Document Intelligence Pipeline for Legal Analytics

Group 13

High-Level Overview

The Problem

- Manual review of invoices, contracts, emails & minutes
- Time-consuming, error-prone data extraction
- Inconsistent formats → inconsistent data quality
- Key information buried in PDFs → weak search & reporting

Tech Stack (PoC)

python

pdfplumber

Pydantic

Qwen 2.5 (7B)
[Local LLM]

pandas

Our Solution (PoC)

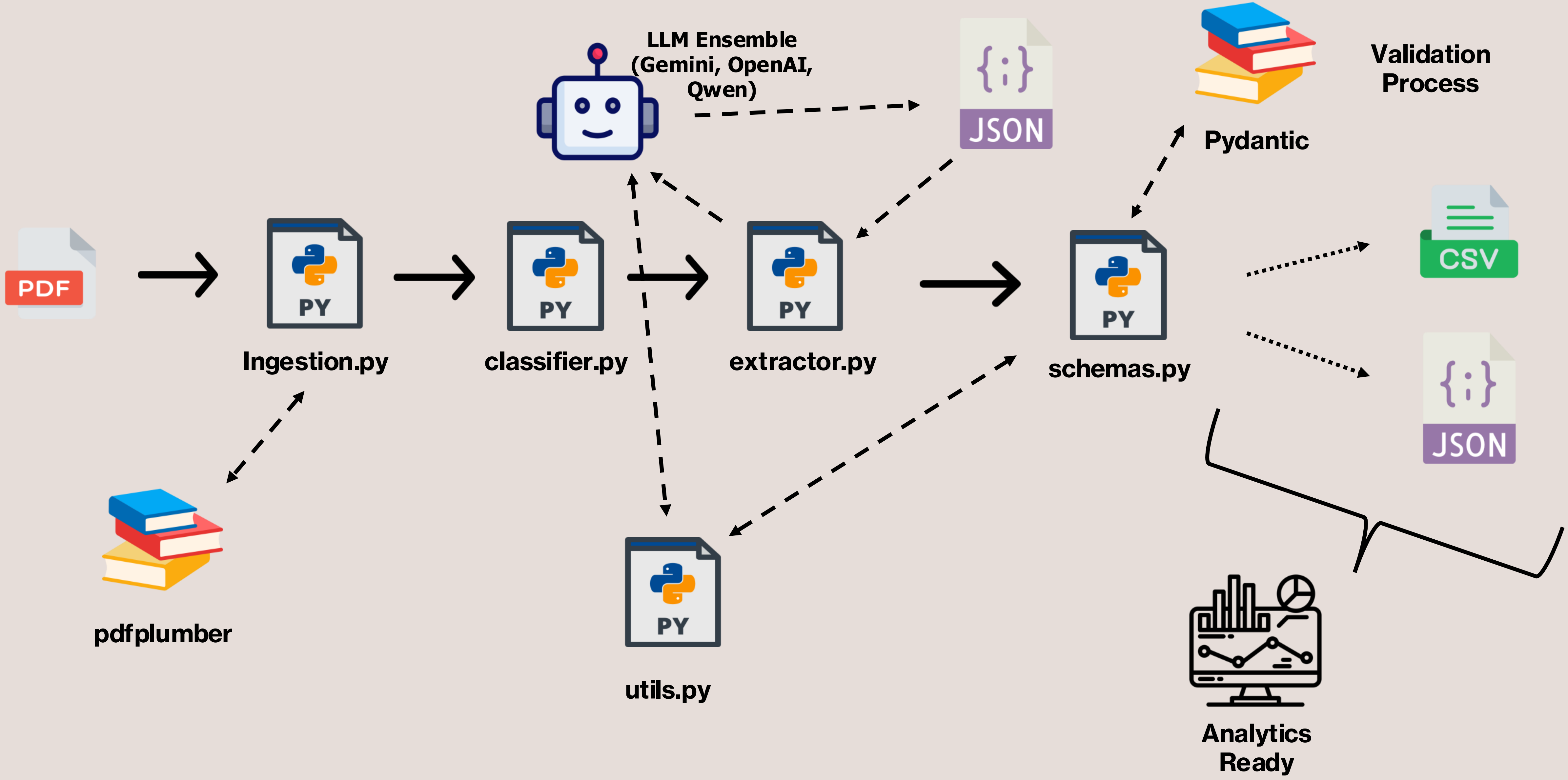
Automated document intelligence Pipeline that:

- Ingests PDF documents from different sources
- Classifies them by type (invoice, contract, email, meeting minutes)
- Uses an **ensemble of LLMs** (OpenAI, Gemini, Qwen) to extract key fields into structured JSON
- Validates & standardizes data with strict schemas through **multi-agent orchestration**
- Stores outputs for **reporting, search, summarization & aggregation**

Platform-Level View

1. *Inputs:* PDF documents (invoices, contracts, emails, meetings minutes)
2. *Core platform:*
   - Ingestion layer: Read PDF & extract text
   - AI layer: Classifies documents + extract key fields (LLM)
   - Validation layer: cleans, normalizes, and validates data with schemas
   - Storage layers: saves structured data (JSON/CSV)
3. *Outputs:* Search | Reporting | Summaries | Analytics

Detailed Approach to Data Pipeline (Local)



Proposed Technical Architecture Beyond POC

Production Architecture

Data Lakehouse & Orchestration

- Multi-source ingestion (PDFs, emails, shared drives)
- S3 / Azure Blob storage for raw documents
- Airflow / Prefect for workflow orchestration (poll-based)
- Azure and GCP for access to the latest OpenAI and Gemini models while maintaining secure governance
- Batch processing to minimize costs w/o impacting value

Parallel Processing & AI Services

- Apache Spark to parallelize processing
- LLM multi-agent orchestration for classification & extraction
- Top-grade GPUs for inference

Storage & Analytics

- SQL Data Warehouse
- Audit-ready log storage
- Agentic chart production for targeted queries through MCPs with SQL functions in a chat UI

Downstream Applications

Financial Reporting

- Vendor-level spend aggregation
- Trend analysis & forecasting

Smart Search

- “Find all WeWork invoices”
- Instant semantic search & filters

Compliance

- Track contract expiry / deadlines
- Automated alerts & reminders

Legal Research

- Cross-document linking by party names
- Case-related document grouping

Summarization

- “Show all invoices related to X in July”
- Automated monthly summaries

Thank you!