

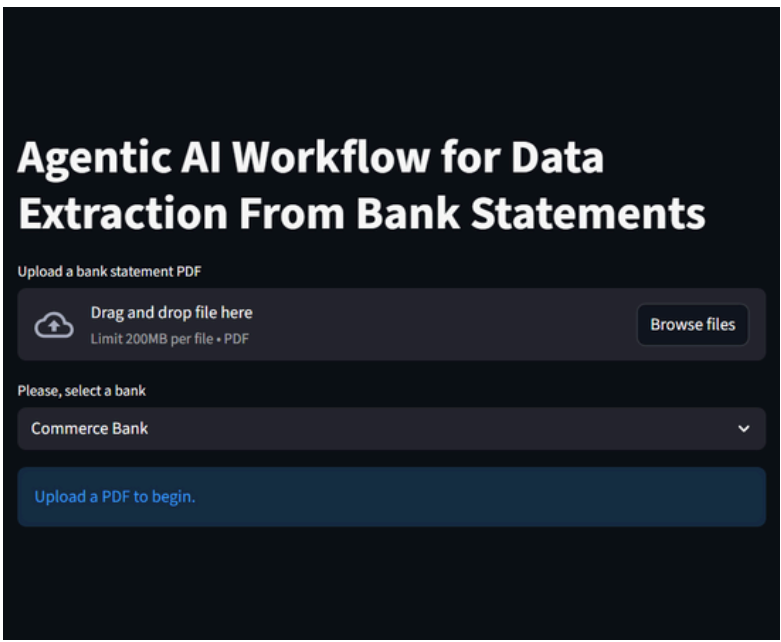


AGENTIC AI WORKFLOW FOR DATA EXTRACTION FROM BANK STATEMENTS PDFs

USING OPEN AI GPT 4.1 MINI

Developed by Angel A. Barrera

USER INTERFACE (UI)



- 1 Upload PDF (Bank statement)
- 2 Extract & Validate

ROUTER AGENT

pdf_router.py

Decides whether the PDF is machine-readable (TEXT) or scanned/image-based (VISION)



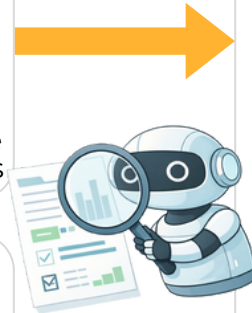
If the PDF is
TEXT or **VISION**

VISION OCR AGENT

OpenAI Vision

Render PDF pages to PNG
via PyMuPDF

This converts scanned
statements into AI-readable
images for accurate analysis



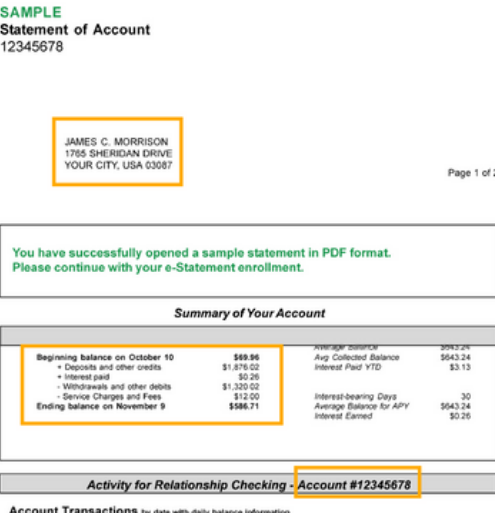
OCR EXTRACTION AGENT

orchestrator_vision.py

Converts document images
into readable text for data
extraction

PyMuPDF

PDF → Image renderer



REFLECTION AGENT

Validate each extracted field
against the source

PASS / FAIL / UNCERTAIN



Agent job is to explain its reasoning for the decision

status	extracted_value	evidence_quotes	suggested_value	confidence	reason
PASS	12345678	Statement of Account 12345678	12345678	1	Account number '12345678' found clearly in the statement headings.
PASS	11/09	Ending balance on November 9	11/09	1	The statement ending date corresponds to November 9, which matches extracted date.
PASS	49.96	Beginning balance on October 10	49.96	1	Beginning balance exactly matches statement text line.
PASS	1,876.02	+ Deposits and other credits	1,876.02	1	Deposits and other credits total matches the explicit summary line.
PASS	1,320.02	- Withdrawals and other debits	1,320.02	1	Withdrawals and other debits total matches the explicit summary line.
PASS	12.00	- Service Charges and Fees	12.00	1	Service charges and fees total is clearly stated in the summary section.
PASS	586.71	Ending balance on November 9	586.71	1	Ending balance matches the labeled total balance at statement end.
PASS	SAMPLE	CONFID: SAMPLE	SAMPLE	1	Bank name is configured from UI selection.

FAIL / UNCERTAIN

Re-validate

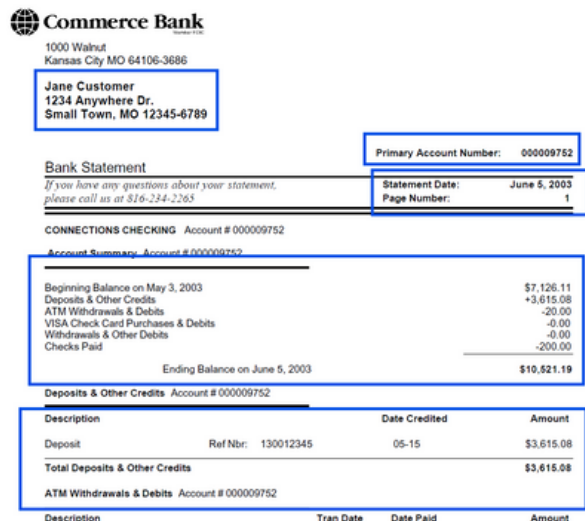
TEXT EXTRACTION AGENT

orchestrator_text.py

PDFPlumber reads and
extracts the values from the
PDF

Schema Extraction

PDFPlumber



REPAIR AGENT

- Applies suggestions given by the Reflection Agent
- Automatically corrects issues and refines extracted data
- Sends it back for review



FINAL OUTPUTS

- Produced a structured table (downloadable as a .csv/Excel file)
- Validation report
- Metadata
- Can be connected to a database



PyMuPDF
MuPDF with Python



IMAGE-BASED PDF RESULTS

bank_name	account_num	statement_date	beginning_balance	total_deposits_and_credits	total_withdrawals_and_debits	total_service_charges_fees	ending_balance
SAMPLE	12345678	9-Nov	69.96	1,876.02	1,320.02	12	586.71

bank_name	customer_name	customer_address	account_number	statement_date	initial_balance	deposits_and_credits	ATM Withdrawals Debits	VISA Check Card Purchases Debits	Withdrawals Other Debits	total_checks_paid	ending_balance
SAMPLE	Jane Customer	1234 Anywhere Dr., Small Town, MO 12345-6789	9752	5-Jun-03	7,126.11	3,615.08	20	0	0	305	10,521.19

TEXT-BASED PDF RESULTS

bank_name	customer_name	customer_address	account_number	statement_date	initial_balance	deposits_and_credits	ATM Withdrawals Debits	VISA Check Card Purchases Debits	Withdrawals Other Debits	total_checks_paid	ending_balance
Commerce Bank	Jane Customer	1234 Anywhere Dr., Small Town, MO 12345-6789	9752	5-Jun-03	7,126.11	3,615.08	20	0	0	305	10,521.19