

Intelligent Invoice & Document Processor

Project Overview

SMBs spend enormous time on manual data entry from invoices, purchase orders, and receipts. Finance teams key the same data into accounting systems by hand — introducing errors, delaying cash flow visibility, and burning hours that should go to higher-value work.

Designed and developed an Intelligent Invoice & Document Processor that automatically extracts, classifies, and validates data from invoices, purchase orders, and expense receipts using OCR and large language models — posting structured, validated results directly into the client's accounting system.

The solution handles the full document lifecycle: from ingestion of scanned PDFs, mobile photos, and email attachments through to confident, rule-validated posting — with a human-in-the-loop review queue catching anything the model flags as low-confidence.

Architecture & Solution

Documents are ingested via email, upload, or automated inbox monitoring and processed through an OCR pipeline that normalises varying layouts, languages, and scan quality. Extracted fields are passed to an LLM layer that classifies document type, extracts structured data, and enriches records with context from the client's vendor and chart-of-accounts data.

A validation engine checks extracted data against configurable business rules — duplicate detection, amount thresholds, vendor whitelist checks — before posting to QuickBooks, Xero, or a custom ERP via REST API. Low-confidence extractions are routed to a human review queue, keeping humans in the loop without slowing down the majority of documents that clear automatically.

Business Impact

The solution dramatically reduced manual data entry effort, shortened invoice processing cycles, and improved accuracy in accounting records. Finance teams shifted from keying data to reviewing exceptions, enabling significantly higher document volumes without adding headcount. Faster processing also improved cash flow visibility and reduced late-payment risk.

Key Capabilities

- OCR-based document ingestion (PDFs, photos, email attachments)
- LLM-powered field extraction and document classification
- Business rule validation and duplicate detection
- Automated posting to QuickBooks, Xero, and custom ERPs
- Human-in-the-loop review queue for low-confidence extractions
- Vendor and chart-of-accounts enrichment
- Audit trail and processing analytics
- Configurable extraction rules per document type

Technologies

Python, Azure Document Intelligence, OpenAI APIs, LangChain, REST APIs, Docker, OAuth 2.0, Prompt Engineering.