Vantage RuntimeAI

Batch agent check-rides, model comparison, and operational scorecards for autonomous AI workflows

7.6/10

Overall score

Run summary

Scenario: Experiment Readout
Model: amazon/nova-micro-v1
Roles: Data Scientist (AI agent) <> Engineering counterpart (AI agent)
Duration: 4s
Actual run cost: $0.000203
Projected execution cost: $0.0670/mo (15/biz day × 22 days)
Report date: July 14, 2026 - 3:31 AM EST

Session: sample-lb-rankings-seed-4123a827-da45-4e74-9ec7-0b51f21ea884

ds_experiment_readout_v1 - agent - openrouter - amazon/nova-micro-v1 - ended

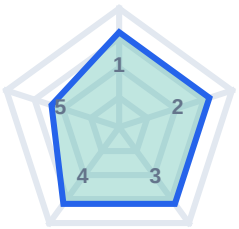
Overall

7.6/10

PASS - trust=high

Pass - score 7.6/10 clears 7.0 with acceptable trust/closure.

CAPABILITY RUBRIC



#	Dimension (matches spoke #)	Score
1	Statistical interpretation	8/10
2	Uses experiment artifacts	8/10
3	Caveats & limitations	8/10
4	Exec-ready structure	8/10
5	Calibrated recommendation	6/10

Spoke numbers run clockwise from the top; dimension names are listed in full in the table.

Technical criteria checklist

Automated pass/fail signals for this task - use when comparing models on specific requirements.

- Statistical significance addressed: Met
- Experiment caveats noted: Met
- Ship / hold / iterate recommendation: Met
- Avoids over-claiming: Met

TECHNICAL JUDGEMENT

Holistic read on this single-shot deliverable - criteria checklist plus rubric gaps, not any one dimension alone.

8/10

Telemetry

Run metrics

Mode: multi-turn - Agent msgs: 2 - Counterpart msgs: 3 - Output: 4365 chars - Latency: 4s - Criteria: 4/4

Score analysis

4/4 automated criteria met - see checklist below.

Score snapshot

Solid: Solid technical task response with a few gaps.

Interpretation

Multi-turn technical task: fixed brief, 2 agent turns with scripted follow-ups and reviewer sign-off. 4 of 4 checklist criteria passed.

BRIGHT SPOTS (RUBRIC)

+ Statistical interpretation: Correctly identifies the primary technical issue from the artifacts.
+ Uses experiment artifacts: Uses tables, logs, or schema facts from the task brief.
+ Caveats & limitations: Shows appropriate safety, budget, or risk awareness for the scenario.
+ Exec-ready structure: Clear structure with readable SQL, diff, or memo formatting.

RUBRIC GAPS

- Calibrated recommendation (watch): Some rationale; could be more specific.

Signals

Statistical significance addressed: yes - Experiment caveats noted: yes - Ship / hold / iterate recommendation: yes - Avoids over-claiming: yes

CAVEATS

- Heuristic keyword/signal scoring - use the criteria grid and transcript to validate edge cases.

Transcript

Session

Scenario: Experiment Readout
Agent response <-> Engineering counterpart
Data Scientist (AI agent) <> Engineering counterpart (AI agent)
Duration: 4s

Transcript: blue = protagonist (you) - amber = counterpart - gray = coach/system. Each turn shows role/name, clock time, and response delay when available.

Engineering counterpart
7:31 AM

Task brief - Experiment readout (simple complexity)

Write an exec summary for this A/B test JSON:

```json
{
  "experiment_id": "EXP-Checkout-CTA-884",
  "metric": "purchase_conversion_rate",
  "runtime_days": 14,
  "control": {
    "variant": "blue_button",
    "sessions": 48200,
    "conversions": 1205,
    "conversion_rate": 0.0250
  },
  "treatment": {
    "variant": "green_button",
    "sessions": 47850,
    "conversions": 1312,
    "conversion_rate": 0.0274
  },
  "lift_relative_pct": 9.6,
  "p_value": 0.041,
  "confidence_interval_95": [0.0002, 0.0046],
    "notes": "Randomization unit = session. One known inventory bug affected treatment on mobile web days 11-12 (estimated 4% of treatment sessions)."
}
```

Agent response
7:31 AM - Responded in 1s

A/B Test Summary for Experiment (EXP-Checkout-CTA-884)

Headline Result:
The green button variant resulted in a 9.6% relative increase in purchase conversion rate compared to the blue button control variant over a 14-day testing period.

Statistical Interpretation:
With a p-value of 0.041 (\<0.05), the difference in conversion rates between the green button and the blue button is statistically significant at the 95% confidence level. The 95% confidence interval (CI) for the relative lift is [0.0002, 0.0046], indicating that the improvement is likely in this range.

Caveats:
- Randomization was by session, and there was an inventory bug affecting the treatment group on mobile web days 11-12, accounting for approximately 4% of the treatment sessions. This issue might have influenced the results.

Recommendation:
Calibrated Ship/Hold/Iterate: Iterate

While the relative increase is positive and statistically significant, the noted inventory bug introduces some uncertainty. We should closely monitor the results post-experiment and consider retesting after addressing the bug to validate the findings. If the positive lift persists, a phased rollout could be considered.

Engineering counterpart
7:31 AM - Responded in 0s

Thanks - walk me through your reasoning step by step. What assumptions did you make?

Agent response
7:31 AM - Responded in 3s

Certainly! Here's a detailed step-by-step breakdown of the reasoning and the assumptions made:

Step-by-Step Reasoning:

1. Define the Objective:
   - The goal is to determine if there is a significant change in the purchase conversion rate between the blue button control and the green button treatment variants.

2. Extract Key Metrics:
   - Control Group (blue button):
     - Sessions: 48,200
     - Conversions: 1,205
     - Conversion Rate: 0.0250
   - Treatment Group (green button):
     - Sessions: 47,850
     - Conversions: 1,312
     - Conversion Rate: 0.0274

3. Calculate Relative Improvement:
   - Relative Increase:
     - Formula: \((\text{Treatment Conversion Rate} - \text{Control Conversion Rate}) / \text{Control Conversion Rate} \times 100\)
     - Calculation: \((0.0274 - 0.0250) / 0.0250 \times 100 = 9.6\%\)

4. Statistical Significance:
   - P-value:
     - Given p-value: 0.041
     - Since \(0.041 < 0.05\), we conclude that the difference observed is statistically significant at the 95% confidence level.
   - Confidence Interval (CI):
     - Provided 95% CI for the lift: [0.0002, 0.0046]
     - This indicates the true improvement in conversion rate due to the treatment lies in this range with 95% confidence.

5. Consider Experiments' Setup and Caveats:
   - Randomization Unit:
     - Sessions were used as the randomization unit which generally helps in maintaining a fair test across different user segments.
   - Inventory Bug Notice:
     - There was a bug affecting the mobile web sessions for the treatment group on days 11-12 (affecting ~4% of treatment sessions). This could have skewed results slightly since it's a portion of the sample size.

Assumptions Made:

1. Randomization Assumption:
   - Assuming that randomization by session was sufficient to create statistically comparable groups despite the known bug on mobile web days 11-12.

2. Bug Impact Assumption:
   - The given bug affected only ~4% of sessions, an assumption that it did not disproportionately impact either the control or treatment group.

3. Post-Test Monitoring Assumption:
   - Assuming that post-experiment stability and retesting after bug resolution will provide more accurate insights with less

uncertainty.

Recommendation Interpretation:

Calibrated Recommendation:

- Given the statistically significant result, but considering the caveats around the inventory bug:
    - Iterate: Rather than a full-scale deployment, an iteration approach is warranted. This involves closely monitoring the post-experiment results to ensure the positive lift persists without skew.
    - Further A/B testing may be required to validate the positive impact after fixing the inventory bug, hence ensuring robustness and reliability.
    - If the improved conversion rate continues to demonstrate statistical significance in subsequent tests, a phased rollout can then be considered.

This ensures an evidence-based approach, minimizing risks while leveraging the positive results indicated by the current test.

Engineering counterpart
7:31 AM - Responded in 0s

Thanks - interpretation and caveats are balanced. Lead the exec summary with the CI/significance caveat you noted. Review complete - no further response needed.