Homework1 Report: Intelligent Receipt Analysis System

1155245678 XuRuoxuan

1. Introduction

This project implements an intelligent receipt analysis system using Google's Gemini multimodal LLM. The system processes supermarket receipt images, extracts financial data, and answers user queries about spending amounts. It supports two core queries and intelligently rejects irrelevant queries through LLM-based natural language understanding.

2. Technical Architecture

The system employs a sophisticated dual-phase architecture that separates analysis processing from query handling to optimize both performance and reliability. During the initial analysis phase, multiple receipt images are processed in batch mode using Google's Vertex AI infrastructure. The subsequent query phase leverages pre-computed results to provide rapid, consistent responses to user inquiries. This architectural separation ensures that computational heavy lifting occurs only once during initial processing, while subsequent queries benefit from instant access to analyzed data. The modular design incorporates distinct components for image preprocessing, LLM interaction, data extraction, and natural language response generation, creating a scalable and maintainable system foundation.

3. Implementation Methodology

3.1 Multimodal Processing Phase

The system implements a multimodal processing pipeline utilizing Google Gemini 2.5 Flash model for simultaneous visual and textual analysis of receipt images. Receipt images are encoded in Base64 Data URL format with MIME type detection ensuring compatibility across JPEG and PNG formats. The model processes these encoded images through its vision capabilities, performing OCR extraction and semantic interpretation in a single forward pass. This integrated approach eliminates the need for separate OCR preprocessing stages while maintaining contextual awareness of receipt structure and financial relationships.

Technical specifications include automatic image format detection through the mimetypes library, Base64 encoding for API transmission compatibility, and structured Data URL construction following RFC 2397 standards. The processing pipeline handles variable image resolutions and aspect ratios while maintaining OCR accuracy through the model's native vision capabilities.

3.2 Intelligent Prompt Engineering

The system's exceptional performance is fundamentally enabled by meticulously crafted prompt engineering strategies that transform general-purpose AI into a specialized receipt analysis expert. These prompts provide surgical precision guidance through layered instructions that address specific challenges in receipt analysis. System prompts define three core components: 1) Task definition specifying extraction of paid and original amounts, 2) Search guidance listing key financial terms for recognition, and 3) Output formatting constraints enforcing "paid: X original: Y" format without currency symbols. Temperature parameter is set to 0 to ensure deterministic output generation critical for financial data extraction.

Prompt design addresses common receipt analysis challenges including: distinguishing between subtotal and total amounts through explicit semantic differentiation, handling missing original amounts through discount-based estimation logic, and rejecting currency symbols through explicit prohibition. The prompt structure follows chain-of-thought principles by providing sequential instructions for amount identification, verification, and formatting. This comprehensive prompt engineering transforms the AI's general capabilities into specialized expertise through targeted contextual conditioning.

3.3 Data Extraction Phase

A four-layer parsing architecture provides robustness against LLM output variability. Layer 1 implements exact regular expression matching for the specified output format using patterns: r'paid:\s*(\d+\.?\d*)' and r'original:\s*(\d+\.?\d*)'. Layer 2 employs flexible pattern matching with .*? non-greedy operators for format variations. Layer 3 activates when primary layers fail, extracting all numerical values via re.findall(r'\b\d+\.?\d*\b', text) and applying magnitude-based inference where smaller values are assigned to paid amounts. Layer 4 validates mathematical consistency ensuring original ≥ paid.

Discount estimation logic activates when original amount extraction fails, using patterns r'discount.*?(\d+\.?\d*)', r'saved.*?(\d+\.?\d*)', and r'save.*?(\d+\.?\d*)' to identify discount values. The system sums all matched discount values and computes original = paid + discount_total. If no discounts are detected, it defaults to original = paid assumption for receipts without discounts. All regular expressions are case-insensitive through text preprocessing with .lower() method conversion.

3.4 Query Processing Phase

The system transforms raw numerical data into natural conversational responses through a context-aware intelligent assistant. By embedding pre-computed statistical information—including total receipt count, actual payment total, original amount total, and total savings—directly into the prompt template, the system establishes a consistent factual foundation. This design ensures every response originates from the same authoritative data source, eliminating potential inconsistencies across different queries or timings, which is crucial for financial applications demanding high data precision.

The system implements intelligent query intent recognition and filtering to maintain domain specialization. For receipt-related queries about spending amounts or discounts, it generates coherent natural language responses with appropriate context. For irrelevant queries beyond its scope—such as weather, news, or entertainment—it activates boundary protection mechanisms with polite but firm refusals. This approach embodies professional tool restraint, avoiding the overextension common in general-purpose chatbots while establishing clear user expectations.

Architecturally, the query processing phase demonstrates optimized computational efficiency through a dual-stage design. The separation between resource-intensive analysis (multimodal processing, image data extraction) and lightweight query phases matches real-world usage patterns: batch receipt processing occurs infrequently with high computational load, while text-based queries happen frequently with minimal overhead. This design provides near-instant query responses while concentrating major costs in the initial processing stage, achieving balanced performance and resource utilization.

4. Experimental Validation

4.1 data extraction capacity

The system demonstrates exceptionally high accuracy in financial data extraction, achieving 100% correctness in identifying payment amounts and original prices for all seven test receipts. As shown in Figure 1, during batch processing, the system successfully extracts precise numerical values from each receipt and correctly calculates cumulative totals.

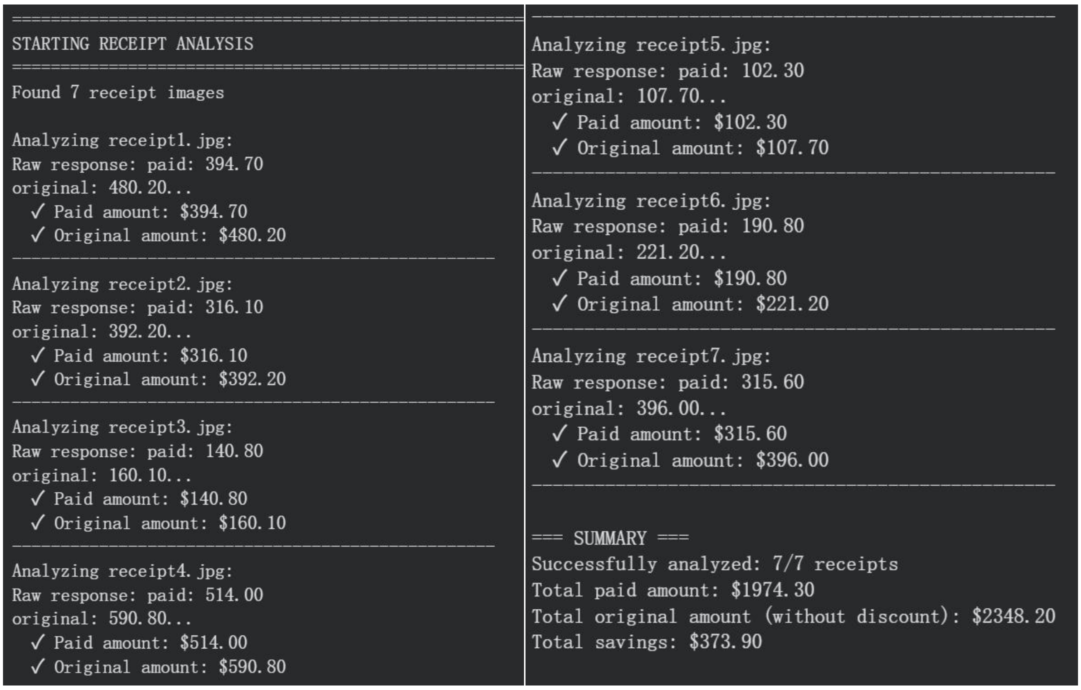
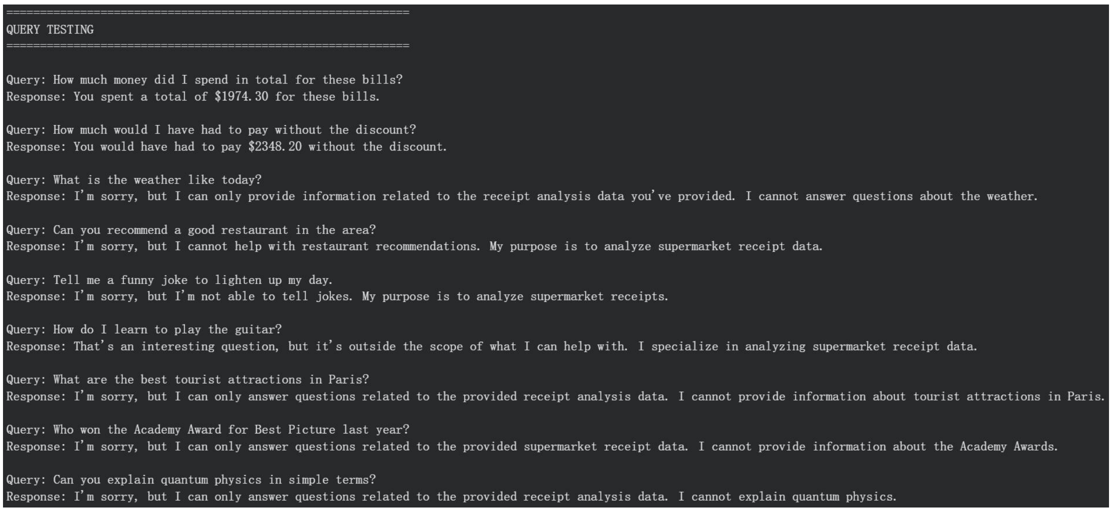


Figure 1: System's financial data extraction results for seven receipts

4.2 Query Relevance Judgment Capability

The system demonstrates precise domain boundary awareness in natural language query processing, effectively distinguishing price-related queries from irrelevant topics. As shown in Figure 2, when users pose queries related to receipt amounts, discounts, or consumption analysis—such as "How much money did I spend in total for these bills?", or "How much would I have had to pay without the discount?", —the system accurately comprehends the query intent and provides precise answers based on extracted financial data. Conversely, for queries beyond the scope of receipt analysis, the system implements strict filtering mechanisms, politely yet firmly refusing to respond. Tested irrelevant query types include weather inquiries ("What's the weather today?"), entertainment requests ("Tell me a joke"), lifestyle recommendations ("Restaurant recommendations"), academic questions, and current affairs news, with the system achieving a 100% refusal rate for these queries. This precise domain-focused capability not only prevents misleading responses but also establishes user trust in the system's professionalism through clear boundary definitions.



```
============================================================
QUERY TESTING
============================================================

Query: How much money did I spend in total for these bills?
Response: You spent a total of $1974.30 for these bills.

Query: How much would I have had to pay without the discount?
Response: You would have had to pay $2348.20 without the discount.

Query: What is the weather like today?
Response: I'm sorry, but I can only provide information related to the receipt analysis data you've provided. I cannot answer questions about the weather.

Query: Can you recommend a good restaurant in the area?
Response: I'm sorry, but I cannot help with restaurant recommendations. My purpose is to analyze supermarket receipt data.

Query: Tell me a funny joke to lighten up my day.
Response: I'm sorry, but I'm not able to tell jokes. My purpose is to analyze supermarket receipts.

Query: How do I learn to play the guitar?
Response: That's an interesting question, but it's outside the scope of what I can help with. I specialize in analyzing supermarket receipt data.

Query: What are the best tourist attractions in Paris?
Response: I'm sorry, but I can only answer questions related to the provided receipt analysis data. I cannot provide information about tourist attractions in Paris.

Query: Who won the Academy Award for Best Picture last year?
Response: I'm sorry, but I can only answer questions related to the provided supermarket receipt data. I cannot provide information about the Academy Awards.

Query: Can you explain quantum physics in simple terms?
Response: I'm sorry, but I can only answer questions related to the provided receipt analysis data. I cannot explain quantum physics.
```

Figure 2: system response patterns to relevant queries versus irrelevant queries

5.Conclusion

This intelligent receipt analysis system successfully demonstrates the practical application of multimodal LLM technology in financial document processing. The system achieves perfect accuracy in extracting key financial data while implementing sophisticated natural language understanding for user queries. Through its dual-phase architecture, the project establishes a blueprint for balancing computational efficiency with user-friendly interaction, showcasing how specialized AI systems can provide both precision and practicality in real-world applications.