# Receipt & Invoice Digitizer

Milestone 1 – Intelligent Document Digitization Pipeline

**Internship Project | Abhay Maurya**

# Problem Statement

- **Manual Handling:** Processing receipts is repetitive, time-consuming, and highly prone to human error.
- **Storage & Search:** Physical bills are difficult to organize, creating data silos and making historical analysis impossible.
- **Legacy Limitations:** Traditional OCR tools often produce unstructured, unreliable text that requires heavy post-processing.

# Project Objective (Milestone 1)

**Goal:** Establish a stable, modular pipeline for converting physical documents into accurate, structured digital data.

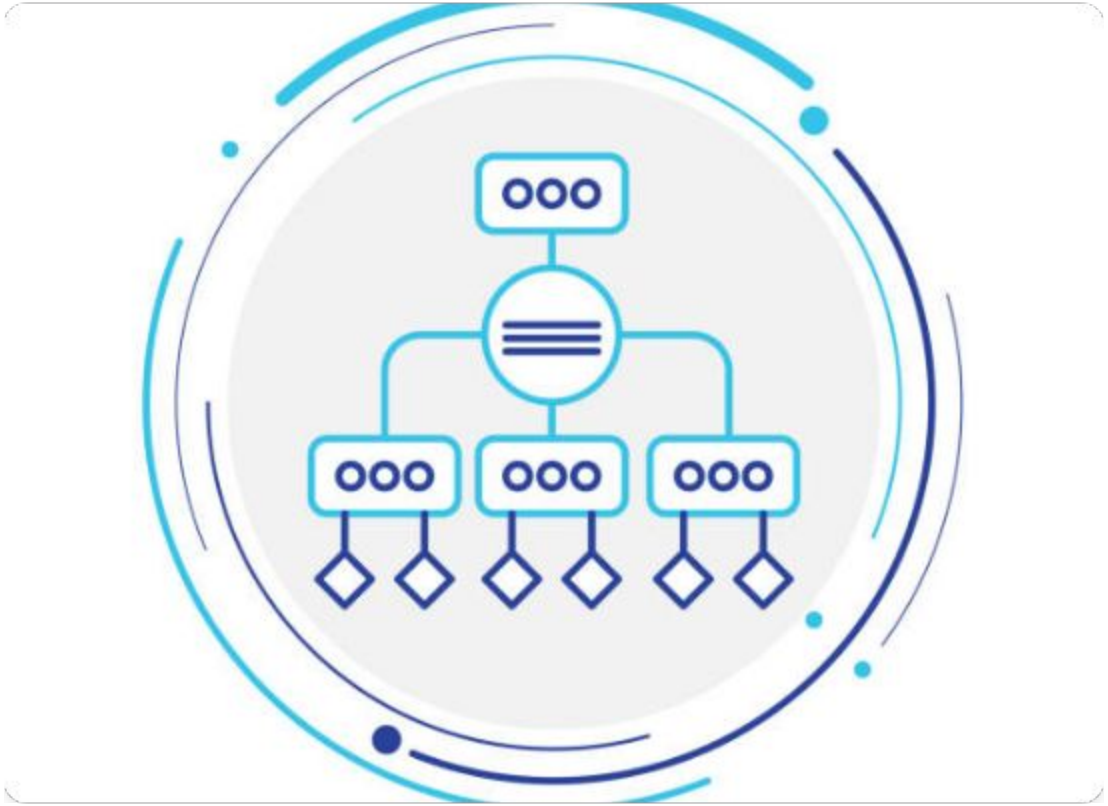
- **Multi-format Ingestion:** Seamless support for JPG, PNG, and multi-page PDF files.
- **Automated Preprocessing:** Intelligent cleaning to handle noise and shadows.
- **Next-Gen OCR:** Single-call AI extraction using Google Gemini.
- **Structured Output:** Direct-to-JSON validation to eliminate regex parsing.
- **Session Persistence:** Streamlit state management to prevent redundant processing.

# System Overview

## The Solution

A web-based Streamlit application powered by AI to automate data extraction.

- **AI-Driven Core:** Leverages Google Gemini 2.5 Flash.
- **Modular Design:** Decoupled ingestion, cleaning, and extraction layers.
- **Scalable:** Stateless logic ready for future database integration.



# Architecture & Data Flow

## User Upload

Accepts JPG, PNG, PDF

## Ingestion

Validation & Conversion

## Preprocessing

Denoise & Normalize

## Gemini OCR

Contextual Extraction

## Streamlit UI

Display Data & JSON

# Core Modules

## Ingestion Module

Handles safe file uploads, detects magic bytes, and converts PDF pages to readable images.

## Preprocessing Module

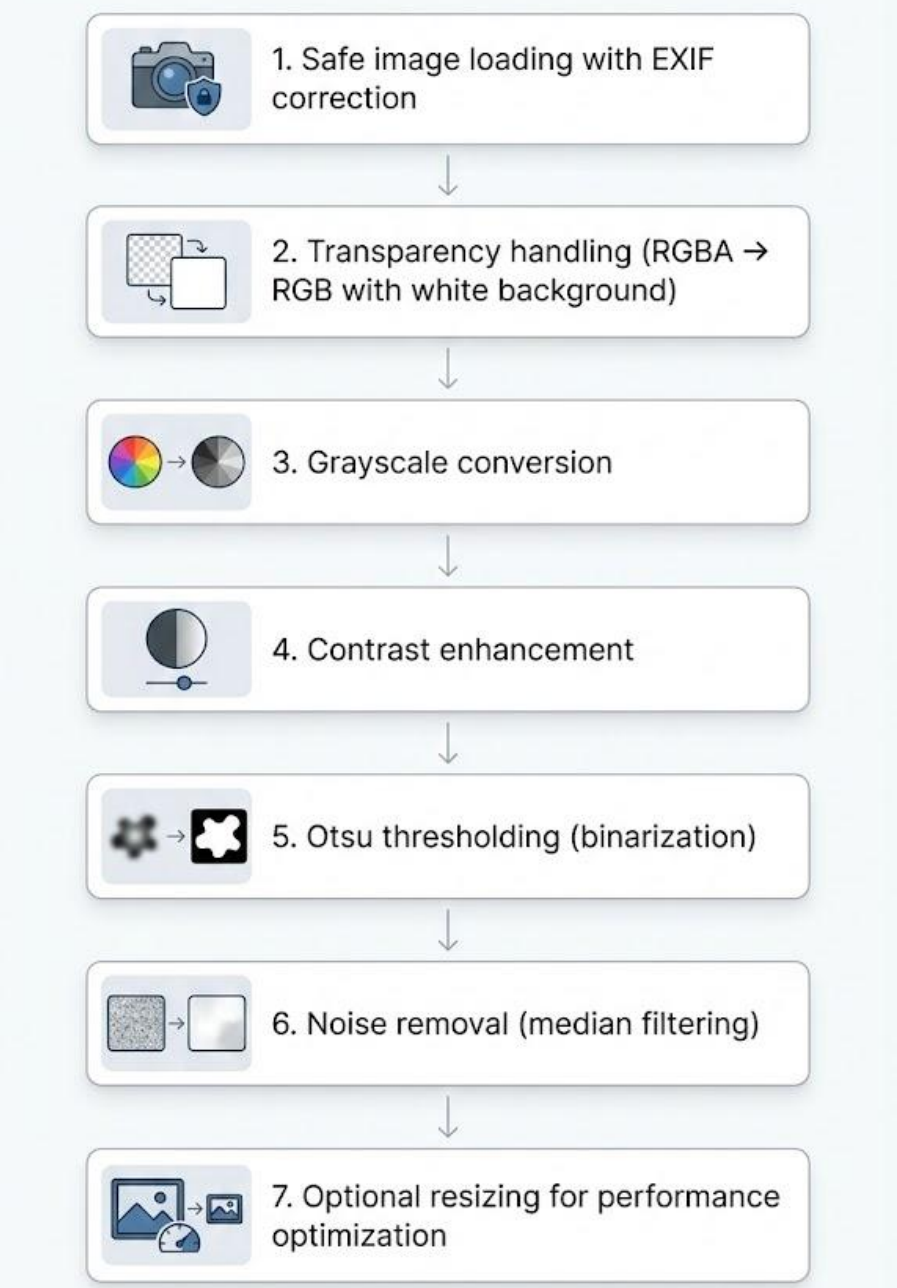
Cleans input data by correcting transparency, removing noise, and resizing to optimize API usage.

## Extraction Module

Sends optimized images to Gemini and receives structured, schema-compliant JSON responses.

# Image Preprocessing

- **Standardization:** Converts all inputs (RGB, RGBA) to a uniform format.
- **Noise Reduction:** Removes background artifacts (grain, shadows) that confuse OCR.
- **Optimization:** Smart-resizing of large images prevents API timeouts.
- **Transparency Handling:** Flattens alpha channels to prevent "black box" errors.



# OCR & Extraction (Gemini AI)

- **Contextual Understanding:** Reads "Total" even on crumpled or faint receipts.
- **Single-Call Efficiency:** Replaces the multi-step pipeline (Detection → Layout → Regex).
- **Structured JSON:** The model is prompted to output strict data, eliminating parsing ambiguity.
- **Deterministic Output:** Configured to minimize hallucinations.

## Structured Output


{
"vendor": "Walmart",
"date": "2023-10-12",
"total": 45.99
}

# Session State & Error Handling

## Session State (Performance)

- Persists data across UI re-renders.
- Prevents expensive API re-calls when switching tabs.
- Maintains user context seamlessly.

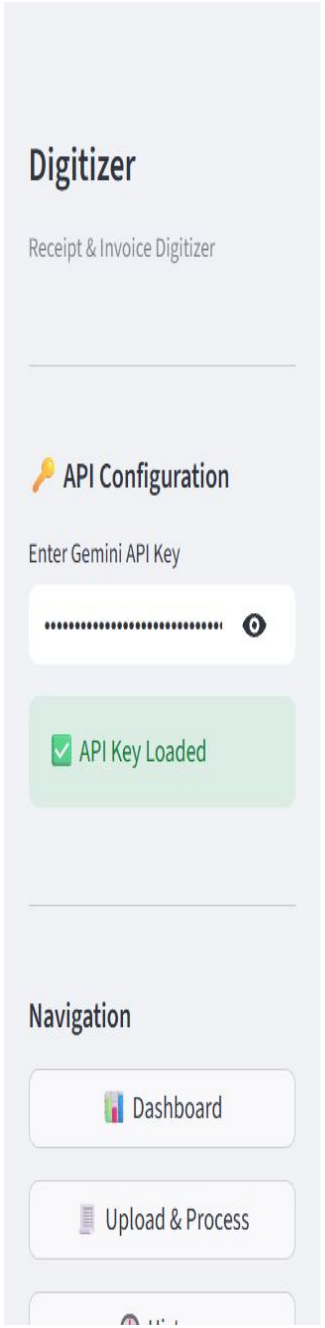
## Error Handling (Robustness)

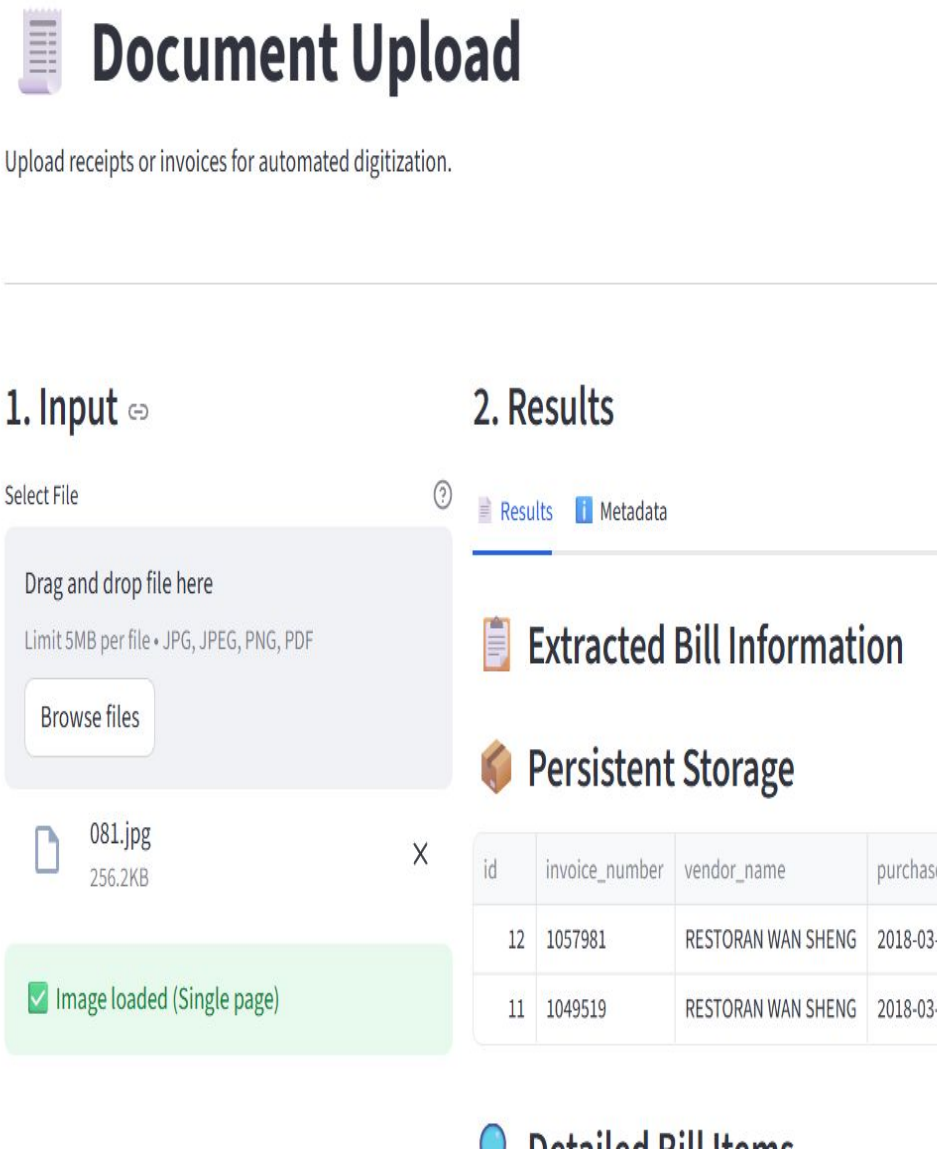
- **Ingestion:** Catches corrupt files/missing libs.
- **API:** Handles timeouts/invalid keys gracefully.
- **Validation:** Enforces schema checks before display.

# User Interface

*Prioritizing clarity and debuggability.*

- **Upload Widget:** Drag-and-drop with validation.
- **Image View:** Processed Image is shown to user.
- **Data Tabs:** Dedicated views for Visuals, JSON, and Metadata.



Deploy

# Digitizer

Receipt & Invoice Digitizer

## API Configuration

Enter Gemini API Key

••••••••••••••••••••••••••••••

✅ API Key Loaded

## Navigation

- Dashboard
- Upload & Process
- History

# 1. Input

Select File

Drag and drop file here

Limit 5MB per file • JPG, JPEG, PNG, PDF

Browse files

081.jpg

256.2KB

✅ Image loaded (Single page)


# 2. Results

Results | Metadata

## 📋 Extracted Bill Information

## 📦 Persistent Storage

| id | invoice_number | vendor_name | purchase_date | purchase_time | payment_method | total_amoun |
|---|---|---|---|---|---|---|
| 12 | 1057981 | RESTORAN WAN SHENG | 2018-03-23 | 13 hours | CASH | $6.70 |
| 11 | 1049519 | RESTORAN WAN SHENG | 2018-03-17 | 15 hours | CASH | $9.60 |

## 🔍 Detailed Bill Items

Select ID to view items:

Bill #11 - RESTORAN WAN SHENG - 2018-03-17

| s_no | item_name | quantity | unit_price | item_total |
|---|---|---|---|---|
| 1 | Kopi (B) | 1 | 2.2 | 2.2 |
| 2 | Teh (B) | 2 | 2.2 | 4.4 |

# Key Outcomes of Milestone 1

- **Stable Pipeline:** Successfully integrated Ingestion → Preprocessing → AI OCR.
- **High Accuracy:** High extraction rate using Gemini 2.5 Flash.
- **Modular Codebase:** Clean separation of concerns (Ingestion, OCR, UI).
- **Security:** Environment variable management for API keys.
- **Foundation Laid:** Architecture is ready for Database integration.

# What's Next (Milestone 2)

## Database Persistence

Storing extracted data in SQL for long-term retention.

## Analytics Dashboard

Visualizing spending trends, categories, and vendor stats.

## History Module

View and manage previously scanned documents.


## Export Functionality

Generate CSV/Excel reports for external accounting.