Name: Liu Zihan Student ID: 1155253144

Agentic AI for Business and FinTech (FTEC5660)

Supermarket Receipt Recognition Agent: Architectural Design and Implementation Report

1. Agent Architectural Design

This project involves the construction of an intelligent agent equipped with multimodal perception and logical reasoning capabilities, designed to automate the processing and analysis of supermarket receipt images. The architecture follows a "Perception, Understanding, Intent Classification, and Execution" pipeline to address the inconveniences associated with traditional manual bookkeeping.

The system logic is divided into four core layers:

(1) **Infrastructure & Data Layer:** Handles the underlying environment configuration, installation of LangChain dependencies, and the Base64 encoding of raw images.
(2) **Core Recognition Layer:** Configures and invokes the Gemini multimodal large language model, utilizing carefully crafted system prompts to perform OCR analysis through semantic understanding.
(3) **Intent Classification Layer:** Categorizes user queries through semantic scanning to identify the specific task required.
(4) **Logical Execution Layer:** Executes specific mathematical aggregation logic or triggers a rejection mechanism based on the identified intent.

This modular design ensures the integrity of information acquisition while providing the flexibility and precision necessary for calculating varied query intents.

2. Image Perception and Preprocessing

In the perception phase, the primary task is to construct a multimodal input environment. Using helper functions such as `image_to_base64` and `get_image_data_url`, we convert local `.jpg` receipt images into Base64-encoded Data URLs.

The rationale behind this approach is that the **Gemini-2.5-flash** model requires image data to be transmitted as a standardized text stream. By encoding multiple receipt images and inputting them into the model simultaneously, the Agent can "examine" all receipts within a single Context Window. Compared to processing images individually, this global perspective maintains higher data consistency and significantly improves the efficiency of aggregating amounts across multiple receipts.

3. Structured OCR Recognition and Extraction Logic

In the core processing stage, the Agent goes beyond simple character recognition by employing a structured extraction logic. Guided by the system prompt, the Agent actively identifies and distinguishes between two key numerical categories:

(1) **Item Amounts:** The specific original price for each individual item on the receipt.
(2) **Discount/Saving Amounts:** Accurate identification of deductions marked as "Savings," "Discount," or denoted by negative signs.

This "Extraction before Calculation" strategy is designed to handle the highly complex formats of supermarket receipts. Since layout styles vary significantly between different retailers , asking a model to calculate the total directly often leads to arithmetic deviations or "hallucinations." By first building a temporary Structured Digital Database, the Agent preserves the spatial layout information (such as amounts listed on the right or specific discount abbreviations), ensuring more accurate recognition and a solid foundation for subsequent calculations.

4. Intent Classification and Task Execution Mechanism

At the final interaction layer, the Agent demonstrates its decision-making capacity. Upon receiving a user query, the Agent first uses keyword recognition (such as "total spend" or "without discount") to determine the user's true intent and dynamically selects a calculation strategy:

(1) **Query 1 (Total Spend):** The Agent identifies "Total" or "Amount Due" at the bottom of each receipt from the structured data and sums them up.
(2) **Query 2 (Original Price without Discount):** The Agent executes a specific algorithmic logic, adding the identified discount amounts back to the actual paid amounts to restore the original price.
(3) **Rejection Mechanism:** If the query is irrelevant to the receipts, the Agent is constrained by the prompt to output a predefined rejection message (e.g., "This is an irrelevant question. Please rephrase your question.").

Regarding model configuration, we utilized Gemini-2.5-flash with the temperature set to 0. This is crucial because receipt analysis requires high determinism and consistency, necessitating the elimination of random variation. Furthermore, classifying user intent beforehand enhances the Agent's robustness and safety. By setting clear operational boundaries in the prompt, the Agent effectively simulates a commercial-grade logic that avoids hallucinations and rejects out-of-domain requests, ensuring rigorous and accurate task completion.

5. Conclusion

Overall, this Supermarket Receipt Agent successfully achieves a closed-loop system ranging from low-level data processing to high-level logical control. Through the design path of "multimodal structured extraction followed by intent-based logic," the Agent demonstrates not only powerful visual recognition but also the intelligence to switch calculation strategies based on user needs. This accurately fulfills all requirements for receipt analysis and the rejection of irrelevant queries.