1. Title

Receipt and Invoice Digitizer

---

2. Introduction

In modern business environments, receipts and invoices are generated at high volumes across retail, logistics, services, and enterprise operations. These documents often exist in unstructured physical or semi-structured digital formats such as scanned images or PDFs. Manual handling of such documents leads to inefficiencies, human errors, loss of records, delayed reimbursements, and limited financial visibility.

The **Receipt & Invoice Digitizer** project aims to solve this problem by providing an automated, AI-assisted system that converts unstructured receipt and invoice documents into machine-readable digital data. The application is implemented as a **multi-page Streamlit web application**, integrating advanced image preprocessing techniques with **Google Gemini AI** for Optical Character Recognition (OCR) and structured information extraction.

This document describes **Milestone 1**, which focuses on building a **robust digitization foundation**—from document ingestion to OCR extraction and UI visualization—without yet emphasizing long-term analytics optimization or enterprise integrations.

---

3. Problem Statement

Organizations and individuals routinely process large volumes of receipts and invoices. The traditional workflow typically involves:

- Manual data entry into spreadsheets or accounting systems
- Storing physical copies for compliance
- High probability of transcription errors
- Inconsistent formatting across vendors

- Difficulty in tracking expenses over time

These challenges result in:

- Increased operational cost
- Reduced accuracy in financial records
- Delays in expense reconciliation
- Poor audit readiness

There is a strong need for an **automated, intelligent, and scalable solution** that can:

- Accept receipts and invoices in common formats
- Extract text accurately
- Convert unstructured content into structured data
- Present results clearly to users

---

4. Milestone 1 Objective

The primary objective of **Milestone 1** is to design and implement a **robust, secure, and extensible document digitization foundation** capable of reliably converting physical receipts and invoices into structured digital data. This milestone focuses on establishing the **core technical pipeline** that will support all subsequent enhancements such as analytics, reporting, and long-term data persistence.

Milestone 1 is intentionally scoped to emphasize **correctness, reliability, and architectural soundness**, ensuring that downstream milestones can be built without refactoring core components. The system is designed to operate consistently across multiple document types while maintaining high OCR accuracy, predictable behavior, and a user-friendly interaction model.

Key Objectives

1. **Multi-Format Document Ingestion**
   Enable secure and reliable ingestion of common receipt and invoice formats, including **JPG, PNG, and PDF files**. The system must correctly identify file types, validate file size and integrity, and handle both single-page and multi-page documents without manual intervention.
2. **Standardized Image Conversion Pipeline**
   Convert all supported document formats into a **uniform, OCR-ready image representation**. PDFs must be safely rendered into individual page images, and all image outputs must be standardized (RGB format, consistent resolution) to ensure predictable downstream processing.
3. **Automated Image Preprocessing for OCR Optimization**

Implement an automated preprocessing layer that enhances OCR performance by applying operations such as **grayscale conversion, noise reduction, contrast enhancement, and binarization**. This preprocessing must run transparently in the background and be resilient to variations in document quality.

4. **Single-Call OCR and Structured Data Extraction**
   Integrate **Google Gemini AI** to perform OCR and semantic understanding in a **single API call per document/page**. The system must extract not only raw text but also **structured bill data**, including vendor details, dates, line items, tax values, totals, currency, and payment method.

5. **Strict Schema Enforcement and Data Normalization**
   Enforce a predefined JSON schema on all extracted data to ensure consistency and database compatibility. Missing or ambiguous values must be handled using safe defaults, and numeric fields must be normalized to valid data types to prevent downstream failures.

6. **Session State Management and Workflow Continuity**
   Maintain application state across Streamlit reruns using structured session state management. This ensures that uploaded files, processed images, extracted results, and user actions persist seamlessly during navigation, reducing redundant computation and improving user experience.

7. **User-Centric and Intuitive Interface Design**
   Provide a **clean, minimal, and user-friendly interface** that clearly guides users through upload, processing, and result inspection. The UI must support image previews, extracted data visualization, and logical navigation without overwhelming the user.

8. **Controlled Error Handling and Fault Tolerance**
   Implement comprehensive error handling across ingestion, preprocessing, OCR, and extraction stages. Failures must be **graceful, informative, and non-destructive**, ensuring that partial errors do not crash the application or corrupt session state.

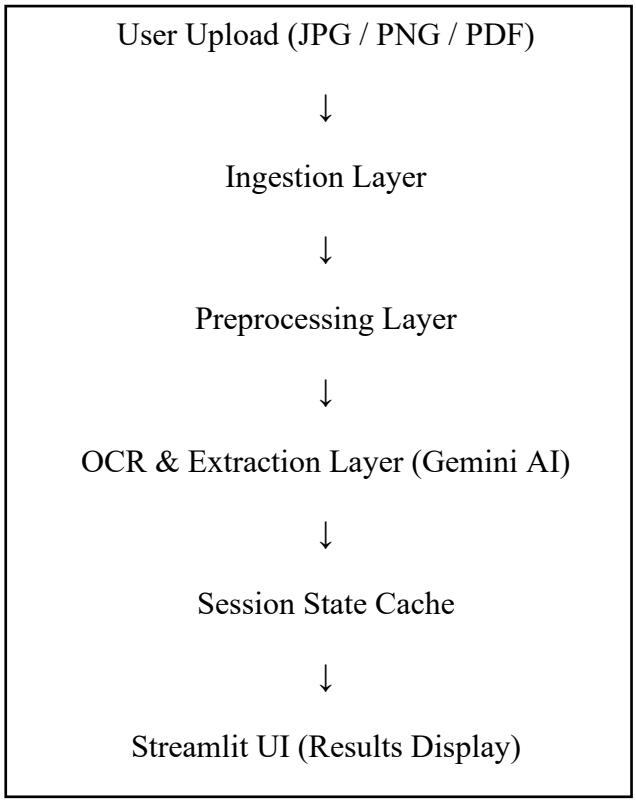
9. **Foundation for Persistent Storage and Analytics**
   Although advanced analytics and reporting are reserved for later milestones, Milestone 1 establishes the **data structures, schemas, and interfaces** required for seamless integration with persistent storage systems and analytical dashboards in future phases.

Outcome of Milestone 1

By the completion of Milestone 1, the system delivers a **production-ready core digitization pipeline** capable of transforming unstructured receipt and invoice documents into reliable, structured digital records. This milestone ensures technical stability, modularity, and scalability, forming a solid base for advanced features in subsequent milestones.

---

5. High-Level Architecture



---

6. Technology Stack

Layer	Technology
Frontend	Streamlit
OCR Engine	Google Gemini 2.5 Flash
Image Processing	Pillow (PIL), OpenCV
PDF Processing	pdf2image (Poppler)
Backend Logic	Python
State Management	Streamlit Session State
Data Handling	Pandas
Database (Planned)	MySQL (used partially in later phase)

---

7. Modules Implemented

7.1 Ingestion Layer (ingestion.py)

Purpose:

The ingestion layer acts as the **gateway** between user uploads and internal processing. Its responsibility is to **safely load documents**, normalize them, and produce a consistent internal representation.

Key Responsibilities

- Detect file type (image vs PDF)
- Validate file integrity
- Convert PDFs into page-wise images
- Apply security limits
- Generate file hash for change detection

Supported Inputs

- Local file paths
- Streamlit UploadedFile objects
- In-memory byte streams (BytesIO)

Security Controls

- Maximum PDF pages (prevents OOM attacks)
- Maximum image pixel threshold (prevents decompression bombs)
- SHA-256 file hashing

Outputs

- List[PIL.Image]
- Metadata dictionary containing:
    - Filename
    - File type
    - Number of pages
    - File hash
    - Truncation status

Design Principle

The ingestion layer never performs preprocessing or OCR. It only prepares data.

7.2 Preprocessing Layer (preprocessing.py)

Purpose:

Raw images captured from cameras or scanners often contain noise, skew, low contrast, or transparency. The preprocessing layer ensures images are **OCR-ready**.

Processing Pipeline:

1. Safe image loading with EXIF correction
2. Transparency handling (RGBA → RGB with white background)
3. Grayscale conversion
4. Contrast enhancement
5. Otsu thresholding (binarization)
6. Noise removal (median filtering)
7. Optional resizing for performance optimization

Output:

Clean, binary PIL image optimized for OCR

Design Principle:

Preprocessing is **purely visual** and **content-agnostic**.

7.3 OCR & Extraction Layer (ocr.py)

Purpose:

This layer performs **single-call OCR and structured extraction** using Google Gemini AI.

Why Single-Call Design?

- Reduces API cost
- Avoids synchronization issues
- Improves consistency
- Faster response time

Extracted Fields:

- Vendor name
- Purchase date
- Purchase time
- Currency

- Line items (description, quantity, unit price, total)
- Tax amount
- Total amount
- Payment method
- Raw OCR text

Key Characteristics:

- Deterministic output (temperature = 0)
- Strict JSON schema
- Defensive parsing and normalization
- Graceful failure handling

Design Principle:

OCR output must be **immediately database-ready**.

---

8. Streamlit Application Design

8.1 Session State Management

Streamlit reruns the entire script on every interaction.
Session state is used to preserve:

- Uploaded file context
- API key
- Preprocessed images
- OCR results
- Navigation state
- Save status

This prevents:

- Redundant OCR calls
- Data loss on reruns
- UI inconsistency

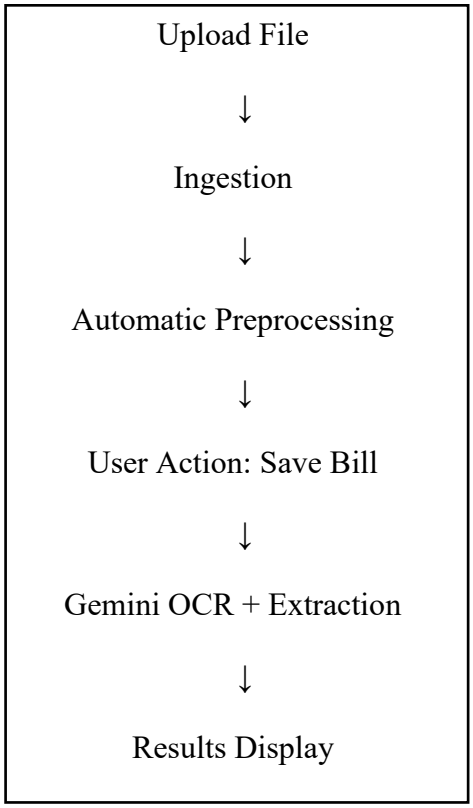
8.2 Sidebar Navigation

The sidebar manages:

- **Gemini API key input**
- **Page navigation**
- **Application metadata**

The API key is stored only in session memory and never logged or persisted.

8.3 Upload & Process Workflow



Supported Scenarios:

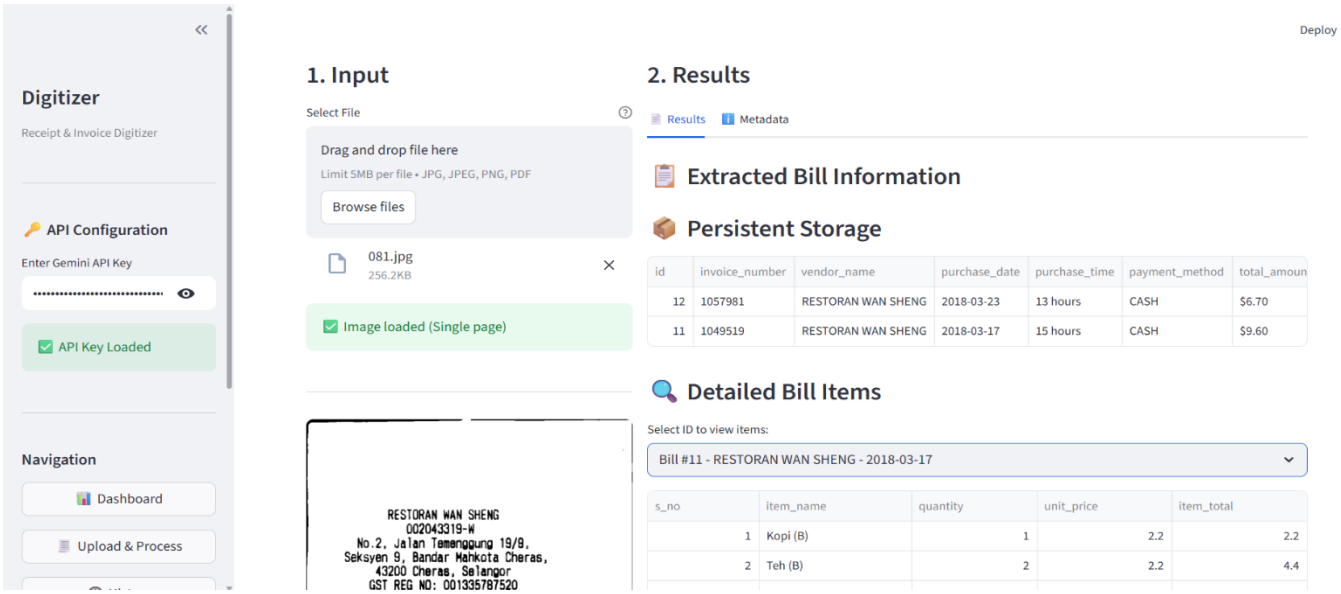
- Single image receipts
- Single-page PDFs
- Multi-page PDFs (page-wise processing)

8.4 Results Display

Results are presented using:

- Preprocessed image previews
- Structured bill data
- File metadata

Tabbed layout improves readability and user experience.



---

9. Error Handling & Validation

Milestone 1 prioritizes **controlled failures** over silent errors.

Handled cases include:

- Missing API key
- Unsupported file formats
- Oversized uploads
- Corrupted images
- PDF conversion failures
- Invalid or partial AI responses

All errors are surfaced with **human-readable messages**.

10.Security & Stability Considerations

- File size limits enforced
- Page limits enforced on PDFs
- No secrets stored or logged
- No untrusted file writes
- Hash-based file change detection
- Defensive JSON parsing

11.Conclusion

Milestone 1 successfully delivers a production-grade foundation for receipt and invoice digitization. The system demonstrates strong architectural separation, reliable preprocessing, accurate OCR extraction, and a user-friendly interface.

This milestone lays a solid technical base for:

- Persistent storage optimization
- Advanced analytics
- Multi-user workflows
- Enterprise integrations