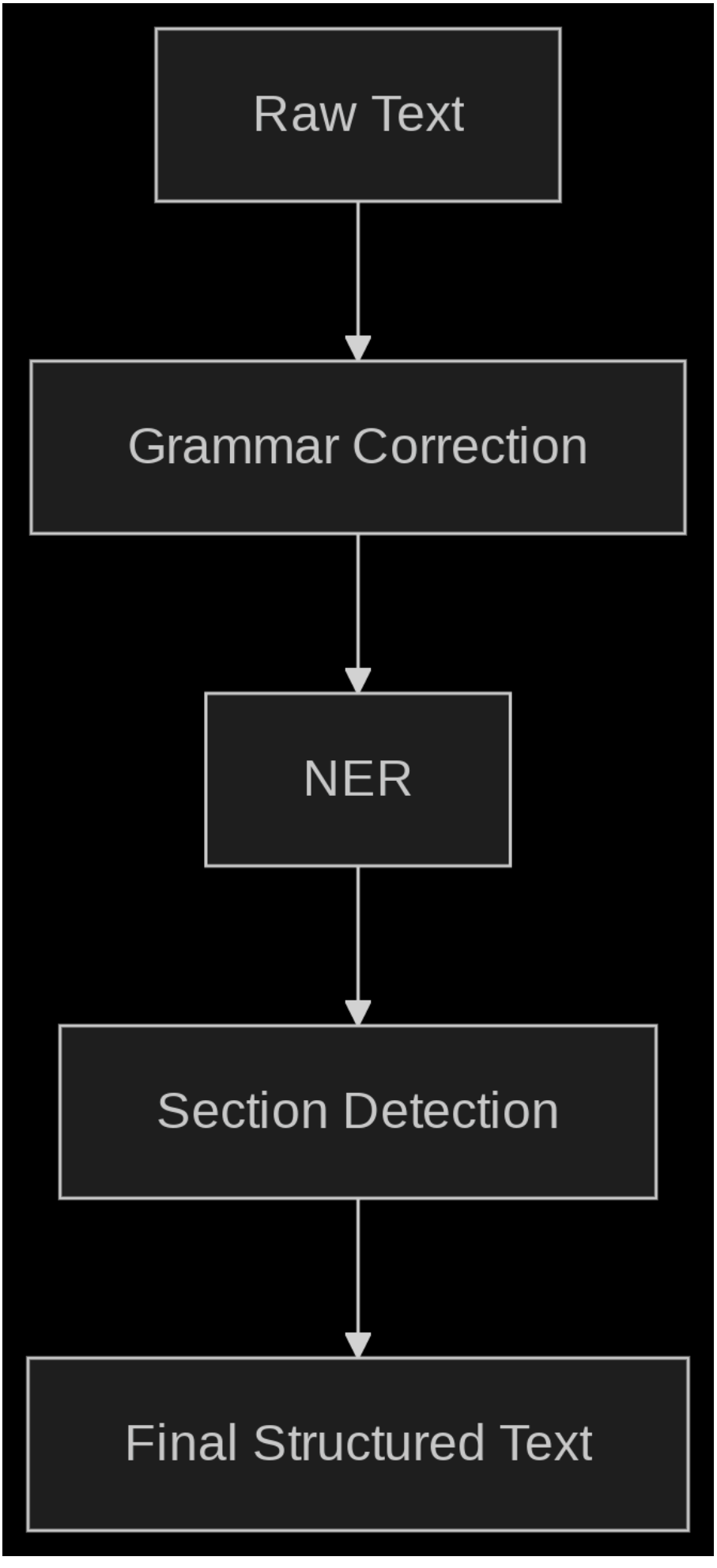
- As the number of pages and complexity increases, the processing time scales accordingly.

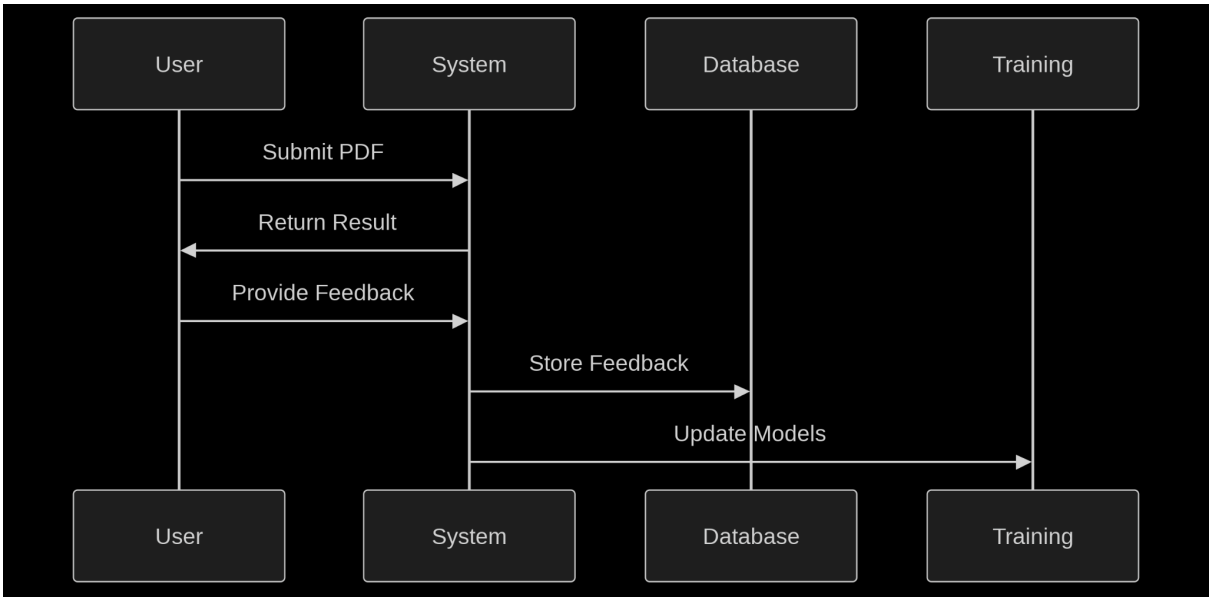
- Simple, text-only PDFs are processed almost instantly.

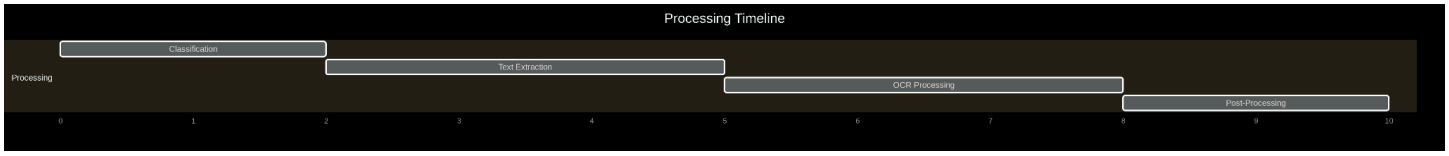

Input PDF
Meta-Model Controller
Document Type Classifier
Text-Dominant
Image-Dominant
Text Extraction Model
Document OCR Model
Text Post-Processing Model
Structured Output Generator
Final JSON Output

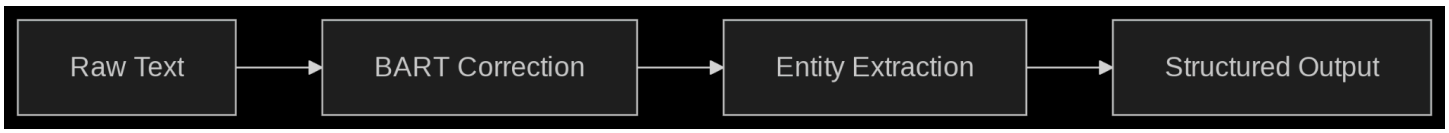

Document Classification
10%
15%
75%
Text Elements
Images
Tables


Input Image
Donut Model
Tesseract OCR
Combiner
Final Text


Raw Text
Grammar Correction
NER
Section Detection
Final Structured Text


User
System
Database
Training
Submit PDF
Return Result
Provide Feedback
Store Feedback
Update Models
User
System
Database
Training


Processing Timeline
Processing
Classification
Text Extraction
OCR Processing
Post-Processing
0
1
2
3
4
5
6
7
8
9
10


Raw Text
BART Correction
Entity Extraction
Structured Output