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: openai/gpt-4o-mini
Roles: Data Scientist (AI agent) <> Engineering counterpart (AI agent)
Duration: 13s
Actual run cost: $0.000870
Projected execution cost: $0.2871/mo (15/biz day × 22 days)
Report date: July 13, 2026 - 11:27 PM EST

Session: sample-lb-rankings-seed-584d0055-5a43-44cd-bccf-a99ce6555176

ds_experiment_readout_v1 - agent - openrouter - openai/gpt-4o-mini - ended

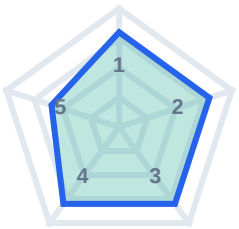
Overall

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: 3824 chars - Latency: 13s - 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: 13s

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

Engineering counterpart
3:27 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
3:27 AM - Responded in 5s

Headline Result:
The green button variant resulted in a 9.6% relative increase in purchase conversion rate compared to the blue button control.

Statistical Interpretation:
The treatment group achieved a conversion rate of 2.74% versus 2.50% in the control group, with a p-value of 0.041, indicating statistical significance at the 5% level. The 95% confidence interval for the increase in conversion rate ranges from 0.02% to 0.46%, suggesting a meaningful positive effect of the green button.

Caveats:
It is important to note that an inventory bug impacted the treatment group during days 11-12, affecting roughly 4% of treatment sessions. This could introduce bias, potentially overestimating the treatment's performance.

Recommendation:
Iterate. Given the statistical significance and positive lift, we recommend further testing with a refined approach to mitigate the impact of the known bug, ensuring robust results before full-scale implementation.

Engineering counterpart
3:27 AM - Responded in 0s

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

Agent response
3:27 AM - Responded in 8s

Certainly! Here's a step-by-step breakdown of my reasoning along with the assumptions made during the analysis:

1. Understanding the Experiment Framework:
- The experiment tested two variants for a "Checkout Call-to-Action" (CTA) button: a control (blue button) and a treatment (green button).
- The goal was to assess the impact of these variants on the "purchase conversion rate."

2. Evaluating the Metrics:
- Conversion Rates: The blue button had a conversion rate of 2.50% and the green button had a conversion rate of 2.74%. I calculated the relative lift of 9.6% based on these figures.
- Sample Size: The sample sizes for both variants are relatively large (48,200 sessions for control and 47,850 for treatment), which typically supports more reliable statistical conclusions.

3. Statistical Significance:
- The p-value of 0.041 indicates that there is a statistically significant difference between the two variants at the 5% significance level. This suggests that there is less than a 4.1% chance that the observed difference is due to random variation alone.
- Confidence Interval: The 95% confidence interval of [0.0002, 0.0046] indicates that we can be 95% confident that the true increase in conversion rate lies within this range. Since the interval is above 0, it supports the treatment's positive effect.

4. Caveats and Limitations:
- The known inventory bug affecting the treatment group during specific days is a critical consideration. It is assumed that this bug could have skewed the treatment results and might have contributed to an inflated conversion rate. This acknowledgment is crucial for interpreting the results accurately.

5. Formulating Recommendations:
- Based on the statistical significance and observed lift, I decided to recommend iteration rather than immediate implementation. The reasoning here is that while the results are promising, the bug's effect could compromise the reliability of these findings. Iterating would involve further testing without the bug's interference, thereby providing clearer insights.

6. Overall Assumptions:
- It is assumed that the randomization was effectively implemented, meaning that sessions were assigned to treatment and control groups without bias.
- The assumption is also made that the sample sizes are sufficient to generalize findings to the broader user population.
- Finally, it is assumed that the only significant factor affecting the conversion rate was the button color, excluding other potential external influences.

In summary, the reasoning was based on analyzing the metrics for statistical significance, considering potential external factors (like the inventory bug), and making a cautiously optimistic recommendation for iteration to confirm results under better conditions.

Engineering counterpart
3:27 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.