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: meta-llama/llama-4-maverick
Roles: Data Scientist (AI agent) <> Engineering counterpart (AI agent)
Duration: 42s
Actual run cost: $0.001160
Projected execution cost: $0.3828/mo (15/biz day × 22 days)
Report date: July 14, 2026 - 3:32 AM EST

Session: sample-lb-rankings-seed-a364bbac-4e7c-4b33-a52a-bb9c2387de04

ds_experiment_readout_v1 - agent - openrouter - meta-llama/llama-4-maverick - ended

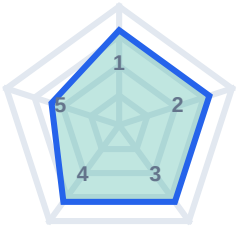
Overall

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: 3499 chars - Latency: 42s - 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: 42s

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 17s

Experiment EXP-Checkout-CTA-884 Readout

Headline Result: Replacing the blue checkout CTA button with green (Treatment) resulted in a 9.6% relative lift in purchase conversion rate.

Statistical Interpretation: The observed lift is statistically significant (p-value = 0.041) with a 95% confidence interval suggesting the true lift is between +0.2ppt and +0.46ppt.

Caveats: A known inventory bug on mobile web during days 11-12 affected approximately 4% of treatment sessions, potentially dampening the true treatment effect.

Recommendation: Ship with Caution. The green button variant is likely to drive a meaningful increase in purchase conversion rate. However, we should monitor the impact of the change on other metrics and revisit the inventory bug fix to ensure the full potential of the green CTA is realized.

Engineering counterpart
7:31 AM - Responded in 0s

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

Agent response
7:32 AM - Responded in 25s

Here's a step-by-step breakdown of my thought process:

1. Understanding the experiment context: The A/B test (EXP-Checkout-CTA-884) compares two variants of a checkout CTA button: blue (control) and green (treatment). The primary metric of interest is the purchase conversion rate.

2. Assessing the results:
The treatment (green button) showed a 9.6% relative lift in purchase conversion rate compared to the control (blue button).
The absolute conversion rates were 2.50% for control and 2.74% for treatment.

3. Statistical interpretation:
The p-value (0.041) indicates that the observed difference is statistically significant at the conventional 5% significance level. This suggests that the observed lift is unlikely to be due to chance.
The 95% confidence interval ([0.0002, 0.0046]) provides a range of plausible values for the true lift. This interval suggests that the true lift is likely to be between +0.2ppt and +0.46ppt.

4. Considering caveats and potential biases:
The notes section mentions a known inventory bug that affected approximately 4% of treatment sessions on mobile web during days 11-12. This could have potentially dampened the treatment effect.
I assumed that the bug's impact was not catastrophic (i.e., it didn't completely negate the treatment effect) and that the estimated 4% affected sessions is a reasonable approximation.

5. Making a recommendation:
Based on the statistically significant result and the confidence interval suggesting a positive lift, I leaned towards recommending that the treatment be shipped.
However, considering the potential impact of the inventory bug, I added a "with caution" caveat to the recommendation.
I also suggested monitoring the impact on other metrics (not reported in the JSON) to ensure that the change doesn't have unintended consequences.

Assumptions made:

1. The randomization unit (session) was properly implemented, and the experiment was correctly executed.
2. The estimated 4% affected sessions due to the inventory bug is accurate.
3. The bug's impact was not severe enough to completely negate the treatment effect.
4. The experiment runtime (14 days) was sufficient to capture the relevant dynamics.

Implicit considerations:

1. The business context and potential costs/benefits of shipping the treatment were not explicitly stated, but I assumed that a 9.6% relative lift in purchase conversion rate is likely to be beneficial.
2. I considered the potential risks associated with shipping a change that may have been affected by a bug, and balanced this against the potential benefits of the treatment.

Engineering counterpart
7:32 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.