



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-3.1-8b-instruct

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

Duration: 9s

Actual run cost: \$0.000206

Projected execution cost: \$0.0680/mo (15/biz day × 22 days)

Report date: July 13, 2026 - 11:26 PM EST

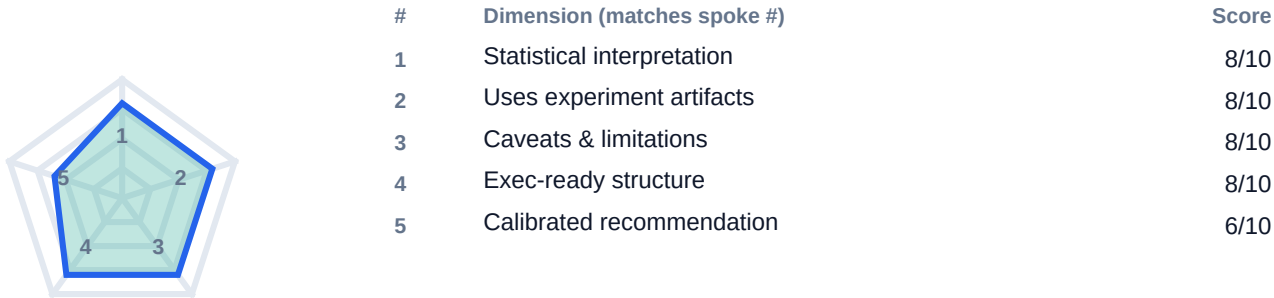
Overall

7.6/10

PASS - trust=high

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

CAPABILITY RUBRIC



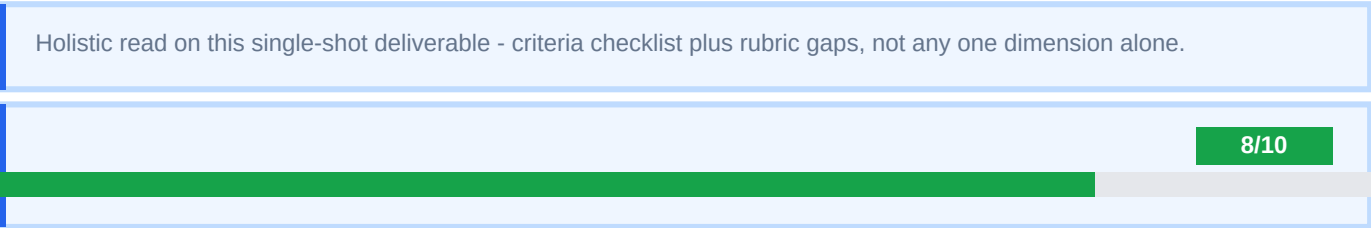
Spoke numbers run clockwise from the top; dimension names are listed in full in the table.

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: Met

TECHNICAL JUDGEMENT



Telemetry

Run metrics

Mode: multi-turn - Agent msgs: 2 - Counterpart msgs: 3 - Output: 3019 chars - Latency: 9s - 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: 9s

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

Engineering counterpart

3:26 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

3:26 AM - Responded in 3s

Here is the exec summary:

Headline Result:

Treatment with the green "green_button" CTA resulted in a 9.6% relative lift in purchase conversion rate over the blue "blue_button" control.

Statistical Interpretation:

The difference in purchase conversion rates between the treatment and control groups is statistically significant (p-value = 0.041), with a 95% confidence interval of [0.0002%, 0.0046%]. This indicates that the observed lift is likely to be real and not due to chance.

Caveats:

- Randomization unit was at the session level.
- Inventory bug on mobile web affected treatment sessions on days 11-12, potentially impacting 4% of treatment sessions.

Recommendation:

Ship the updated CTA button with the green design to production, as the lift is statistically significant and appears to be a real improvement. However, keep a close eye on mobile web traffic to mitigate any potential inventory-related issues.

Engineering counterpart

3:26 AM - Responded in 0s

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

Agent response

3:26 AM - Responded in 5s

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

1. Headline Result:

I started by highlighting the key finding: a 9.6% relative lift in purchase conversion rate. This is a clear and concise way to communicate the main effect of the test.

Assumptions:

- I assumed that the experiment's primary objective was to measure the impact of the CTA button design on purchase conversion rate.

2. Statistical Interpretation:

Next, I interpreted the results in a statistical context.

Assumptions:

- I assumed that the experiment was properly randomized, which is implicitly stated in the notes but not explicitly in the JSON data. If the randomization was flawed, the results might not be reliable.
- I took the p-value at face value (0.041) to determine statistical significance. I assumed that the experiment had sufficient power to detect a meaningful effect, given the sample size and runtime.
- I considered the 95% confidence interval [0.0002%, 0.0046%] to determine the plausible range of the true effect size. This indicated that the observed lift is likely to be real, as the interval does not contain zero.

3. Caveats:

I highlighted two potential limitations of the experiment.

Assumptions:

- I assumed that the inventory bug on mobile web affected a small proportion of treatment sessions (4%), which I considered relatively minor.
- I assumed that the bug did not introduce a large or systematic bias that would affect the overall interpretation of the results.

4. Recommendation:

Based on the analysis, I recommended shipping the updated CTA button to production. I considered the following factors:

- The statistically significant and non-trivial lift (9.6%) justified the recommendation.
- The inventory bug was a minor issue that could be monitored and addressed separately.

Assumptions:

- I assumed that the benefits of the updated CTA button outweigh the potential risks of shipping with a known inventory issue.
- I assumed that the team could effectively monitor and mitigate the impact of the inventory bug on mobile web traffic.

Engineering counterpart

3:26 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.