Homework 1 Report: Supermarket Bill Analysis using Multimodal LLMs

Course: FTEC 5660 - Agentic AI for Business and FinTech

Student ID: 1155246851

1. Problem Description

The objective of this assignment is to develop an AI model capable of processing multiple supermarket receipt images and answering specific user queries based on the extracted data. The model must handle three primary tasks:

1. **Query 1:** Calculate the total amount spent across all provided receipts.
2. **Query 2:** Calculate the total amount that would have been paid without any applied discounts.
3. **Irrelevant Query Handling:** Recognize and reject queries that do not pertain to the provided receipt data.

2. Methodology and Solution Design

2.1 Technology Stack

The solution is built using the following components:

- **Large Language Model:** `gemini-2.5-flash` via Google Vertex AI. This model was chosen for its multimodal capabilities, allowing it to interpret text and image data simultaneously.
- **Framework:** `LangChain` was used to manage prompt templates and pipeline execution.
- **Environment:** Google Colab for development and testing.

2.2 Image Processing

Receipt images are retrieved from a central repository, unzipped, and converted into Base64 encoded strings. These encoded strings are then formatted into Data URLs suitable for processing by the Gemini API.

2.3 Prompt Engineering

Separate prompt pipelines were designed for each query type to ensure accuracy and limit the model's scope:

- **Query 1 Pipeline:** Instructs the model to identify and sum the final paid amounts (Subtotals) from each receipt.

- **Query 2 Pipeline:** Specifically directs the model to look for "Original Prices" or sum the subtotal with identified "Discounts" to find the pre-discount total.
- **Negative Constraint:** Both pipelines include instructions to reject out-of-domain queries with a polite explanation that the model's functionality is limited to shopping receipt data.

3. Implementation Details

The implementation follows a modular structure:

1. **Helper Functions:** `image_to_base64` and `get_image_data_url` handle the necessary data transformations for multimodal input.
2. **Model Initialization:** The Gemini model is configured with a low `temperature` of 0.2 to ensure deterministic and accurate numerical output.
3. **Pipeline Construction:** Using LangChain's `ChatPromptTemplate`, the system prompt and user query are combined with the Base64 image data to form a comprehensive input for the LLM.

4. Evaluation and Results

The model was tested against several receipt samples from Fusion (ParknShop HK).

- **Accuracy:** The model correctly identified itemized prices, subtotals, and specific discount labels (e.g., "Buy 2 Save $12.8", "5% OFF (CU)”). They are aligned within +- 2 dollar within the actual data
- **Robustness:** When presented with irrelevant queries like "What's the weather like?", the model correctly triggered the rejection logic, stating its inability to process non-receipt data.

5. Conclusion

By leveraging the multimodal capabilities of `gemini-2.5-flash` and the structured pipeline of `LangChain`, the developed solution effectively automates the extraction and analysis of financial data from unstructured image sources. The approach demonstrates high reliability in calculating totals and identifying complex discount structures commonly found in retail environments.