FTEC 5660 Homework 1
Receipt Understanding Using Multimodal Large Language Models

Shi Zhan
Student ID: 1155209445

January 17, 2026

1 System Overview

A high-level workflow of the system is shown below:

1. Receipt images are loaded and converted into Base64 data URLs.
2. Each image is sent to the Gemini model together with a carefully designed prompt.
3. The model outputs structured financial information in JSON format.
4. Extracted data from multiple receipts is aggregated to answer user queries.

2 Financial Information Extraction

For each receipt image, the model extracts the following fields:

- Total amount paid
- Subtotal
- Discounts (if any)

The raw JSON strings returned by the model are parsed and converted into Python dictionaries. Special care is taken to handle formatting issues such as Markdown-style code blocks.

3 Query Classification and Aggregation

To support user interaction, queries are classified into predefined categories:

- Total money spent across all receipts.
- Total amount without discounts.

After classification, the system performs aggregation over all parsed receipts:

- The total spent is computed by summing the `total_paid` values.
- The total without discounts is computed by adding back the extracted discount amounts.

Irrelevant queries are safely ignored.

4 Experimental Results

Using multiple receipt images as input, the system successfully:

- Extracted structured financial information from each receipt.
- Correctly computed the total expenditure.
- Answer the pre-defined queries and refuse to answer the irrelevant queries.

The results demonstrate that multimodal LLMs can effectively replace traditional OCR and rule-based pipelines for receipt understanding tasks.