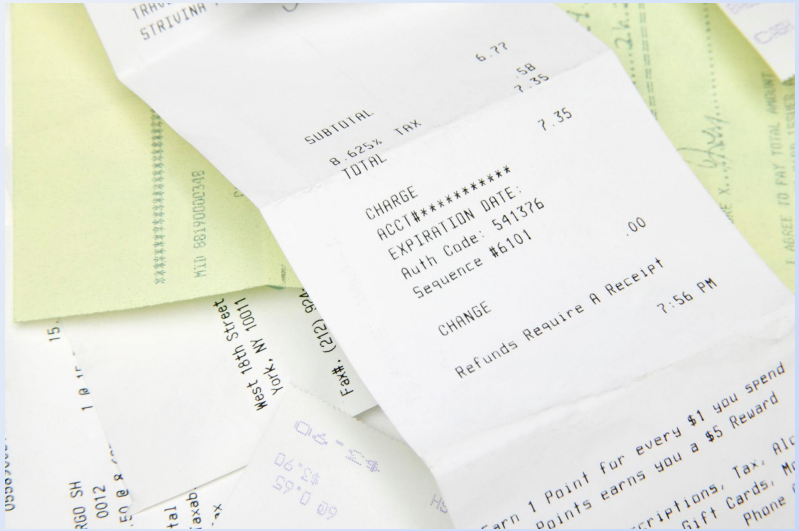
BillBuddy: Receipt and Invoice Digitizer

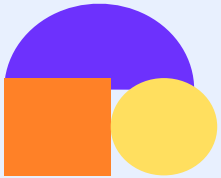
Problem Statement

- System should automatically scans, extracts, and digitizes information using:
  - OCR (Optical Character Recognition)
  - NLP-based field extraction.
- Digitized data is stored in a structured format,
  - easy to search, analyze
  - Integratable with accounting or ERP systems.



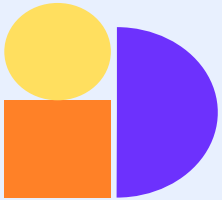
Existing Solutions and Limitations

- **Veryfi** leads in speed (2.8s average) and accuracy (98.7%),
- **Google Cloud Vision** prioritizes cost efficiency ($320/1000 docs).
- **Rossum and Nanonets** excel in AI/ML-based extraction for unstructured data.
- **Koncile** claiming near-100% field accuracy via computer vision and LLMs.



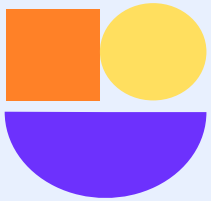
Accuracy Challenges

Existing systems misread 1 in 10 characters



Handling Variability

Layouts vary wildly by vendor, country, and format.



Manual Intervention and Costs

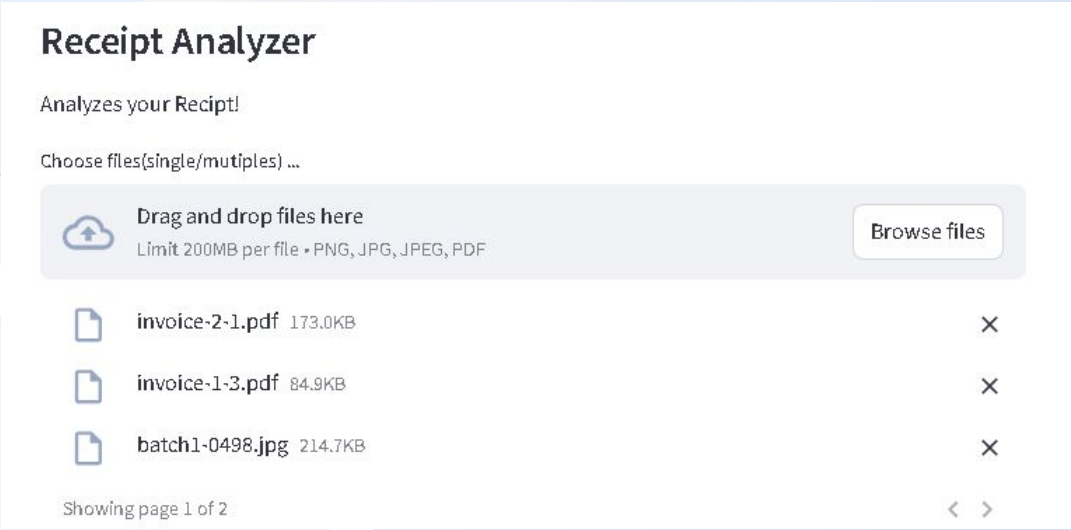
Even top tools like standalone OCR demand regular manual corrections for 10-15% of documents.

Proposed Solution

- AI-driven system that automates the digitization and processing of physical and digital receipts and invoices using:
  - Optical Character Recognition (OCR)
  - Natural Language Processing (NLP),
- System combines computer vision (OpenCV), OCR (Tesseract), and lightweight NLP
- Transform unstructured documents into structured, actionable insights user-friendly web interface.

Input & Processing: OCR Engine

- Robust multi-stage OCR pipeline
- Supports both images and PDFs
    - Uploaded files converted to grayscale images,
    - Reduces noise using **Gaussian blur**
    - Uses **Adaptive thresholding** for clean black-and-white images.
    - PDFs processed using **pdf2image (200 DPI)**
    - Text extracted using **Tesseract OCR (PSM mode 1)**
    - Regex-based parsing for structured fields.



Extracted Text

Invoice no: 22529030

Date of issue: 02/02/2018

—— Seller: Client:
Hogan, Adkins and Rogers Johnson LLC
2917 Young Cape 3913 Frank Inlet
South James, PA 13959 Port Kathy, GA 28821
Tax Id: 960-87-3385 Tax Id: 938-83-8149

IBAN: GB480ZWQ30024053292452

——— ITEMS

No. Description Qty UM. Net price Net worth VAT [%]

✅ Invoice saved to History Vault!

Seller:

Hogan, Adkins and Rogers
2917 Young Cape
South James, PA 13959

Tax Id: 960-87-3385
IBAN: GB48OZWQ30024053292452

Client:

Johnson LLC
3913 Frank Inlet
Port Kathy, GA 28821

Tax Id: 938-83-8149

ITEMS

No.	Description	Qty	UM	Net price	Net worth	VAT [%]	Gross worth
1.	PS2 Slim Console System SCPH-77000 PINK Playstation 2 Tested	4.00	each	252.00	1 008.00	10%	1 108.80
2.	Nintendo Wii Console With Cables And Accessories And New Super Mario Bros Disc!	5.00	each	29.99	149.95	10%	164.94
3.	Nintendo Switch Gaming Console W/ Neon Blue & Red Controllers,Carry Case,Cable	1.00	each	300.00	300.00	10%	330.00
4.	Sony PlayStation Gray PS1 Console! CONSOLE ONLY! Tested Working!	2.00	each	21.99	43.98	10%	48.38
5.	B408 Nintendo Gameboy Color console Ice Blue Toysrus Limited Japan GBC REAR x	4.00	each	74.99	299.96	10%	329.96
6.	Nintendo DS Lite Onyx Black System Console & 6 games Mario Charger Stylus	1.00	each	69.99	69.99	10%	76.99
7.	Nintendo 64 N64 Game Console System + Controller Cords WORKING	3.00	each	100.00	300.00	10%	330.00

SUMMARY

	VAT [%]	Net worth	VAT	Gross worth
	10%	2 171.88	217.19	2 389.07
Total		$ 2 171.88	$ 217.19	$ 2 389.07

Original

Seller:

Hogan, Adkins and Rogers
2917 Young Cape
South James, PA 13959

Tax Id: 960-87-3385
IBAN: GB48OZWQ30024053292452

Client:

Johnson LLC
3913 Frank Inlet
Port Kathy, GA 28821

Tax Id: 938-83-8149

ITEMS

No.	Description	Qty	UM	Net price	Net worth	VAT [%]	Gross worth
1.	PS2 Slim Console System SCPH-77000 PINK Playstation 2 Tested	4,00	each	252,00	1 008,00	10%	1 108,80
2.	Nintendo Wii Console With Cables And Accessories And New Super Mario Bros Disc!	5,00	each	29,99	149,95	10%	164,94
3.	Nintendo Switch Gaming Console W/ Neon Blue & Red Controllers,Carry Case,Cable	1,00	each	300,00	300,00	10%	330,00
4.	Sony PlayStation Gray PS1 Console! CONSOLE ONLY! Tested Working!	2,00	each	21,99	43,98	10%	48,38
5.	B408 Nintendo Gameboy Color console Ice Blue Toysrus Limited Japan GBC REAR x	4,00	each	74,99	299,96	10%	329,96
6.	Nintendo DS Lite Onyx Black System Console & 6 games Mario Charger Stylus	1,00	each	69,99	69,99	10%	76,99
7.	Nintendo 64 N64 Game Console System + Controller Cords WORKING	3,00	each	100,00	300,00	10%	330,00

SUMMARY

	VAT [%]	Net worth	VAT	Gross worth
	10%	2 171,88	217,19	2 389,07
Total		$ 2 171,88	$ 217,19	$ 2 389,07

Preprocessed

Data Storage & Persistence

- All results stored in a **user-scoped SQLite database**
- Ensures data privacy and session persistence
- Structured extracted fields
- Validation status and metadata
- Invoices saved automatically after upload
- Vault management at Ease.

User's History

↑ id	filename	vendor	invoice_id	date	total_amount	upload_time
1	batch1-0001.jpg	Invoice		04/13/2013		2026-01-14 11:31
2	batch1-0002.jpg	Invoice		03/03/2012		2026-01-14 11:32
3	batch1-0003.jpg	Invoice		04/09/2014		2026-01-14 11:32
4	batch1-0001.jpg	Invoice		04/13/2013		2026-01-14 11:36
5	batch1-0002.jpg	Invoice		03/03/2012		2026-01-14 11:36
6	batch1-0003.jpg	Invoice		04/09/2014		2026-01-14 11:36
7	invoice-1-3.pdf					2026-01-14 11:38
8	batch1-0497.jpg	Invoice		07/25/2017		2026-01-14 11:42

View details

Select an invoice...

Select invoices to delete:

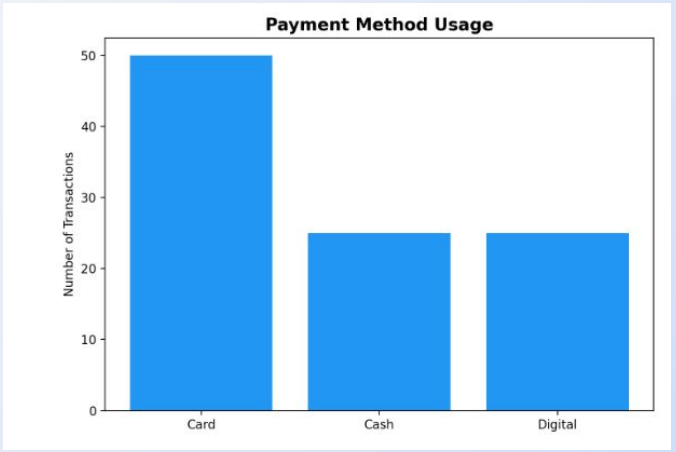
Choose options

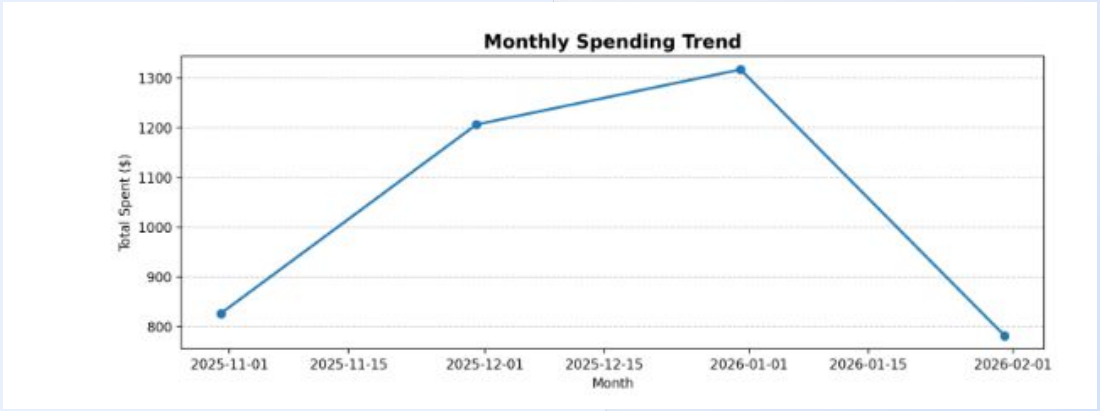
Delete Selected

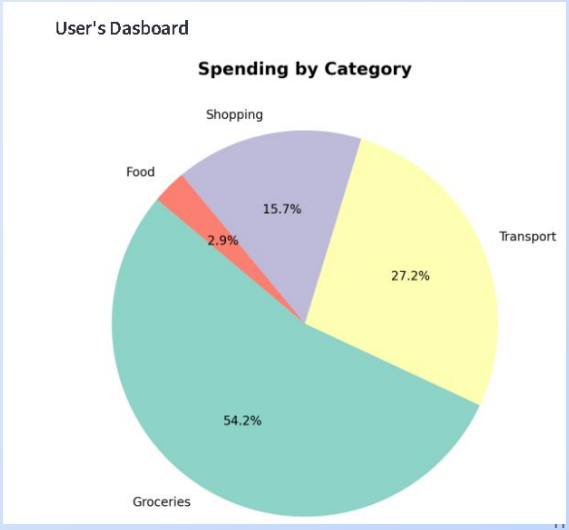
User Interface & Data Visualisation

- Built using **Streamlit.**
- Converts processed invoices into insights
- Fully automated analysis pipeline.
- Charts used:
    - **Bar Graph, Line Graph**, **Pie Charts**
- Integrated **Gemini-powered chat interface** for conversational support


Login
Username
demo_user
demo_user
Go to
Home
Analyzer
Dashboard
History
Support
Settings
Receipt Invoice and Digitizer
About the Project
The Receipt and Invoice Digitizer is an AI-driven system that automates the digitization and processing of physical and digital receipts and invoices using Optical Character Recognition (OCR) and Natural Language Processing (NLP), reducing manual effort and improving accuracy in financial operations.


Payment Method Usage
Number of Transactions
0
10
20
30
40
50
Card
Cash
Digital


Monthly Spending Trend
Total Spent ($)
800
900
1000
1100
1200
1300
2025-11-01
2025-11-15
2025-12-01
2025-12-15
2026-01-01
2026-01-15
2026-02-01
Month


User's Dasboard
Spending by Category
Shopping
Food
Transport
Groceries
15.7%
2.9%
27.2%
54.2%

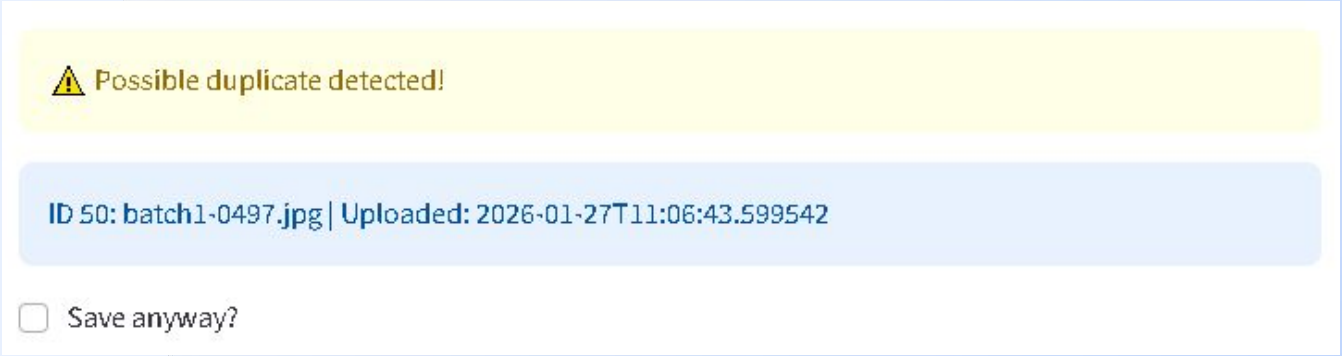
Validation & Duplication:

1. Validates total amount plausibility (e.g., > $0, reasonable range):

   Field match (same vendor + invoice ID + total ±2%)

2. Detects duplicates using:

   Image hash (same visual document).



Field Extraction

- Uses regex patterns for structured data (invoice ID, amounts)
- Enhances with spaCy NLP to identify vendors and dates from context.
- Enhances with spaCy NLP to identify vendors and dates from context.
- Handles OCR noise by normalizing text (fixing “O” vs “0”, “l” vs “1”) and using flexible patterns.

15	invoice-2-1.pdf	Date	0012820		10	2026-01-14 13:08
50	batch1-0497.jpg	Bcbgeneration Black Sleeveless		06055	615,88	2026-01-27 11:06
51	invoice-1-3.pdf	El Yunque Colorado Qty		1213		2026-01-27 11:09

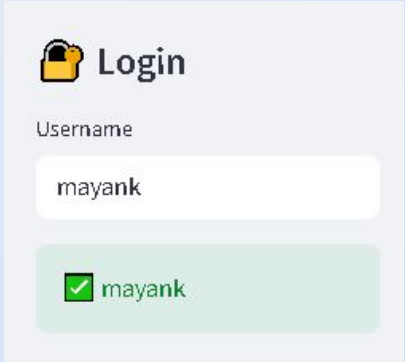
Structured Storage

Saves raw file, extracted fields, validation status, and user notes in SQLite

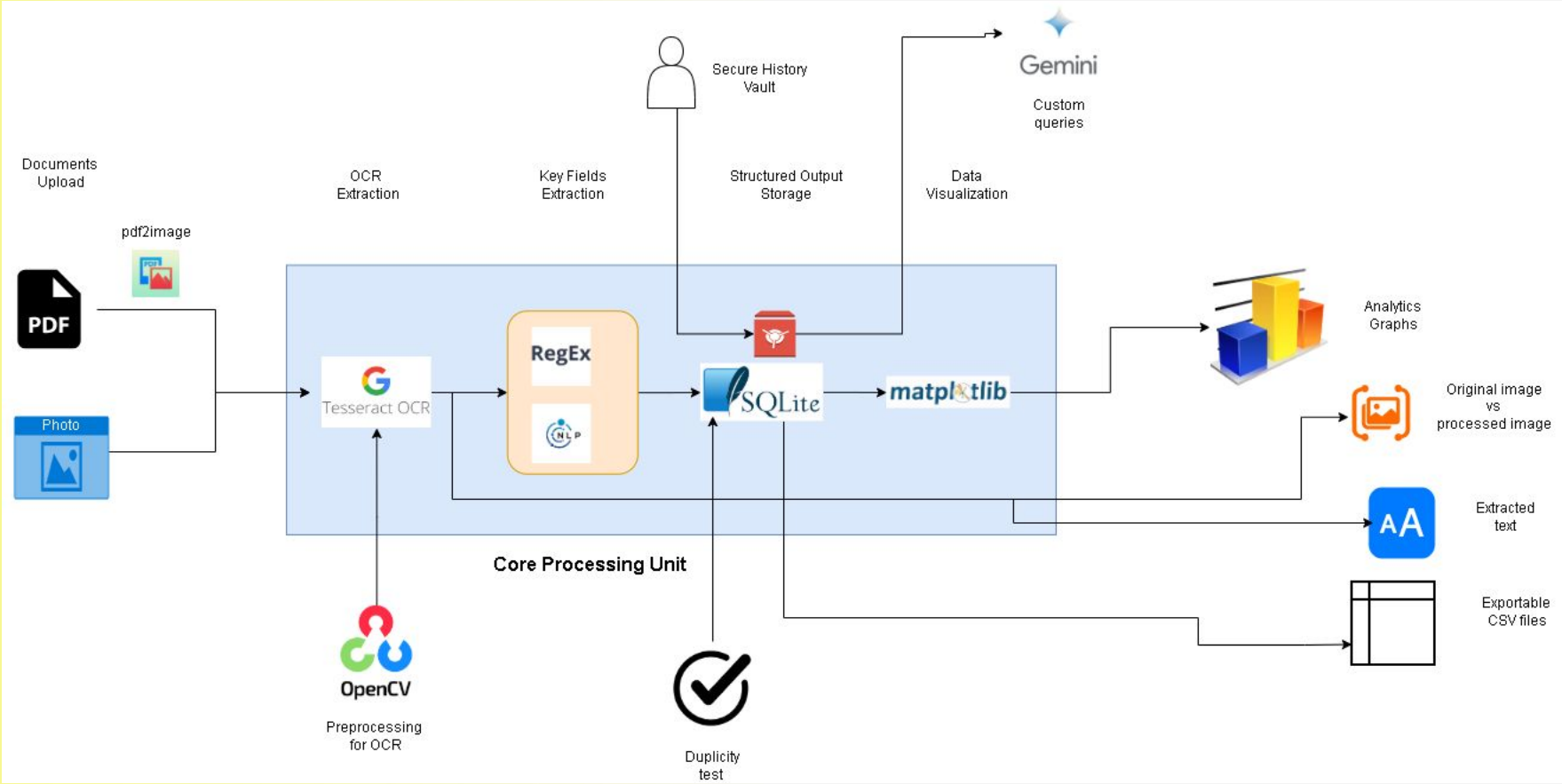
User can view the items in n line format..

id	filename	vendor	invoice_id	date	total_amount	upload_time
12	batch1-0497.jpg	Invoice		07/25/2017		2026-01-14 13:07
13	batch1-0498.jpg	Invoice		01/23/2020		2026-01-14 13:08
14	invoice-1-3.pdf					2026-01-14 13:08
15	invoice-2-1.pdf	Date	0012820		10	2026-01-14 13:08
50	batch1-0497.jpg	Bcbgeneration Black Sleeveless		06055	615,88	2026-01-27 11:06
51	invoice-1-3.pdf	El Yunque Colorado Qty		1213		2026-01-27 11:09

Each user gets an isolated, persistent history vault



Architecture Diagram



Tools and resources

Frontend

Streamlit

Core Libraries and Frameworks

- OpenCV
- Tesseract OCR + pytesseract
- Pdf2image
- Pillow (PIL)
- Matplotlib, Seaborn
- Regex

Database and data management

- SQLite
- Python Standard Library (json, io, datetime)

External APIs

Google Gemini API

Thank you

Any questions? Ask away!