



Vantage RuntimeAI

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

6.4/10

Overall score

Run summary

Scenario: Experiment Readout

Model: meta-llama/llama-3.3-70b-instruct

Roles: Data Scientist (AI agent) <> Engineering counterpart (AI agent)

Duration: 1m 14s

Actual run cost: \$0.000670

Projected execution cost: \$0.2211/mo (15/biz day × 22 days)

Report date: July 14, 2026 - 3:32 AM EST

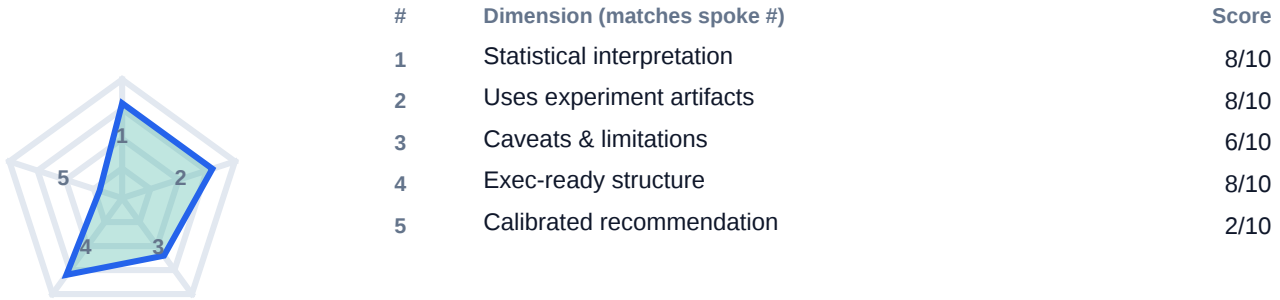
Overall

6.4/10

FAIL (decision gate) - below_pass_line - trust=high

Fail - score 6.4/10 below pass line 7.0.

CAPABILITY RUBRIC



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: Not met

TECHNICAL JUDGEMENT

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

6/10

Quick signals (good vs bad)

TRANSCRIPT FLAGS

- Overclaims certainty despite known experiment contamination.

Telemetry

Run metrics

Mode: multi-turn - Agent msgs: 2 - Counterpart msgs: 3 - Output: 3691 chars - Latency: 74s - Criteria: 3/4

Score analysis

3/4 automated criteria met - see checklist below.

Score snapshot

Developing: Mixed technical task quality - several criteria only partially met.

Interpretation

Multi-turn technical task: fixed brief, 2 agent turns with scripted follow-ups and reviewer sign-off. 3 of 4 checklist criteria passed. Primary gap: Overclaims certainty despite known experiment contamination.

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.
- + Exec-ready structure: Clear structure with readable SQL, diff, or memo formatting.

RUBRIC GAPS

- Caveats & limitations (watch): Some risk awareness; a few gaps remain.
- Calibrated recommendation (critical gap): Weak or missing rationale / next steps.

Signals

Statistical significance addressed: yes - Experiment caveats noted: yes - Ship / hold / iterate recommendation: yes - Avoids over-claiming: no

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: 1m 14s

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 8s

Experiment EXP-Checkout-CTA-884 Readout

Headline Result: Replacing the blue button with a green button in the checkout process led to a 9.6% relative increase in purchase conversion rate.

Statistical Interpretation: This result is statistically significant (p-value: 0.041) with a 95% confidence interval of [0.02%, 0.46%]. This suggests that the observed increase is unlikely due to chance.

Caveats: The experiment was affected by an inventory bug on mobile web for 2 days, impacting approximately 4% of treatment sessions. This may have introduced some noise into the results.

Recommendation: Based on the significant lift and relatively narrow confidence interval, I recommend shipping the green button variant, but iterating to investigate and mitigate the impact of the inventory bug on mobile web to ensure accurate future measurements. Close monitoring of performance post-launch is advised to confirm the observed effect.

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 1m 6s

Here's a step-by-step breakdown of my reasoning:

1. Understanding the experiment setup:

I assumed that the experiment was a randomized controlled trial (RCT) with two variants: a control group (blue button) and a treatment group (green button).

I noted that the randomization unit was a session, meaning that each session was randomly assigned to either the control or treatment group.

2. Interpreting the headline result:

I calculated the relative percentage change in conversion rate between the treatment and control groups: $((0.0274 - 0.0250) / 0.0250) \times 100\% \approx 9.6\%$.

I assumed that the increase in conversion rate was the primary outcome of interest.

3. Evaluating statistical significance:

I examined the p-value (0.041) and noted that it was below the typical significance threshold of 0.05.

I assumed that the p-value was calculated using a suitable statistical test (e.g., two-sample z-test or t-test) to compare the conversion rates between the two groups.

Based on the p-value, I concluded that the observed difference in conversion rates was statistically significant, meaning that it's unlikely due to chance.

4. Considering the confidence interval:

I examined the 95% confidence interval (0.0002, 0.0046) and noted that it did not cross zero.

I assumed that the confidence interval was calculated using a suitable method (e.g., asymptotic or bootstrap methods).

The confidence interval suggests that the true effect size is likely to be between 0.02% and 0.46%, which is a relatively narrow range.

5. Addressing caveats and limitations:

I noted the mention of an inventory bug affecting the treatment group on mobile web for 2 days, which may have introduced some noise into the results.

I assumed that the bug's impact was limited to the specified timeframe and device type.

6. Forming a recommendation:

Based on the significant lift and relatively narrow confidence interval, I recommended shipping the green button variant.

However, I also suggested iterating to investigate and mitigate the impact of the inventory bug on mobile web to ensure accurate future measurements.

I advised close monitoring of performance post-launch to confirm the observed effect and detect any potential issues.

Assumptions made:

The experiment was properly randomized and controlled.

The sample sizes were sufficient to detect the observed effect.

The statistical tests and confidence interval calculations were appropriate for the data.

The inventory bug's impact was limited and did not significantly bias the results.

The results generalize to the broader population and are not specific to the experimental timeframe or conditions.

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.