

EXTERN + OUTAMATION  
EXTERNSHIP

---

# ADVANCED AI-POWERED DOCUMENT INSIGHT AND DATA EXTRACTION PROJECT

---

SPRINT 4 STATUS UPDATE

by Lashawn Fofung, MBA, PMP, CSM, CSPO

LF

## SKILLS

- Python
- AI/ML
- Data Scrapping
- Prompt Engineering
- LLMs
- Text Extraction
- Document Intelligence
- Python Automation
- Optical Character Recognition (OCR)
- Retrieval-Augmented Generation (RAG)

Helping businesses improve, automate, and architect critical tasks (such as document handling, customer service, and data management) by saving thousands of hours and boosting accuracy.



# SPRINTS



01

AI Document Intelligence

02

Python & Google Colab

03

Python Data Extraction

04

Optimizing OCR

05

Implementing RAG

06

Optimizing RAG Pipelines

07

Blob Processing & Classification

08

Interactive Chatbot

09

AI-Powered Document Automation Platform

---

PROJECT TIMELINE

---

10 WEEKS

# Optimizing OCR

Business Automation

# Background: 10-Week Project

The Outamation Advanced AI-Powered Document Insights and Data Extraction Externship is a comprehensive 10-week program. Its core goal and purpose are to master the required skills to create an AI-Powered Document Automation Platform. The program focuses on bridging foundational technologies (ML, LLMs, NLP, Computer Vision) with high-value industry applications, such as solving complex, document-heavy challenges in mortgage automation, ultimately transforming slow, manual review processes into automated, efficient workflows.

A wireframe diagram illustrating the structure of a bank statement and a mortgage application. On the left, a wireframe of a bank statement form is shown with fields for name, gross pay, hire date, and account number. On the right, a wireframe of a mortgage application form is shown with fields for name, gross pay, hire date, account number, and account balance. The wireframes are interconnected by a network of lines, symbolizing the flow of data between different systems.



# Challenge & Solution: Building The Document Intelligence (AI) Pipeline

- **The Business Challenge:**
  - Automating the highly manual and time-consuming process of mortgage underwriting by handling 200-300 pages of unstructured documents.
- **The Solution:**
  - Designing and implementing a robust, AI-Powered Document Automation Platform to transform raw files into structured, intelligent insights.
- **The 4-Step Document Intelligence Workflow:**
  - Step 1. Document Acquisition & Pre-Processing (The Input)
  - Step 2. Data Extraction & Preparation
  - Step 3. Knowledge Base & Modeling (The AI Engine)
  - Step 4. Output, Review, & Integration (The Results)

This sprint mastered the critical front-end of the AI pipeline: acquiring unstructured documents, pre-processing images for clarity, extracting text via OCR, and preparing clean, structured data for analysis.

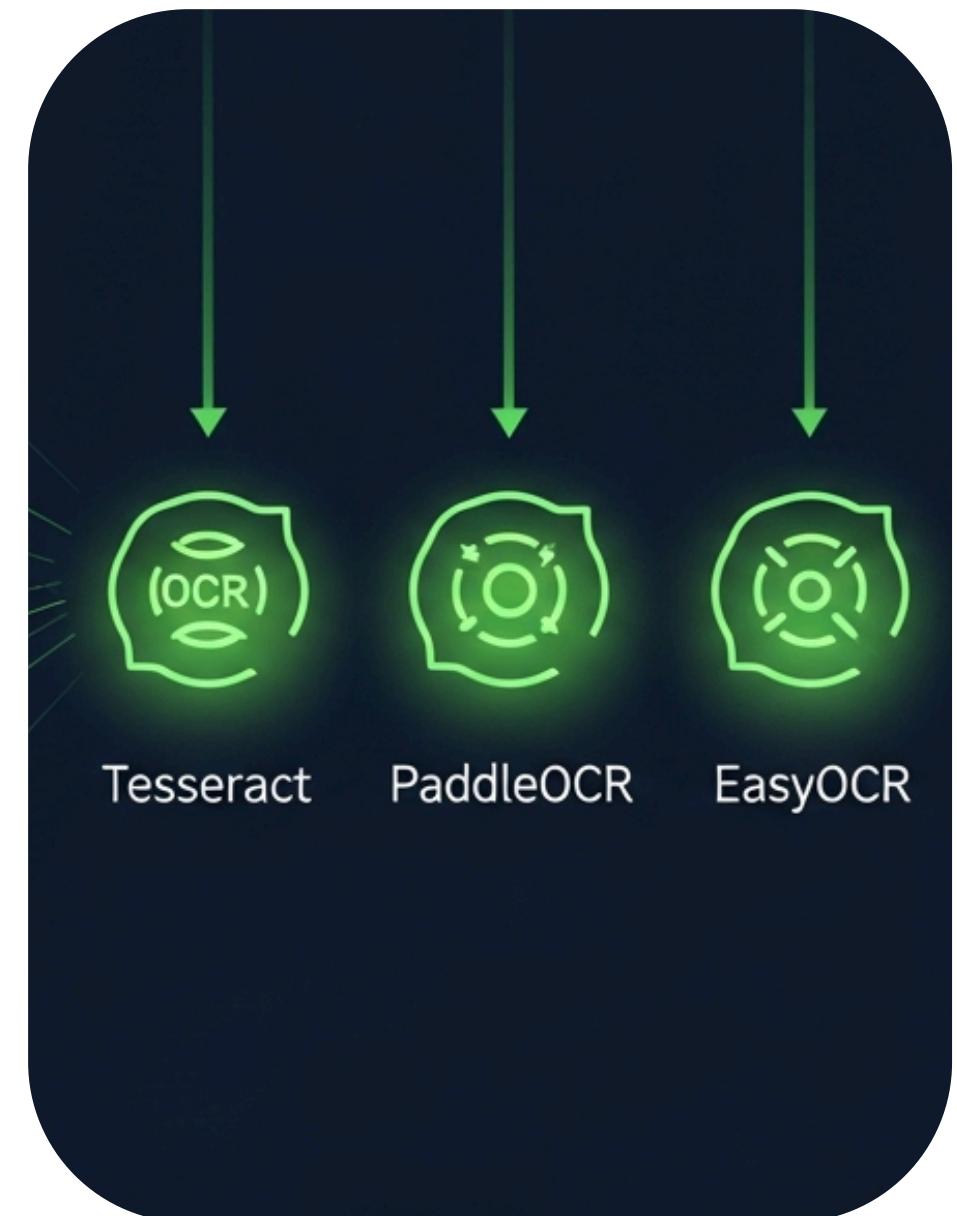
# SPRINT 4 FOCUS

Steps 1 & 2: End-to-End OCR  
(Acquisition, Pre-processing, Extraction, & Preparation)

# Overall Status: Sprint 4 Complete

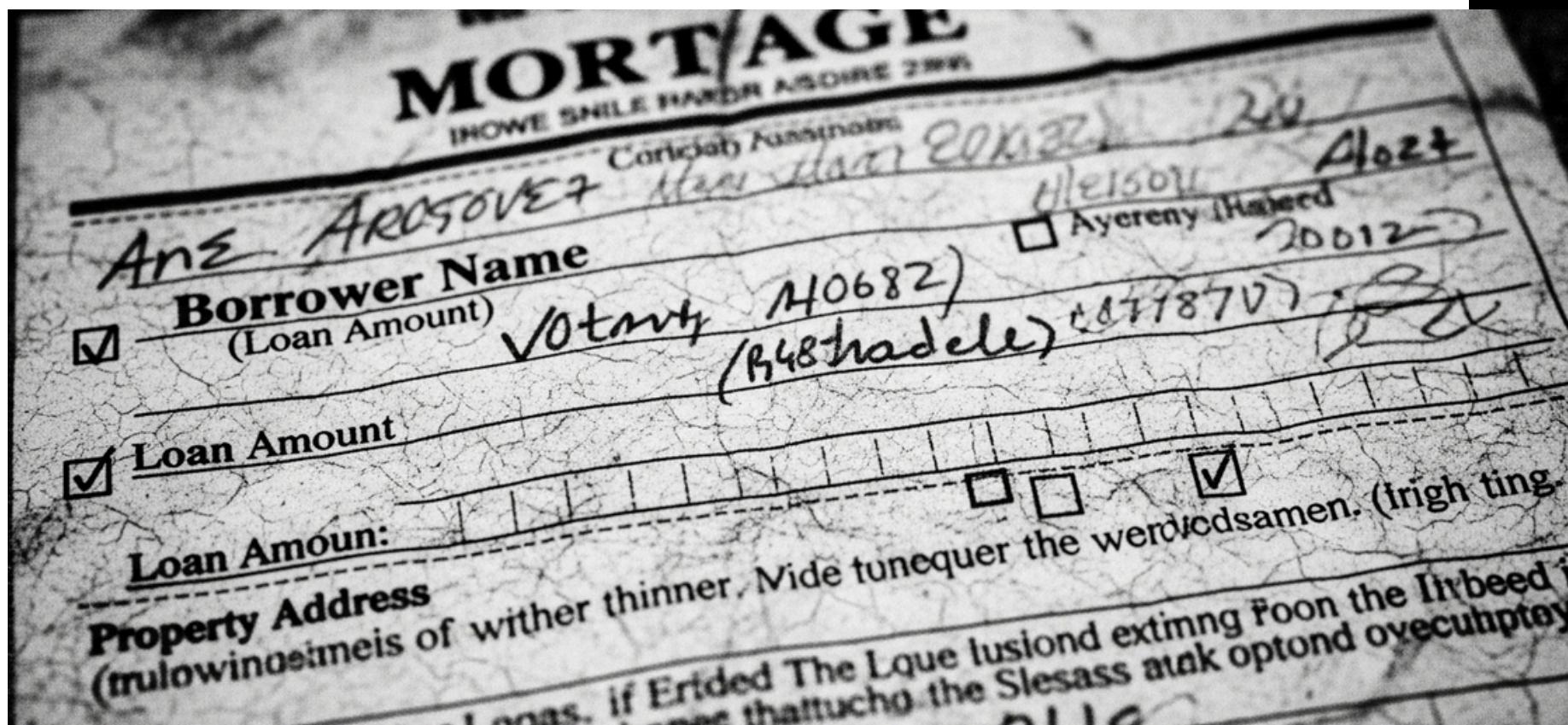
I've successfully concluded Sprint 4, focused on mastering Advanced OCR Comparison and Layout-Aware Extraction for the AI-Powered Document Automation Platform.

- **Comparative Engine Analysis:** Completed rigorous, head-to-head benchmarking of Tesseract, PaddleOCR, and EasyOCR to select the optimal, high-accuracy engine.
- **Pipeline Phase Complete:** Finalized the Document Acquisition & Pre-Processing phase, establishing a high-fidelity input source for downstream AI/LLM modeling.
- **Structured Data Output:** Advanced ability to transform complex, multi-page scans into clean, layout-aware JSON outputs, directly solving the input quality challenge.
- **Reliability Foundation:** Successfully established a robust OCR foundation critical for ensuring the overall accuracy and reliability of the AI-Powered Document Automation Platform.



# Key Accomplishments - Baseline Pipeline Established

- Built an end-to-end workflow using Tesseract, refining my skills in image pre-processing, text extraction, and structuring data into JSON.
- The initial objective was to establish a working, end-to-end framework for the document intelligence platform using a basic, known OCR engine as a starting point.



## Key tools and inputs:

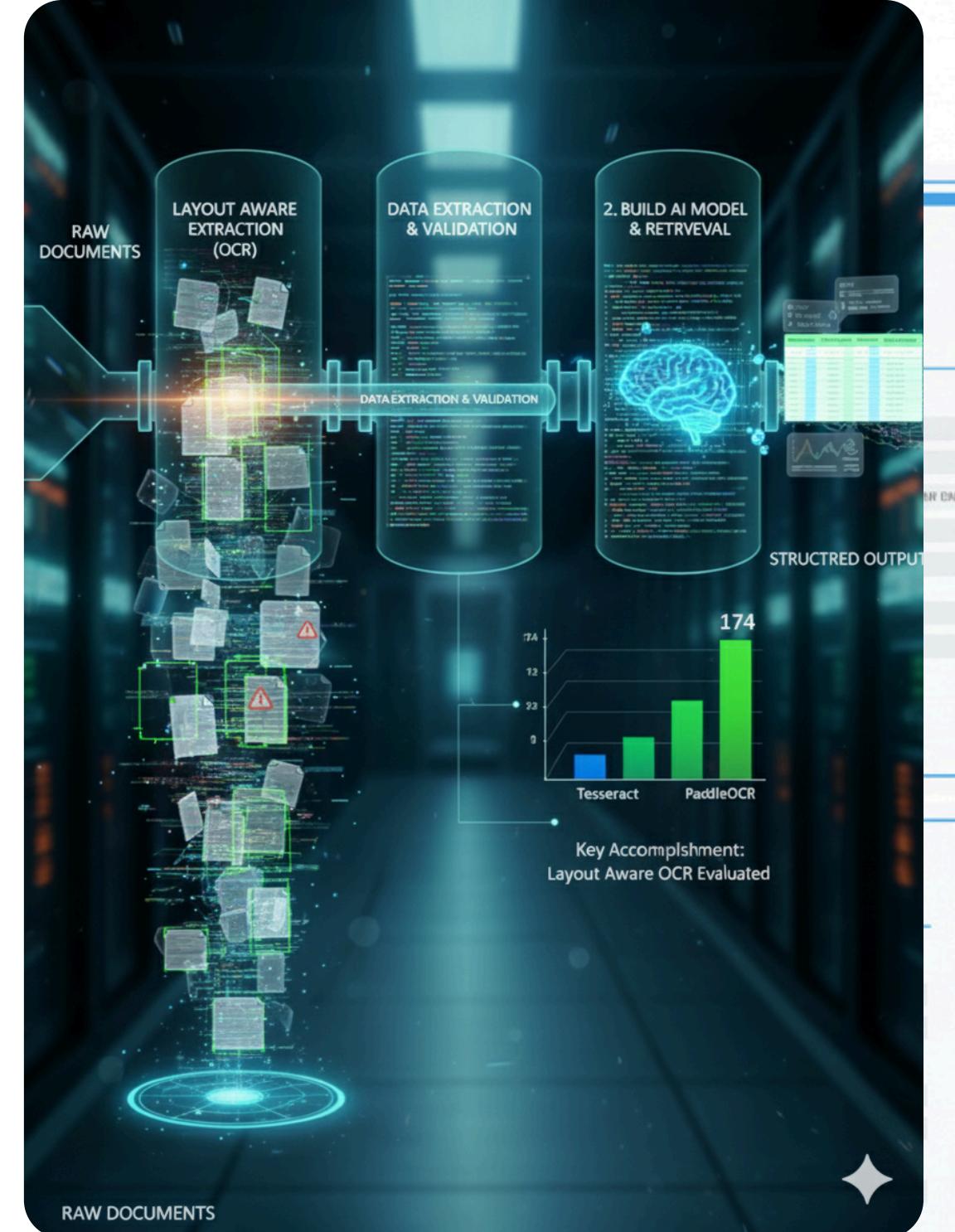
- **Tool:** Tesseract (Baseline PoC)
- **Goal:** Validate the architectural flow from image input to structured JSON output.
- **Input:** Raw, messy, scanned mortgage document PDFs.

# Baseline Pipeline Established – Key Process Steps

- **Pre-processing**
  - Implemented image quality adjustments (de-skewing, binarization) to optimize Tesseract input.
- **Text Extraction**
  - Ran raw OCR to get plain text output.
- **Spatial Structuring**
  - Parsed Tesseract's HOGR output to associate every word with its x/y coordinates.
- **Initial Finding**
  - Tesseract's simple output confirmed the pipeline flow but highlighted the severe need for layout -aware extraction methods.

# Key Accomplishments - Layout-Aware OCR Evaluated

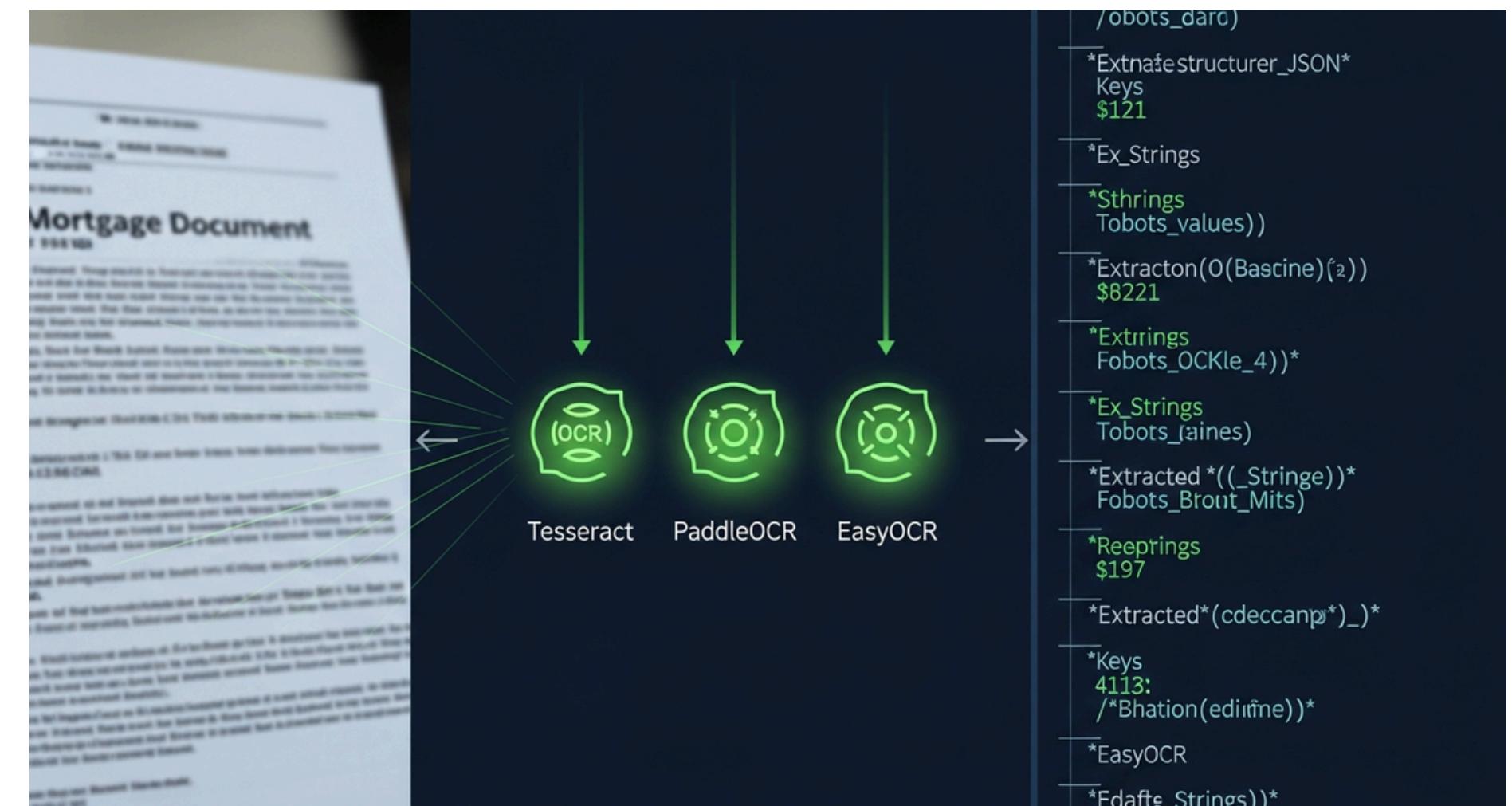
- **Advanced Engine Deployment**
  - Successfully deployed and tested advanced layout-aware OCR engines (PaddleOCR & EasyOCR), moving beyond basic text extraction.
- **Structured Data Validation**
  - Proved the capability to handle complex documents (e.g., mortgage forms), tables, and fields by focusing on structural intelligence, not just character recognition.
- **Quantitative Success**
  - Directly validated the performance gap, resulting in 174 items found by PaddleOCR compared to only 3 items by the baseline Tesseract, confirming superior data capture.
- **Production Readiness**
  - Provided a data-driven recommendation to adopt the superior layout-aware tool, ensuring high data integrity and completeness for the final AI production pipeline.



# Key Accomplishments – Comparative Analysis Complete

Conducted a head-to-head benchmark of all three engines to provide a data-driven recommendation for production.

- **PaddleOCR:** Gave the best structure overall, especially with longer text and labels, though it was harder to install.
- **EasyOCR:** Handled the form layout well and was fast to set up.
- **Tesseract:** Easy to use but gave very messy output — lots of broken lines and missed sections.



# Key Accomplishments -

## OCR Engine Comparison

Metric	Tesseract OCR	PaddleOCR	EasyOCR
<b>Accuracy</b>	<b>Low</b> (Captured least data points: 3 data points)	<b>High</b> (Captured the most data points: 174 items)	<b>Good</b> (Captured almost the same data points: 165 items)
<b>Layout Awareness</b>	<b>Basic</b> (Requires pre-processing for complex documents/tables).	<b>Excellent</b> (Built-in tools for structure, table recognition, and reading order).	<b>Limited/Basic</b> (Focuses on text detection; minimal structure understanding).
<b>Ease of Use</b>	<b>Easy</b> (Simple command-line use, but fine-tuning is complex).	<b>Moderate/Advanced</b> (Powerful, but requires more configuration).	<b>High</b> (Very simple installation and straight forward Python API).
<b>Output Quality</b>	<b>High</b> quality for clean text; Messy; Structured output requires extra effort.	<b>Superior</b> for structured data extraction and accurate bounding boxes.	<b>Good</b> text and bounding boxes for simple, quick extraction tasks.

# OCR Engines Output

OCR Results - Page 1 (174 items found)

Your actual rate, payment, and cost could be higher. Get an official Loan Estimate before choosing a loan.

**FEES WORKSHEET**

Fee Details and Summary

Applicants	John Q. Smith / Mary A. Smith	Application No.	samplesmith
Prepared By:	XYZ Lender	Date Prepared:	10/05/2015
		Loan Program:	30 YEAR FIXED - Purchase
THIS IS NOT A GOOD FAITH ESTIMATE (GFE). This "Fees Worksheet" is provided for informational purposes ONLY, to assist you in determining an estimate of cash that may be required to close and an estimate of your proposed monthly mortgage payment. Actual charges may be more or less, and your transaction may not involve a fee for every item listed.			
Total Loan Amount:	\$ 380,000	Interest Rate:	4.250 %
Term/Due In:	360 / 360 mths		
Fee	Paid To	Paid By (Fee Split*)	Amount PFC/F/POC
<b>ORIGINATION CHARGES</b>			
Underwriting Fee	XYZ Lender	Borrower	\$ 550.00 ✓
Wire Transfer Fee	XYZ Lender	Borrower	\$ 75.00 ✓
Administration Fee	XYZ Lender	Borrower	\$ 445.00 ✓
<b>OTHER CHARGES</b>			
Appraisal Fee	XYZ Lender	Borrower	\$ 525.00 ✓
Credit Report Fee	XYZ Lender	Borrower	\$ 25.00 ✓
Fax Fee	XYZ Lender	Borrower	\$ 25.00 ✓
File Certification Fee	XYZ Lender	Borrower	\$ 20.00 ✓
Closing/Escrow Fee	Settlement Agent	Borrower	\$ 860.00 ✓
Document Preparation Fee	Settlement Agent	Borrower	\$ 150.00 ✓
Notary Fee	NOTARY	Borrower	\$ 150.00 ✓
Lender's Title Insurance		Borrower	\$ 650.00 ✓
Title - Courier Fee	Settlement Agent	Borrower	\$ 50.00 ✓
Electronic Document Delivery Fee	Settlement Agent	Borrower	\$ 50.00 ✓
Pest Inspection Fee	PEST CONTROL	Borrower	\$ 50.00 ✓
Home Inspection	HI COMPANY	Borrower	\$ 450.00 ✓
Mortgage Recording Charge		Borrower	\$ 150.00 ✓
Daily Interest Charges	XYZ Lender	Borrower	\$ 44,861.11 x 25 day(s) \$ 1,121.53 ✓
Hazard Insurance Premium		Borrower	\$ 39.58 x 12 mth(s) \$ 475.00

TOTAL ESTIMATED FUNDS NEEDED TO CLOSE:		TOTAL ESTIMATED MONTHLY PAYMENT:	
Purchase Price (+)	475,000.00	Loan Amount (-)	380,000.00
Affiliations (+)		Cash Deposit	5,000.00
Land (+)		Principal & Interest	1,869.37
Refi (inv. debts to be paid off) (+)		Other Financing (P & I)	
Refi Prepaid Items/Reserves (+)	1,121.53	Hazard Insurance	39.58
Est. Closing Costs (+)	4,520.00	Real Estate Taxes	400.00
Total Estimated Funds needed to close	95,641.53	Total Monthly Payment	2,308.95

\* PFC = Prepaid Finance Charge      \*\* B = Borrower      S = Seller      Br = Broker      L = Lender      TP = Third Party      C = Correspondent  
Cals Form - fees.frm (09/2015)

PaddleOCR

174 Items Found

EasyOCR Results - 165 text segments found

Your actual rate, payment, and cost could be higher. Get an official Loan Estimate before choosing a loan.

**FEES WORKSHEET**

Fee Details and Summary

Applicants	John Q. Smith / Mary A. Smith	Application No.	samplesmith
Prepared By:	XYZ Lender	Date Prepared:	10/05/2015
		Loan Program:	30 YEAR FIXED - Purchase
THIS IS NOT A GOOD FAITH ESTIMATE (GFE). This "Fees Worksheet" is provided for informational purposes ONLY, to assist you in determining an estimate of cash that may be required to close and an estimate of your proposed monthly mortgage payment. Actual charges may be more or less, and your transaction may not involve a fee for every item listed.			
Total Loan Amount:	\$ 380,000	Interest Rate:	4.250 %
Term/Due In:	360 / 360 mths		
Fee	Paid To	Paid By (Fee Split*)	Amount PFC/F/POC
<b>ORIGINATION CHARGES</b>			
Underwriting Fee	XYZ Lender	Borrower	\$ 550.00 ✓
Wire Transfer Fee	XYZ Lender	Borrower	\$ 75.00 ✓
Administration Fee	XYZ Lender	Borrower	\$ 445.00 ✓
<b>OTHER CHARGES</b>			
Appraisal Fee	XYZ Lender	Borrower	\$ 525.00 ✓
Credit Report Fee	XYZ Lender	Borrower	\$ 25.00 ✓
Tax Service Fee	XYZ Lender	Borrower	\$ 80.00 ✓
Flood Certification Fee	XYZ Lender	Borrower	\$ 20.00 ✓
Closing/Escrow Fee	Settlement Agent	Borrower	\$ 860.00 ✓
Document Preparation Fee	Settlement Agent	Borrower	\$ 150.00 ✓
Notary Fee	NOTARY	Borrower	\$ 150.00 ✓
Lender's Title Insurance		Borrower	\$ 650.00 ✓
Title - Courier Fee	Settlement Agent	Borrower	\$ 50.00 ✓
Electronic Document Delivery Fee	Settlement Agent	Borrower	\$ 50.00 ✓
Pest Inspection Fee	PEST CONTROL	Borrower	\$ 50.00 ✓
Home Inspection	HI COMPANY	Borrower	\$ 450.00 ✓
Mortgage Recording Charge		Borrower	\$ 150.00 ✓
Daily Interest Charges	XYZ Lender	Borrower	\$ 44,861.11 x 25 day(s) \$ 1,121.53 ✓
Hazard Insurance Premium		Borrower	\$ 39.58 x 12 mth(s) \$ 475.00

TOTAL ESTIMATED FUNDS NEEDED TO CLOSE:		TOTAL ESTIMATED MONTHLY PAYMENT:	
Purchase Price (+)	475,000.00	Loan Amount (-)	380,000.00
Affiliations (+)		Cash Deposit	5,000.00
Land (+)		Principal & Interest	1,869.37
Refi (inv. debts to be paid off) (+)		Other Financing (P & I)	
Refi Prepaid Items/Reserves (+)	1,121.53	Hazard Insurance	39.58
Est. Closing Costs (+)	4,520.00	Real Estate Taxes	400.00
Total Estimated Funds needed to close	95,641.53	Total Monthly Payment	2,308.95

\* PFC = Prepaid Finance Charge      \*\* B = Borrower      S = Seller      Br = Broker      L = Lender      TP = Third Party      C = Correspondent  
Cals Form - fees.frm (09/2015)

EasyOCR

165 Items Found

Your actual rate, payment, and cost could be higher. Get an official Loan Estimate before choosing a loan.

**FEES WORKSHEET**

Fee Details and Summary

Applicants	John Q. Smith / Mary A. Smith	Application No.	samplesmith
Prepared By:	XYZ Lender	Date Prepared:	10/05/2015
		Loan Program:	30 YEAR FIXED - Purchase
THIS IS NOT A GOOD FAITH ESTIMATE (GFE). This "Fees Worksheet" is provided for informational purposes ONLY, to assist you in determining an estimate of cash that may be required to close and an estimate of your proposed monthly mortgage payment. Actual charges may be more or less, and your transaction may not involve a fee for every item listed.			
Total Loan Amount:	\$ 380,000	Interest Rate:	4.250 %
Term/Due In:	360 / 360 mths		
Fee	Paid To	Paid By (Fee Split*)	Amount PFC/F/POC
<b>ORIGINATION CHARGES</b>			
Underwriting Fee	XYZ Lender	Borrower	\$ 550.00 ✓
Wire Transfer Fee	XYZ Lender	Borrower	\$ 75.00 ✓
Administration Fee	XYZ Lender	Borrower	\$ 445.00 ✓
<b>OTHER CHARGES</b>			
Appraisal Fee	XYZ Lender	Borrower	\$ 625.00 ✓
Credit Report Fee	XYZ Lender	Borrower	\$ 25.00 ✓
Tax Service Fee	XYZ Lender	Borrower	\$ 80.00 ✓
Flood Certification Fee	XYZ Lender	Borrower	\$ 20.00 ✓
Closing/Escrow Fee	Settlement Agent	Borrower	\$ 860.00 ✓
Document Preparation Fee	Settlement Agent	Borrower	\$ 150.00 ✓
Notary Fee	NOTARY	Borrower	\$ 150.00 ✓
Lender's Title Insurance		Borrower	\$ 650.00 ✓
Title - Courier Fee	Settlement Agent	Borrower	\$ 50.00 ✓
Electronic Document Delivery Fee	Settlement Agent	Borrower	\$ 50.00 ✓
Pest Inspection Fee	PEST CONTROL	Borrower	\$ 60.00
Home Inspection	HI COMPANY	Borrower	\$ 450.00
Mortgage Recording Charge		Borrower	\$ 150.00
Daily Interest Charges	XYZ Lender	Borrower	\$ 44,861.11 x 25 day(s) \$ 1,121.53 ✓
Hazard Insurance Premium		Borrower	\$ 39.58 x 12 mth(s) \$ 475.00

TOTAL ESTIMATED FUNDS NEEDED TO CLOSE:		TOTAL ESTIMATED MONTHLY PAYMENT:	
Purchase Price (+)	475,000.00	Loan Amount (-)	380,000.00
Affiliations (+)		Cash Deposit	5,000.00
Land (+)		Principal & Interest	1,869.37
Refi (inv. debts to be paid off) (+)		Other Financing (P & I)	
Refi Prepaid Items/Reserves (+)	1,121.53	Hazard Insurance	39.58
Est. Closing Costs (+)	4,520.00	Real Estate Taxes	400.00
Total Estimated Funds needed to close	95,641.53	Total Monthly Payment	2,308.95

\* PFC = Prepaid Finance Charge      \*\* B = Borrower      S = Seller      Br = Broker      L = Lender      TP = Third Party      C = Correspondent  
Cals Form - fees.frm (09/2015)

Tesseract

3 Items Found

# OCR Engines Output

```
Total text segments extracted: 174

All Extracted Text:
=====
1. FEES WORKSHEET
2. Fee Details and Summary
3. Applicants:
4. John Q. Smith / Mary A. Smith
5. Application No:
6. samplesmith
7. Prepared By:
8. XYZ Lender
9. Date Prepared:
10. 10/05/2015
11. Loan Program:
12. 30 YEAR FIXED -Purchase
13. THIS IS NOT A GOOD FAITH ESTIMATE (GFE). Th
14. you in determining an estimate of cash that
15. payment. Actual charges may be more or less
16. Total Loan Amount: $ 380,000
17. Interest Rate:
18. 4.250 %
19. Term/Due In:
20. 360 / 360 mths
21. Fee
```

```
Extracted Text (165 segments):

1. Your actual rate, payment; and cost could
2. FEES WORKSHEET
3. Fee Details and Summary
4. Applicants:
5. John Q. Smith
6. A. Smith
7. Application No:
8. samplesmith
9. Prepared By:
10. XYZ Lender
11. Date Prepared:
12. 10/05/2015
13. Loan Program:
14. 30 YEAR FIXED -Purchase
15. THIS IS NOT
16. A GOOD FAITH ESTIMATE (GFE):
17. This
18. "Fees Worksheet" is provided for informati
19. to assist
20. you in determining
21. an estimate of cash that may be required t
22. and
23. an estimate of your proposed monthly mortg
```

```
{
  "FEES": {
    "text": "FEES",
    "bounding_box": [
      952,
      180,
      140,
      40
    ]
  },
  "LENDER": {
    "text": "LENDER",
    "bounding_box": [
      708,
      1664,
      112,
      37
    ]
  }
}
```

PaddleOCR

EasyOCR

Tesseract

# Recommendation – Paddle OCR

- Recommended engine for the production pipeline
  - Superior performance in accuracy and layout awareness on complex, structured documents
    - (i.e., mortgage forms)
    - Extract text
    - Structure and recognize a higher volume of necessary data points in challenging document layouts
  - Ensures maximum data integrity



# CoLab Notebook

- Review the Extracting text & Bounding Boxes from Scanned PDFs: [HERE](#)
  - Data sample\_mortgage\_document.pdf: [HERE](#)
- Review the Analyze a Scanned PDF (End to End): [HERE](#)
  - Data MTG\_10009588.pdf: [HERE](#)
- Review the Layout OCR Demo: [HERE](#)
  - Data Loan Fees WorksheetNew-2: [HERE](#)
- Review the Compare 3 OCR Engines on a Mortgage PDF: [HERE](#)
  - Data LenderFeesWorksheetNew: [HERE](#)

---

THANK YOU

---

<https://github.com/LashawnFofung/AI-Powered-Document-Automation-Platform>