# System Architecture

This system architecture for invoice data extraction consists of two main approaches, both designed to handle single PDF inputs and multiple PDFs via GitHub repository URLs. The key difference lies in how they process extractable PDFs.

## Approach-1 (Only GEMINI based approach):

This approach uses Google's Gemini AI model for all PDF processing, regardless of whether the PDF is extractable or non-extractable.

**Process flow:**

a. Input: Single PDF or Github Repo URL for multiple PDFs

b. PDF Type Determination

c. For all PDFs:
   - Non-extractable: Convert PDF to image
   - Extractable: Direct processing

d. Google Gemini AI Processing

e. JSON Parsing and Cleaning

f. Data Structuring *(Into Tabular format and JSON format)*

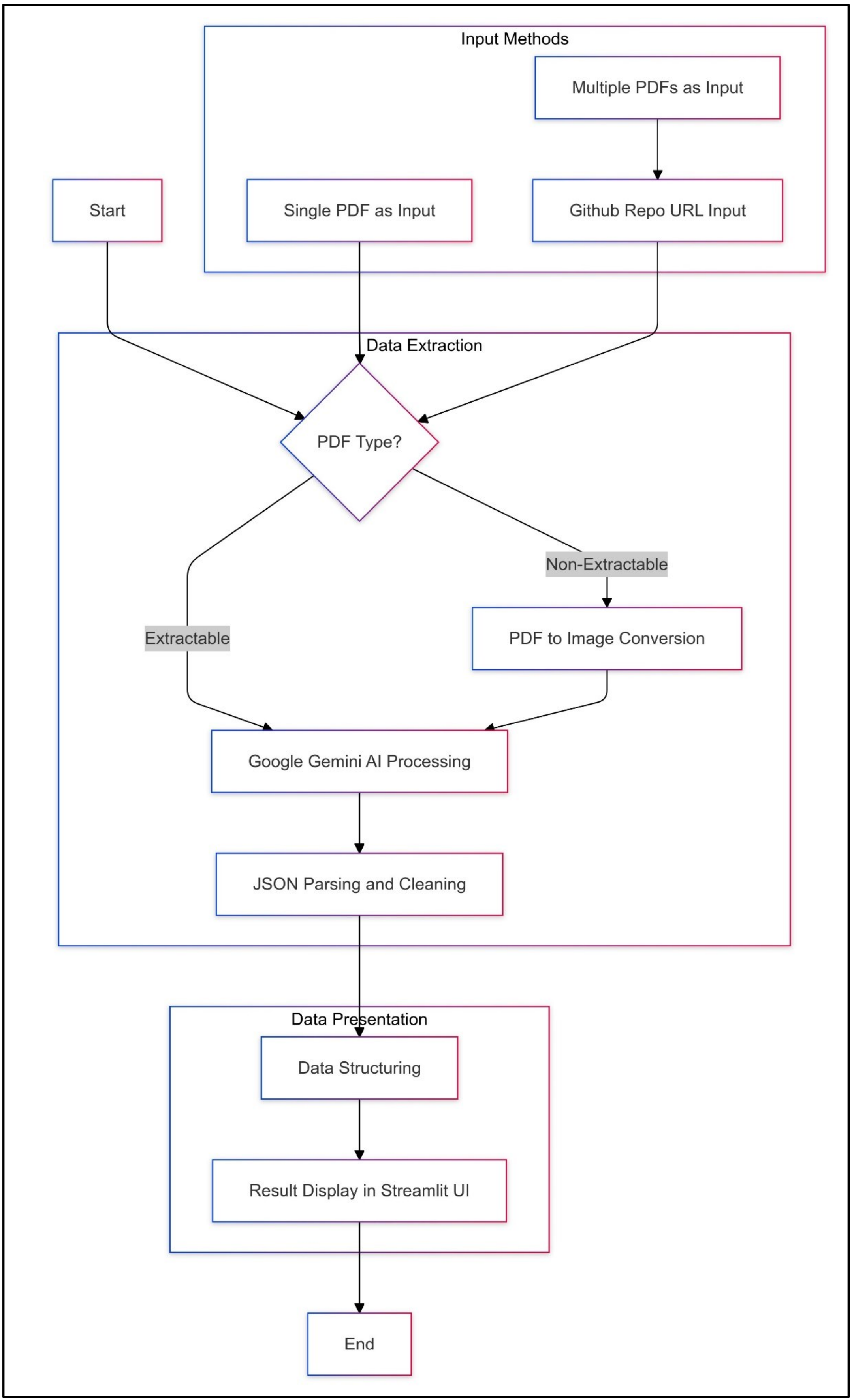
g. Result Display in Streamlit UI

**Advantages:**

a. Consistent processing method for all PDFs
b. Higher accuracy due to its ability to understand context and handle various formats
c. Simplified workflow with fewer components

Disadvantages:

**a. Cost of API tokens** used can increase based on the report’s size



**Approach 1 Architecture**

## Approach-2 (PDFPlumber and GEMINI model together):

This approach combines PDFPlumber for extractable PDFs and Gemini model for non-extractable PDFs.

**Process flow:**

a. Input: Single PDF or Github Repo URL for multiple PDFs

b. PDF Type Determination

c. For extractable PDFs:

   - PDFPlumber Extraction
   - Data Structuring (Into Tabular format and JSON format)
   - Result display in Streamlit UI

d. For Non-extractable PDFs:

   - Gemini model-based processing
   - JSON Parsing and Cleaning
   - Data Structuring (Into Tabular format and JSON format)
   - Result display in Streamlit UI

**Advantages:**

a. This can be cost-effective approach compared with approach 1, since we are using Gemini model only for non-extractable PDFs

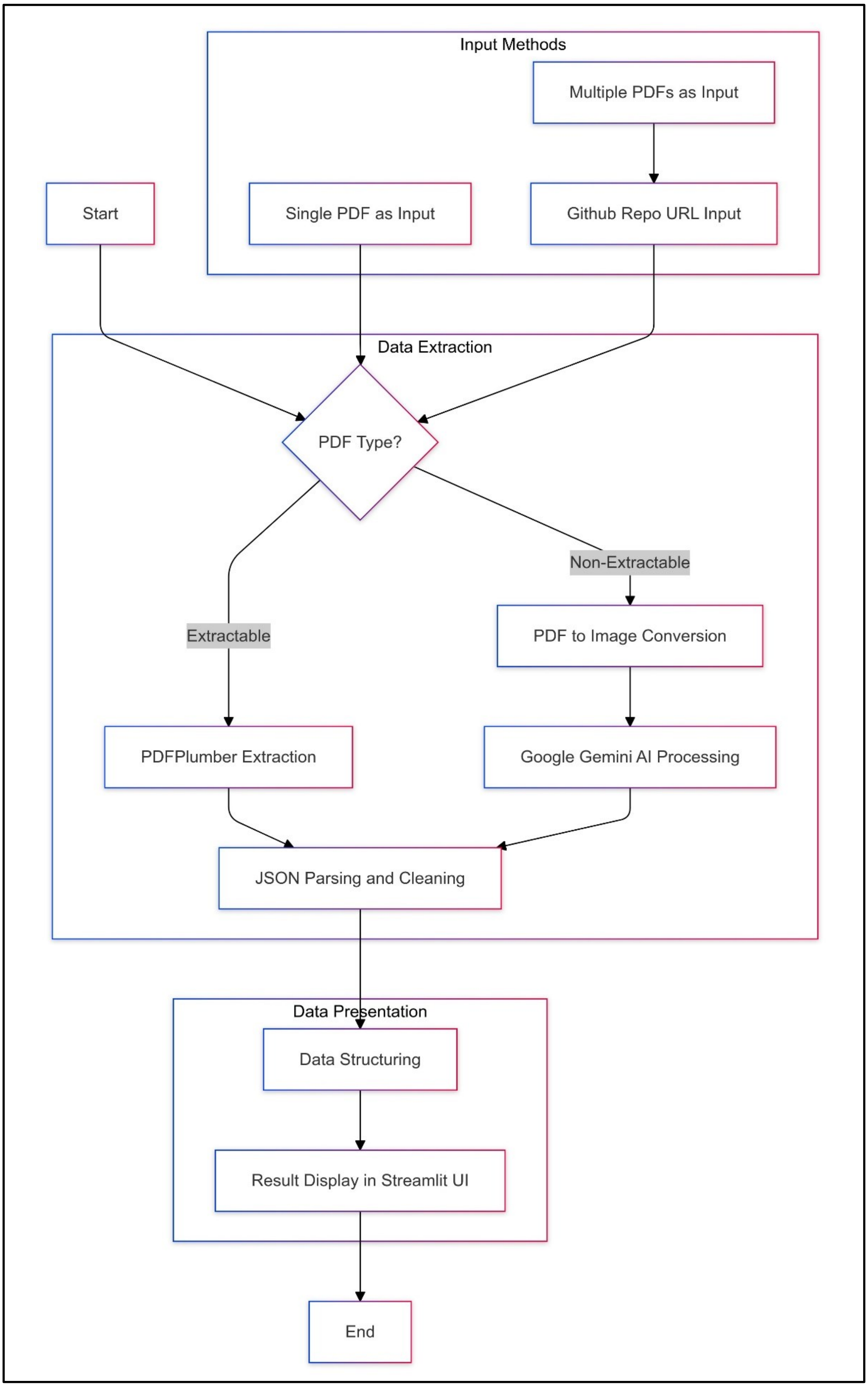
b. This approach still maintains the accuracy for non-extractable PDFs same as the approach 1 at less cost

c. Potentially faster for extractable PDFs

**Disadvantages:**

a. Compromise with accuracy for extractable PDFs to reduce the cost

b. Potential inconsistencies in extraction quality between the two methods



**Approach 2 Architecture**