Vol-2 Issue-2 2024
Scientific Research Journal of Science, Engineering and Technology
ISSN: 2584-0584, Peer Reviewed Journal

Case Study: Centralising Diverse E-Commerce Invoices Using Invoice LLM Model

Dharti Patel, Research Scholar, Computer Engineering, Sardar Patel University, Vallabh Vidya Nagar (Anand)

Dr. H. B. Pandit, Professor, Computer Engineering, Sardar Patel University, Vallabh Vidya Nagar (Anand)

Abstract

E-commerce platforms handle various invoices, including PDFs, handwritten documents, and scanned JPG images. This diversity in invoice formats presents significant challenges when centralising data for accounting and tax purposes. Manual processing leads to operational inefficiencies and limits scalability. This case study discusses how integrating an ***Invoice LLM (Large Language Model)****, combined with* ***Optical Character Recognition (OCR)*** *for handwritten and scanned invoices, helps extract and centralise key entities from different formats, reducing manual intervention and increasing operational efficiency.*

Keywords: E-commerce, Invoice Processing, Large Language Model (LLM), Optical Character Recognition (OCR), Data Extraction, Centralized Data Management, Automation, Scalability

Background

The e-commerce industry handles invoices from thousands of vendors, each with different templates, formats, and delivery methods. These invoices can appear in the following formats:

- **PDF**: Most vendors generate digital invoices in PDF format. These include complex layouts with tables, multiple columns, and embedded images.
- **Handwritten**: Smaller vendors may submit handwritten invoices, which add complexity due to inconsistencies in handwriting styles and legibility.
- **JPG/Scanned**: Some invoices are scanned into JPG format as images of physical invoices or snapshots sent via email or messaging apps.

The lack of a standardised invoice format means the company must process various data inputs. This requires extracting entities from text in PDFs, deciphering handwritten invoices using OCR, and interpreting data from image files like JPGs.

Problem Statement

The diversity in invoice formats complicates extracting consistent data for financial records. Manual intervention is not scalable as the platform grows, and handling different formats—especially handwritten and scanned invoices—requires an automated solution that can effectively recognise and centralise data.

Challenges

1. **Handling Multiple Formats**: The company receives invoices in various formats, such as PDFs, handwritten paper invoices, and image files. Each format presents unique challenges in terms of extraction and centralisation.
2. **Manual Processing**: Employees manually input data from handwritten and scanned invoices, leading to increased error rates and inefficiencies.
3. **OCR for Handwritten Invoices**: Handwritten invoices pose a unique challenge due to varying handwriting styles, which can affect OCR performance.
4. **Inconsistent Data Layouts**: Even digital PDFs may vary in layout, requiring a solution that can interpret different document structures and extract critical information.

Solution: Leveraging Invoice LLM with OCR Integration

To tackle the challenge of diverse invoice formats, the company implemented a two-part solution:

1. **Optical Character Recognition (OCR)**: OCR technology converts handwritten and image-based invoices (JPGs) into machine-readable text. This allows for the extraction of key invoice details from non-digital sources.
2. **Invoice LLM Model**: The LLM model processes the OCR-extracted text from handwritten and scanned invoices, as well as text directly from PDF files, to identify and extract critical data such as invoice numbers, vendor names, item details, tax information, and totals. The extracted data is then mapped to a standardised format for further use.

Implementation Process

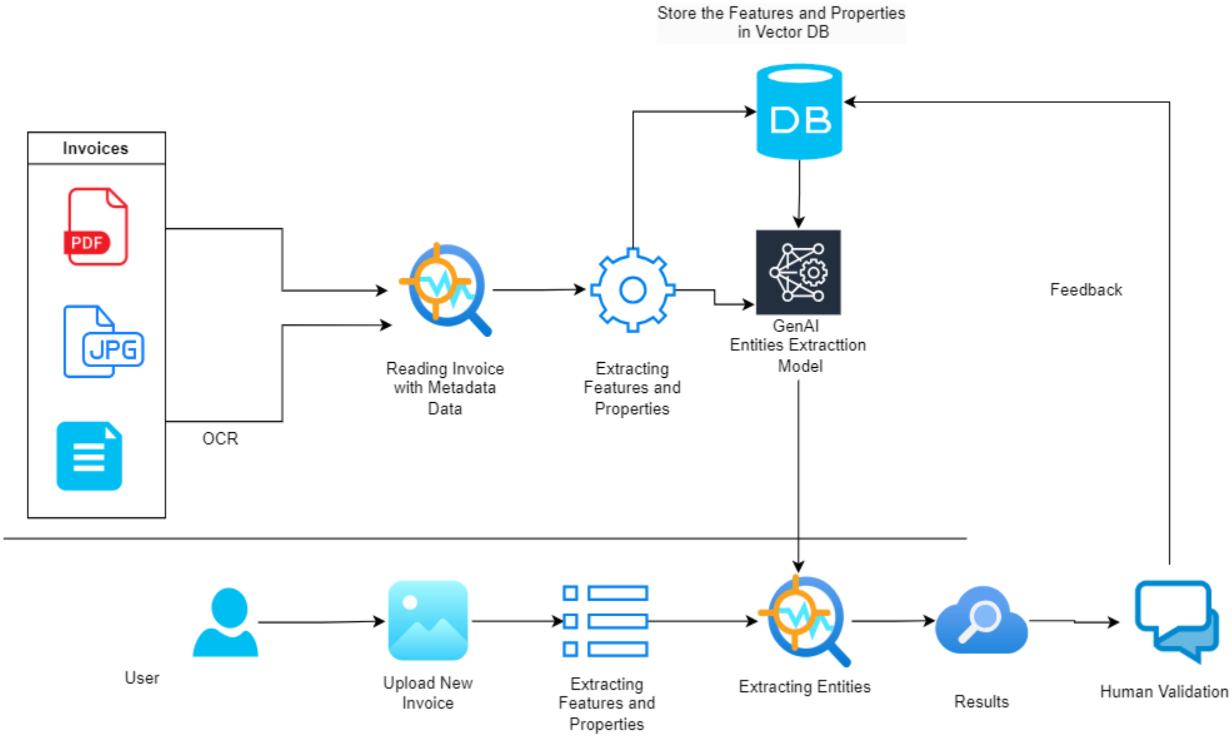
Store the Features and Properties in Vector DB
DB
Invoices
PDF
JPG
OCR
Reading Invoice with Metadata Data
Extracting Features and Properties
GenAI Entities Extracttion Model
Feedback
User
Upload New Invoice
Extracting Features and Properties
Extracting Entities
Results
Human Validation

Figure: Architecture Diagram

The diagram you shared represents a system for centralising diverse e-commerce invoices using an OCR (Optical Character Recognition) process combined with a GenAI (Generalized Artificial Intelligence) model for extracting key entities from different formats of invoices.

Here's a breakdown of the process:

1. **Input Formats (Invoices in PDF, JPG, and Document Files):**
   - The system accepts invoices in formats like PDFs, scanned images (JPG), and document files.
   - These invoices may have different layouts and structures, complicating manual data extraction.
2. **OCR (Optical Character Recognition):**
   - OCR technology is applied to handwritten or scanned invoices (JPGs and images). It converts the non-digital text into machine-readable data.
   - This step ensures that even handwritten or image-based invoices can be processed digitally for further extraction.
3. **Reading Invoice with Metadata:**
   - After the OCR process, the system reads the invoice data, including the metadata (information like date, vendor, etc.), to ensure accurate extraction of features.
4. **Extracting Features and Properties:**
   - The system extracts essential features and properties from the invoices, such as invoice numbers, vendor details, tax amounts, and line items (product descriptions, quantities, etc.).
   - This extraction is necessary to standardise the data from different invoice formats.
5. **GenAI Entity Extraction Model:**
   - A GenAI model is employed to extract entities from the digitised text automatically. It identifies vital information (such as the invoice number, customer details, payment info, etc.) across varying invoice layouts and formats.
   - The AI model is trained on various invoice types to recognise patterns and structures, ensuring reliable data extraction.
6. **Storing in Vector Database:**
   - Once the entities are extracted, the features and properties are stored in a vector database. This database allows easy access to the extracted data and enables faster searches for future use in accounting, audits, and analytics.
7. **Human Validation:**
   - After the automated process, human validation is introduced for quality assurance. If there are any discrepancies or uncertainties in the extraction (for example, with unclear handwriting or ambiguous data), human intervention ensures the correctness of the extracted data.
8. **Feedback Loop:**
   - Feedback from the human validation process is fed back into the GenAI model to improve its performance continuously. This enables the model to learn from its mistakes and better recognise and extract information over time.
9. **End Results and Output:**
   - After successful validation and extraction, the system returns accurate, structured data to the user. This data can be used for various business purposes, such as reporting, financial tracking, and tax compliance.

Results

- **Efficiency Gains**: The automated extraction process, including OCR for handwritten invoices, reduced manual effort by 90%, significantly lowering the time spent on processing invoices from diverse formats.
- **Increased Accuracy**: Error rates in data extraction from handwritten and scanned invoices dropped by 75% compared to manual processing. The LLM model provided high accuracy across different invoice formats.
- **Scalability**: The company can process tens of thousands of invoices daily, regardless of format, without additional labour costs.
- **Data Centralisation**: All extracted data is stored in a centralised database, allowing for streamlined reporting, tax compliance, and financial audits.

Use Case Example

A significant e-commerce platform dealing with over 200,000 vendors worldwide receives invoices in formats ranging from neatly structured PDFs to photos of handwritten invoices sent via messaging apps. Before implementing the Invoice LLM and OCR solution, manual data extraction was slow and prone to errors. After adopting this solution, the company could process 100,000 invoices daily while improving handwritten invoice processing accuracy by 70%.

Conclusion

Combining OCR for handwritten and scanned invoices with an LLM-based entity extraction model offers an effective solution for e-commerce companies with diverse invoice formats. This approach automates the extraction process, enhances accuracy, and enables the centralisation of invoice data, allowing businesses to scale their operations efficiently. By adopting this technology, e-commerce companies can streamline their financial workflows, reduce manual labour, and ensure compliance with reporting standards.

References

1. Desai, Devanshi & Jain, Ansh & Naik, Dhaivat & Panchal, Nishita & Sawant, Dattatray. (2021). Invoice Processing using RPA & AI. SSRN Electronic Journal. 10.2139/ssrn.3852575..
2. D. Baviskar, S. Ahirrao, V. Potdar and K. Kotecha, "Efficient Automated Processing of the Unstructured Documents Using Artificial Intelligence: A Systematic Literature Review and Future Directions," in IEEE Access, vol. 9, pp. 72894-72936, 2021, doi: 10.1109/ACCESS.2021.3072900.
3. Saout, Thomas & Lardeux, Frédéric & Saubion, Frédéric. (2024). An Overview of Data Extraction From Invoices. IEEE Access. PP. 1-1. 10.1109/ACCESS.2024.3360528.
4. Saout, T., Lardeux, F., & Saubion, F. (2024). An Overview of Data Extraction From Invoices. IEEE Access.
5. Bardvall, M., & Hassle, I. (2024). Automating Invoice Recognition: A Comparative Study of Large Language Models and OCR/ML Technologies.
6. Daqqah, B. H. (2024). Leveraging Large Language Models (LLMs) for Automated Extraction and Processing of Complex Ordering Forms (Doctoral dissertation, Massachusetts Institute of Technology).