Supermarket Bill Understanding using Multimodal LLM

Student Name: HU TAINYU
Student ID: 1155249527

1 Introduction

This project develops a multimodal AI system that analyzes supermarket bill images and answers user queries regarding spending. The system supports two types of queries:

- Query 1: How much money did I spend in total for these bills?
- Query 2: How much would I have had to pay without the discount?

In addition, the model is required to reject irrelevant or out-of-domain queries. The system is implemented using a multimodal large language model (Gemini), which takes both images and text as inputs.

2 System Overview

Figure 1 illustrates the overall pipeline of the proposed system.

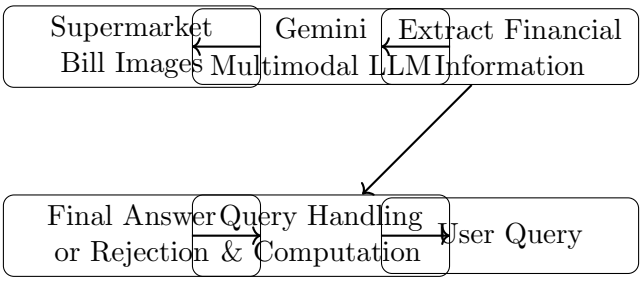

Figure 1: System pipeline for supermarket bill analysis

The system consists of three main modules:

1. Image input module that loads multiple supermarket bill images.
2. Multimodal LLM module that extracts structured information from each bill.
3. Query handling module that interprets user queries and computes final results.

3 Methodology

3.1 Image Understanding with Gemini

Each supermarket bill image is provided to the Gemini multimodal model together with a prompt requesting financial information extraction. The model outputs structured text containing:

- Total paid amount,

- Discount amount (if available),
- Original price before discount.

Figure 2 shows an example of bill image input and extracted information.

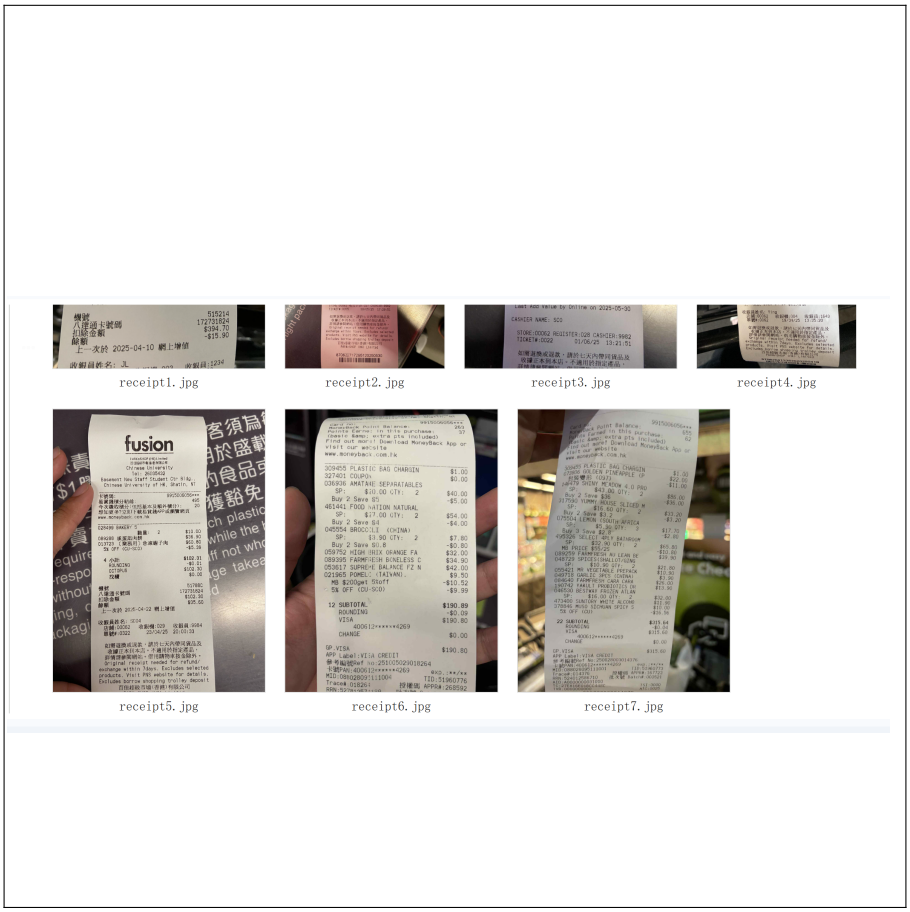

Figure 2: Example bill image and extracted information

3.2 Query Interpretation

The system accepts a natural language query from the user. It checks whether the query matches one of the supported query types (Query 1 or Query 2). If the query does not belong to either category, the system outputs a rejection message indicating that the query is unsupported.

3.3 Computation Logic

Let T_i denote the actual paid amount for bill i, and D_i denote the discount amount for bill i.

For Query 1 (total spending), the computation is:

$$\text{Total Spending} = \sum_i T_i$$

For Query 2 (original cost without discount), the computation is:

$$\text{Original Cost} = \sum_i (T_i + D_i)$$

This approach ensures that the final answer is based on extracted numerical values rather than free-form text generation.

4 Implementation

The system is implemented in Python using a Jupyter Notebook based on the provided starter code. The main steps include:

1. Loading supermarket bill images.
2. Sending images to the Gemini multimodal API.
3. Parsing the model outputs to extract numerical values.
4. Aggregating values across multiple bills.
5. Answering user queries according to predefined rules.

5 Experimental Results

5.1 Query 1: Total Spending

The system successfully extracts the paid amount from each bill image and computes the total spending. Figure 3 shows an example output.

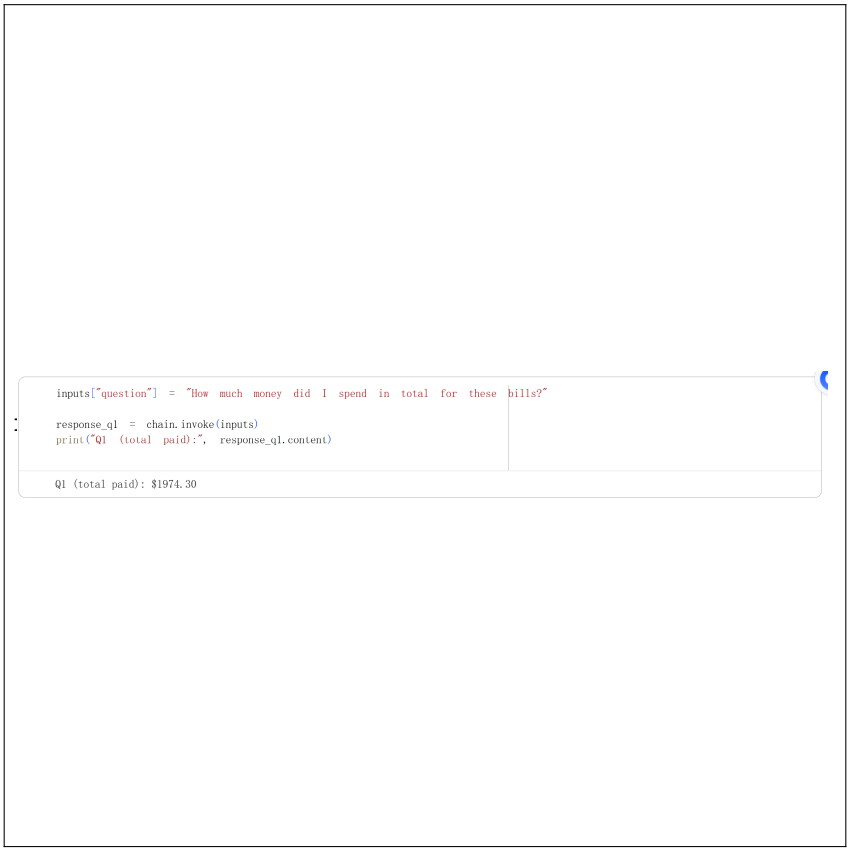

Figure 3: Example result for Query 1

5.2 Query 2: Cost Without Discount

The system reconstructs the original cost by adding discount values to the paid amounts. Figure 4 presents an example output.

Figure 4: Example result for Query 2

5.3 Out-of-Domain Query Rejection

For irrelevant queries, such as unrelated questions about weather or location, the system produces a rejection response. Figure 5 shows an example.

6 Discussion

The results demonstrate that multimodal large language models can effectively extract numerical information from supermarket bills and perform arithmetic reasoning. The system is robust for clearly printed receipts with visible discount information. However, performance may degrade when images are blurry or when discount information is missing or ambiguous.

7 Conclusion

This project presents a multimodal AI system for understanding supermarket bill images and answering spending-related queries. The system supports total spending calculation, original cost estimation without discounts, and rejection of irrelevant queries. The experimental results confirm the feasibility of applying multimodal LLMs to real-world financial document understanding tasks.

Future improvements may include enhanced numeric parsing, more advanced OCR integration, and support for more complex bill layouts.

```
Test 1
Question: Summarize these receipts.
Answer: Sorry, I can only answer Query 1 or Query 2.
--------------------------------------------------
Test 2
Question: How much would I have had to pay without the discount?
Answer: $2399.97
--------------------------------------------------
Test 3
Question: What is the average price per item?
Answer: Sorry, I can only answer Query 1 or Query 2.
--------------------------------------------------
Test 4
Question: How many receipts are there?
Answer: Sorry, I can only answer Query 1 or Query 2.
--------------------------------------------------
Test 5
Question: Tell me the store address.
Answer: Sorry, I can only answer Query 1 or Query 2.
--------------------------------------------------
Test 6
Question: How much money did I spend in total for these bills?
Answer: $1974.30
```

Figure 5: Example of rejected query

References

[1] Google Gemini Multimodal API Documentation.
Course Notebook: Helper Functions and Image Input to Gemini.