Prompt-Guided Multimodal System Workflow

Dai jiaxin 1155241401

Section 1: Introduction

With the rapid advancement of large language models (LLMs), multimodal systems capable of jointly reasoning over visual and textual inputs have become increasingly viable for real-world business applications. This assignment focuses on a multimodal receipt understanding task, where the system analyzes multiple supermarket bill images and responds to user queries related to spending amounts.

Unlike traditional receipt analysis tasks, the primary challenge does not lie in arithmetic computation, but in semantic understanding and task alignment. The system must correctly interpret heterogeneous receipt images, understand intent expressed in free-form natural language queries, and align both modalities to produce accurate and constrained answers. The system accepts multiple receipt images together with an unconstrained user query, while enforcing strict output requirements: only two predefined financial questions—total amount paid and total amount before discounts—are supported, and all other queries must be explicitly rejected. This asymmetry between flexible input and constrained output significantly increases task complexity.

Instead of relying on heavy OCR pipelines or rule-based parsers, this work leverages the reasoning and instruction-following capabilities of a multimodal LLM, using prompt engineering as the primary control mechanism. Carefully designed prompts guide the model's attention and response behavior, making prompt engineering the key determinant of system performance.

Section 2: System Overview

The overall system follows a lightweight and modular design, with prompt engineering serving as the central intelligence layer. At a high level, the processing pipeline can be summarized as:

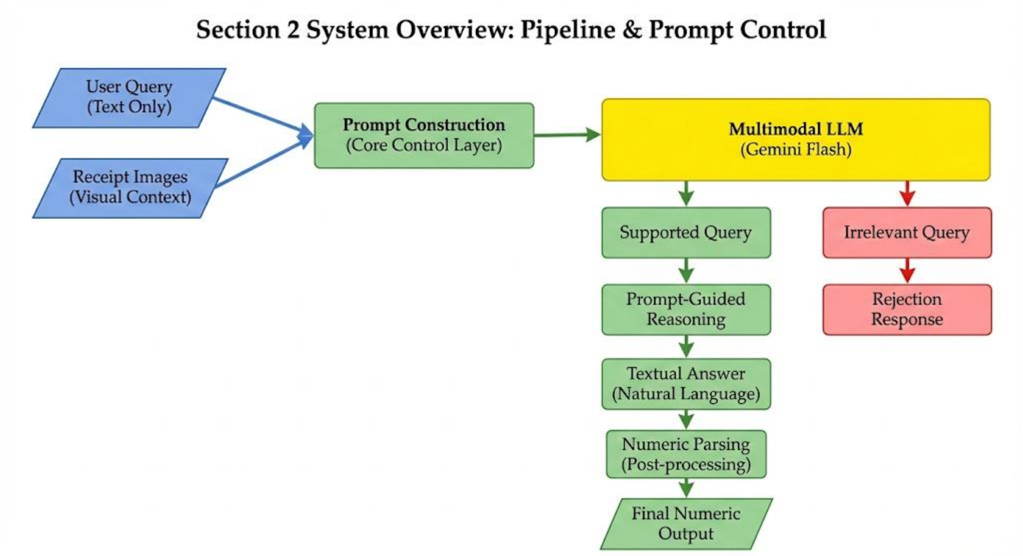
Receipt Images → Multimodal LLM (Gemini) → Prompt-Guided Reasoning → Textual Answer → Numeric Parsing

First, multiple receipt images are provided directly to a multimodal large language model. Instead of explicitly extracting structured text through OCR and post-processing heuristics, the model is expected to directly reason over raw visual inputs, including item prices, totals, and discount information. This design choice avoids the complexity and brittleness of traditional OCR pipelines, especially when dealing with diverse receipt layouts and image quality variations.

Second, all task-specific intelligence is concentrated in the prompt design. The prompt defines the task boundaries, clarifies the expected reasoning objective, and constrains the model's response behavior. Through carefully crafted natural language instructions, the model is guided to (i) distinguish between different types of user queries, (ii) focus on the relevant monetary fields in the receipts, and (iii) reject unsupported or irrelevant questions. No task-specific rules or symbolic logic are hard-coded outside the prompt.

Finally, a minimal post-processing step is applied to the model's textual output. This step uses simple numeric parsing techniques to extract floating-point values from the response. Importantly, this post-processing is not intended to correct reasoning errors or compensate for poor prompts. Its sole purpose is to ensure output stability and consistency during automated evaluation, given the natural variability of language model responses.

Overall, this system design deliberately minimizes architectural complexity and shifts the emphasis toward prompt engineering as a controllable and extensible interface between user intent, visual evidence, and model reasoning.


Section 2 System Overview: Pipeline & Prompt Control
User Query (Text Only)
Receipt Images (Visual Context)
Prompt Construction (Core Control Layer)
Multimodal LLM (Gemini Flash)
Supported Query
Irrelevant Query
Prompt-Guided Reasoning
Rejection Response
Textual Answer (Natural Language)
Numeric Parsing (Post-processing)
Final Numeric Output

Section 3: Prompt Engineering Design

Prompt engineering is the central component of the proposed system. Rather than treating the prompt as a simple instruction string, this work adopts a design-driven approach, where the prompt is explicitly engineered to control multimodal grounding, task scope, reasoning behavior, and output format. This section outlines the design objectives and explains how the prompt structure enables accurate and robust performance under strict task constraints.

3.1 Prompt Design Objectives

Before constructing the prompt, several high-level design objectives were defined to ensure predictable and task-aligned model behavior. These objectives guided the prompt design and its structural choices.

Multimodal Grounding.
The prompt must enable the model to associate the user's textual query with visual evidence from multiple receipt images, encouraging joint reasoning over text and images rather than treating them independently.

Task Disambiguation.
The system supports exactly two financial queries: total spending after discounts (Query 1) and total spending before discounts (Query 2). The prompt guides the model to distinguish between these tasks based on query semantics, without explicit flags or hard-coded logic.

Numerical Reasoning Constraint.
The prompt constrains the model to focus on numerical aggregation across receipts, prioritizing monetary values over descriptive or contextual responses.

Out-of-Domain Rejection.
The prompt instructs the model to identify and reject queries outside the supported task scope, treating rejection as a core requirement rather than an edge case.

By explicitly defining these objectives, the prompt design emphasizes intentional control and task alignment instead of ad hoc experimentation.

3.2 Multimodal Prompt Structure

A key design choice is the explicit structuring of the prompt as a natural language query followed by multiple receipt images:

Textual query → Visual evidence (multiple receipts)

Placing the query first establishes user intent as the primary reasoning anchor, guiding the model's attention when processing the images. Presenting multiple receipts in parallel enables cross-image aggregation, which is essential for answering queries involving total spending across bills. This structured input encourages joint reasoning over textual intent and visual content.

3.3 Task-Specific Prompting Strategy

The prompt differentiates between Query 1 and Query 2 through implicit linguistic cues rather than explicit rules or formulas. For Query 1, the model is guided to focus on final payable amounts, while for Query 2 it is implicitly encouraged to reason about pre-discount prices by interpreting discount-related fields. This approach leverages natural language instruction alignment, allowing task differentiation to emerge from query semantics. As a result, the system avoids hard-coded conditional logic and remains robust to variations in query phrasing.

3.4 Prompt Robustness and Output Control

Although the prompt encourages concise numeric answers, language models may still produce variable output formats. To ensure evaluation stability, a lightweight post-processing step is applied to extract numerical values from the model's response.

This design reflects a pragmatic engineering principle: prompt engineering governs reasoning behavior, while minimal post-processing ensures output consistency, without correcting or altering the model's underlying reasoning.

Section 4: Handling Irrelevant Queries

4.1 Motivation

In real-world usage, user queries are often unconstrained and may include task-irrelevant questions such as “Which supermarket is this?” or “Is this receipt expensive?”. Although reasonable from a human perspective, such queries fall outside the supported scope of this assignment.

From an evaluation standpoint, incorrectly answering irrelevant queries is penalized as heavily as producing incorrect numerical results. Therefore, the ability to reliably reject out-of-domain queries is a core system requirement. The challenge lies in distinguishing supported financial queries from descriptive or subjective ones, even though all queries are expressed in natural language and may reference the same receipt images.

4.2 Prompt-Level Rejection Strategy

To address this challenge, the system adopts a prompt-based rejection strategy rather than explicit conditional logic or manually defined query patterns. The prompt instructs the language model to assess whether a query falls within the supported task scope before attempting to reason over the receipt images.

Conceptually, the prompt includes guidance equivalent to:

If the question is not related to total spending or discount analysis, respond that the query is not supported.

This instruction functions as a soft guardrail, shifting query validation into the model's semantic reasoning process. As a result, rejection decisions are based on intent understanding rather than surface-level patterns, allowing the system to correctly reject queries involving receipt descriptions or subjective judgments without additional rule engineering.

4.3 Why Prompt-Based Rejection is Preferred

From an AI-for-business perspective, prompt-based rejection offers clear advantages over rule-based approaches. It is scalable, as it generalizes across diverse query phrasings without enumerating keywords; maintainable, since task scope updates only require prompt modification; and cost-efficient, eliminating the need for auxiliary classifiers or preprocessing pipelines.

In contrast, rule-based rejection mechanisms are brittle, difficult to extend, and costly to maintain. Embedding rejection logic directly into the prompt therefore provides a more flexible and future-proof solution.

Section 5: Limitations and Future Improvements

First, the system relies on the language model's implicit understanding of discount semantics. While the prompt successfully guides the model to distinguish between post-discount and pre-discount amounts in most cases, ambiguous receipt layouts or unconventional discount formats may still lead to incorrect interpretations.

Second, robustness may degrade when dealing with extremely low-quality images, such as heavily blurred photos or handwritten receipts. Although multimodal LLMs demonstrate strong visual reasoning capabilities, their performance remains sensitive to input clarity.

Finally, several enhancements could further improve reliability and interpretability. Future work may incorporate chain-of-thought prompting to encourage more structured reasoning, self-consistency techniques to reduce variance across model outputs, or tool-augmented parsing pipelines that combine OCR with LLM-based reasoning for improved numerical accuracy.

Overall, these limitations highlight opportunities to extend the current prompt-centric design into a more comprehensive agentic system while preserving its simplicity and flexibility.