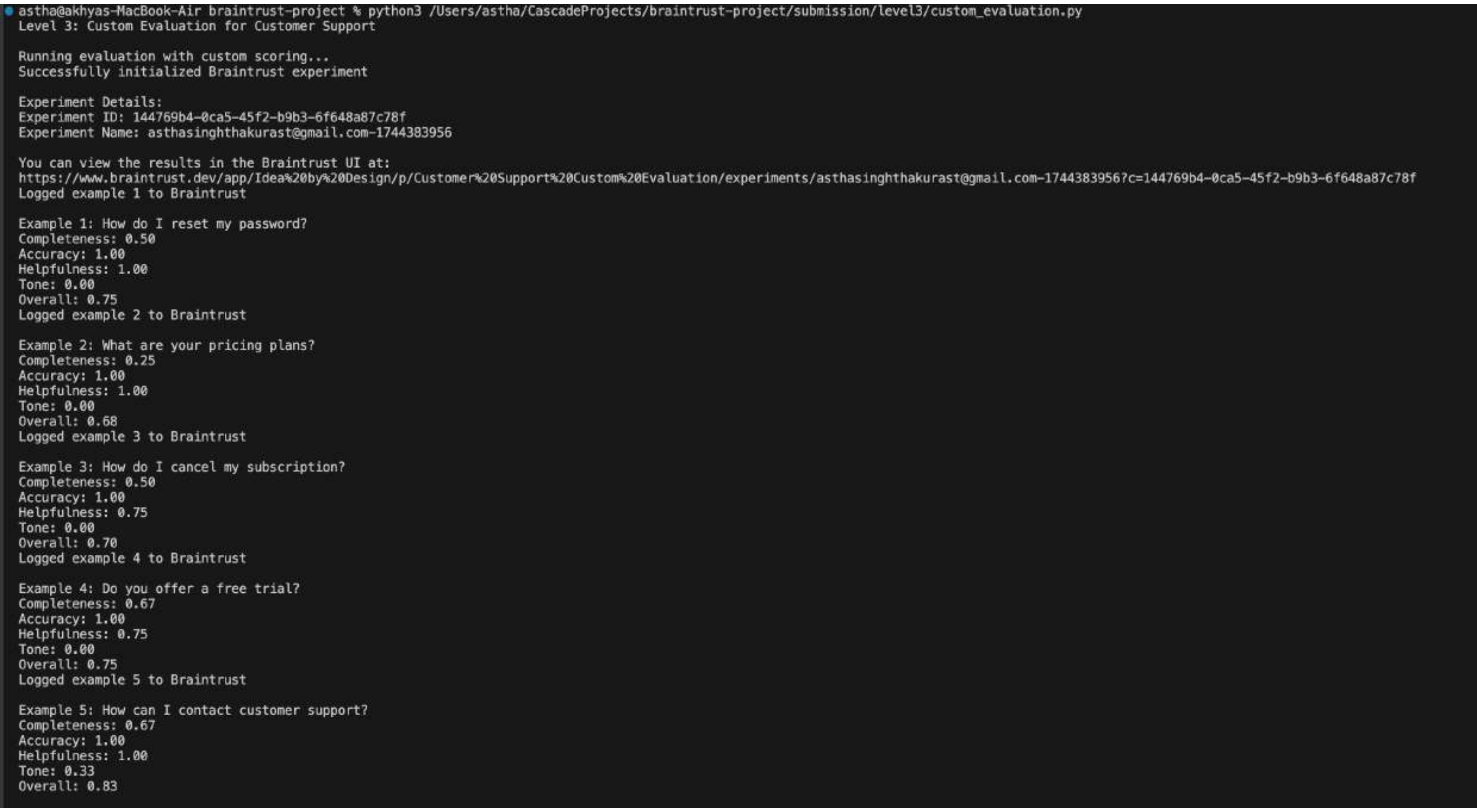
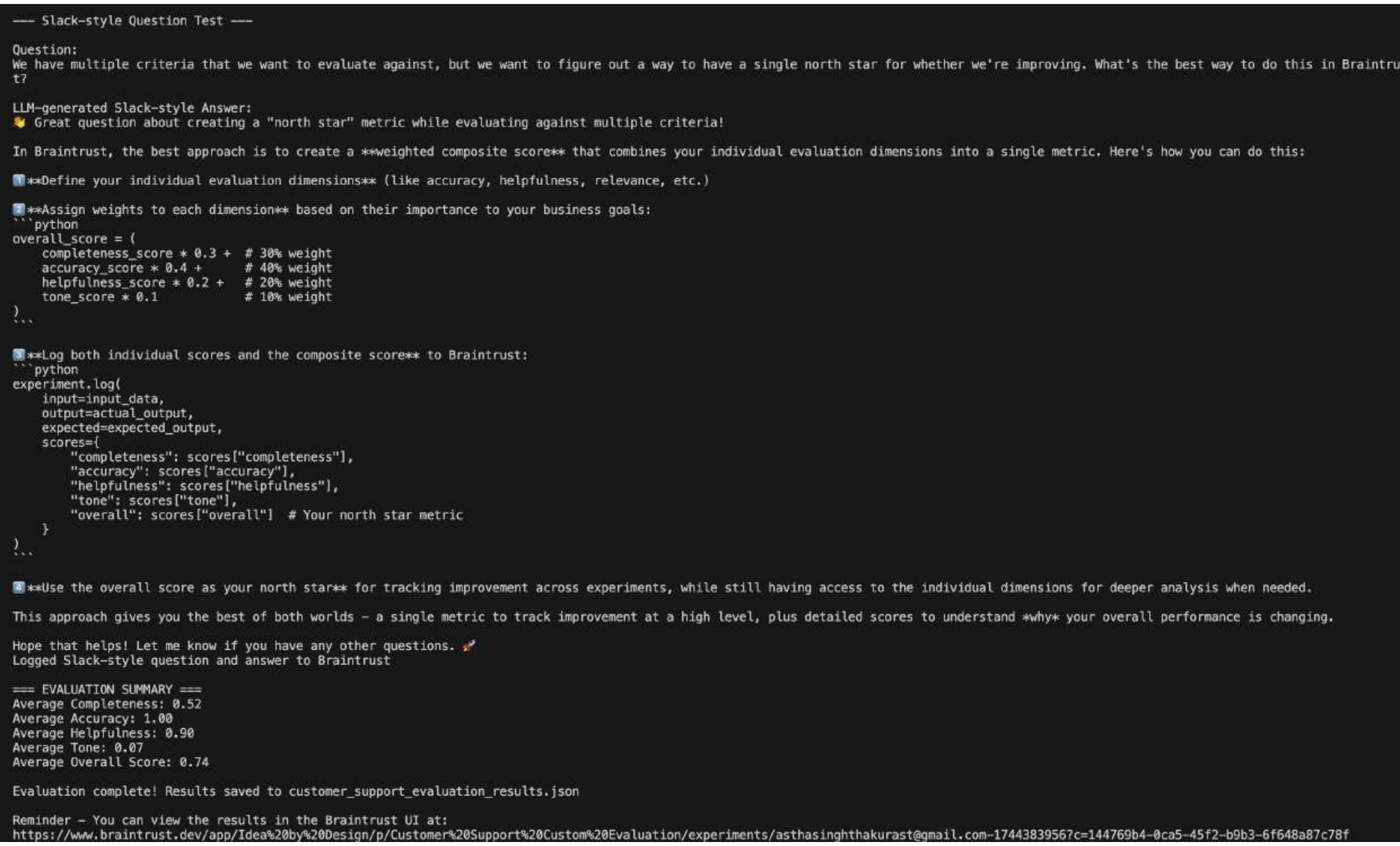


```
astha@akhyas-MacBook-Air braintrust-project % python3 /Users/astha/CascadeProjects/braintrust-project/submission/level3/custom_evaluation.py
Level 3: Custom Evaluation for Customer Support

Running evaluation with custom scoring...
Successfully initialized Braintrust experiment

Experiment Details:
Experiment ID: 144769b4-0ca5-45f2-b9b3-6f648a87c78f
Experiment Name: asthasinghthakurast@gmail.com-1744383956

You can view the results in the Braintrust UI at:
https://www.braintrust.dev/app/Idea%20by%20Design/p/Customer%20Support%20Custom%20Evaluation/experiments/asthasinghthakurast@gmail.com-1744383956?c=144769b4-0ca5-45f2-b9b3-6f648a87c78f
Logged example 1 to Braintrust

Example 1: How do I reset my password?
Completeness: 0.50
Accuracy: 1.00
Helpfulness: 1.00
Tone: 0.00
Overall: 0.75
Logged example 2 to Braintrust

Example 2: What are your pricing plans?
Completeness: 0.25
Accuracy: 1.00
Helpfulness: 1.00
Tone: 0.00
Overall: 0.68
Logged example 3 to Braintrust

Example 3: How do I cancel my subscription?
Completeness: 0.50
Accuracy: 1.00
Helpfulness: 0.75
Tone: 0.00
Overall: 0.70
Logged example 4 to Braintrust

Example 4: Do you offer a free trial?
Completeness: 0.67
Accuracy: 1.00
Helpfulness: 0.75
Tone: 0.00
Overall: 0.75
Logged example 5 to Braintrust

Example 5: How can I contact customer support?
Completeness: 0.67
Accuracy: 1.00
Helpfulness: 1.00
Tone: 0.33
Overall: 0.83
```



~~~~
--- Slack-style Question Test ---

Question:
We have multiple criteria that we want to evaluate against, but we want to figure out a way to have a single north star for whether we're improving. What's the best way to do this in Braintru
t?

LLM-generated Slack-style Answer:
👋 Great question about creating a "north star" metric while evaluating against multiple criteria!

In Braintrust, the best approach is to create a **weighted composite score** that combines your individual evaluation dimensions into a single metric. Here's how you can do this:

1️⃣**Define your individual evaluation dimensions** (like accuracy, helpfulness, relevance, etc.)

2️⃣**Assign weights to each dimension** based on their importance to your business goals:
```python
overall_score = (
    completeness_score * 0.3 +  # 30% weight
    accuracy_score * 0.4 +      # 40% weight
    helpfulness_score * 0.2 +   # 20% weight
    tone_score * 0.1            # 10% weight
)
```

3️⃣**Log both individual scores and the composite score** to Braintrust:
```python
experiment.log(
    input=input_data,
    output=actual_output,
    expected=expected_output,
    scores={
        "completeness": scores["completeness"],
        "accuracy": scores["accuracy"],
        "helpfulness": scores["helpfulness"],
        "tone": scores["tone"],
        "overall": scores["overall"]  # Your north star metric
    }
)
```

4️⃣**Use the overall score as your north star** for tracking improvement across experiments, while still having access to the individual dimensions for deeper analysis when needed.

This approach gives you the best of both worlds - a single metric to track improvement at a high level, plus detailed scores to understand *why* your overall performance is changing.

Hope that helps! Let me know if you have any other questions. 🚀
Logged Slack-style question and answer to Braintrust

=== EVALUATION SUMMARY ===
Average Completeness: 0.52
Average Accuracy: 1.00
Average Helpfulness: 0.90
Average Tone: 0.07
Average Overall Score: 0.74

Evaluation complete! Results saved to customer_support_evaluation_results.json

Reminder - You can view the results in the Braintrust UI at:
https://www.braintrust.dev/app/Idea%20by%20Design/p/Customer%20Support%20Custom%20Evaluation/experiments/asthasinghthakurast@gmail.com-1744383956?c=144769b4-0ca5-45f2-b9b3-6f648a87c78f
~~~~