1.Experimental Setup and Data

1.1 Dataset

Source: Downloaded the receipts.zip file from Google Drive
Content: 7 receipt images (receipt1.jpg to receipt7.jpg)
Format: JPEG format, containing complete receipt information

1.2 Test 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?"
Irrelevant Query Test: "123" (any non-relevant string)

1.3 Evaluation Metrics

Accuracy: Absolute error between calculated total and actual value
Tolerance Range: ±$2 (considering reasonable error margin for receipt recognition)
Rejection Rate: Proportion of irrelevant queries correctly rejected

2.Experimental Results and Analysis

2.1 Accuracy Test Results

Query Type	Expected Total (USD)	System Output (USD)	Absolute Error	Pass/Fail
Query 1	1,974.3	1,974.3	0.0	√
Query 2	2,348.2	2,348.2	0.0	√

Detailed Receipt Parsing Results:

Receipt No.	Actual Payment (USD)	Original Price (USD)

Receipt No.	Actual Payment (USD)	Original Price (USD)
1	394.7	480.20
2	316.1	392.20
3	140.8	160.10
4	514.0	590.80
5	102.3	107.70
6	190.8	221.20
7	315.6	396.00
Total	1,974.3	2,348.2

2.2 Rejection Mechanism Test

Test Query	Expected Behavior	Actual Behavior	Result

Test Query	Expected Behavior	Actual Behavior	Result
"123"	Reject and return prompt message	"I can only answer questions about receipt totals."	√
"What is the weather?"	Reject	Same rejection response	√
"Tell me a joke"	Reject	Same rejection response	√

2.3 Performance Analysis

Processing Speed: Average 2-3 seconds per receipt parsing
API Calls: 7 receipts + intent recognition = 8 calls
Resource Usage: Stable memory consumption, no significant peaks
Error Recovery: Failure in parsing a single receipt does not affect overall process