AI-Powered Invoice Automation

A prototype that automates PDF invoice processing into JSON using OCR, LLMs, and vector search.

Students

Noah Adrian Fleischmann

Dominik Rüegg

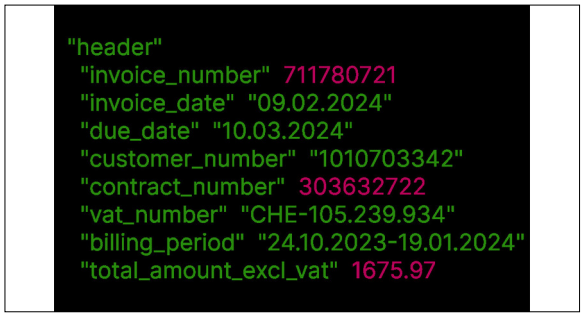
Introduction: Etter Consulting Partners (ECP) processes invoices from hundreds of energy and utility providers manually. The lack of standardization and the diversity of formats make extracting data into a unified structure a time-consuming and error-prone task. This project addresses these challenges by developing a prototype to automate the extraction of detailed information from PDF invoices and convert it into a structured, machine-readable format. The extracted data is formatted as JavaScript Object Notation (JSON), facilitating seamless integration into subsequent analytical processes.

Approach / Technology: The prototype consists of a Python program that integrates multiple technologies to tackle the complexities of invoice processing. Docling OCR plays a crucial role in converting diverse PDF formats, including scanned invoices, into structured Markdown, forming the foundation for data extraction. Large Language Models (LLMs) are central to the process, identifying invoice providers and extracting key details from the text. QR code recognition complements this by directly extracting provider information and relevant data embedded within QR codes when available. Additionally, vector similarity search is employed to identify the invoice provider by comparing document embeddings to known provider profiles. Each component was optimized to address specific challenges. A dataset of manually parsed invoices from ECP served as the benchmark for evaluating the pipeline's accuracy and reliability.

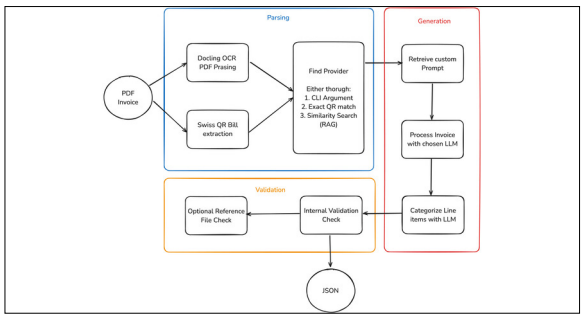
Result: The project successfully achieved its primary objective of accurately parsing invoice data and converting it into structured JSON. Both Claude Sonnet 3.5 and OpenAI GPT-4 demonstrated strong performance, with a best-case accuracy of up to 94% in specific cases, and overall accuracies of 66.66% and 63.66%, respectively. Llama 3.3-70b reached an overall accuracy of 60.68%. These accuracy metrics were determined using a custom scoring system developed for this project and validated across a large dataset, confirming the software's capability to reliably automate data extraction.
Challenges were identified in the categorization of line items, where the LLM occasionally assigned incorrect categories due to limited contextual information. Providing additional data for each possible category is expected to improve categorization accuracy. Especially, the prompt plays a significant role in influencing results: with better prompts, the system's accuracy improves substantially.

The project not only delivered a robust prototype but also highlighted key areas for further refinement such as augmenting category information and meticulous tuning of each provider-specific prompt, enhancing system scalability and precision. The final solution is a functioning Python CLI, supporting parsing through OpenAI GPT, Anthropic Claude, and Llama, and can be used in any environment capable of executing Python.

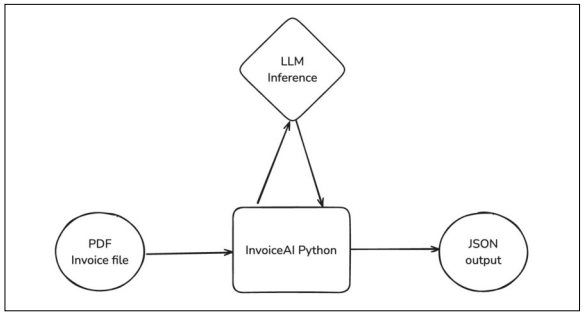
Sample JSON
Own presentment



Flowchart
Own presentment



Simplified Architecture
Own presentment



Advisors
Prof. Dr. Mitra Purandare, Benjamin Plattner

Subject Area
Artificial Intelligence, Software

Project Partner
etter consulting partners AG, Zürich, ZH