



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: sao10k/l3-lunaris-8b

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

Duration: 11s

Actual run cost: \$0.000155

Projected execution cost: \$0.0511/mo (15/biz day × 22 days)

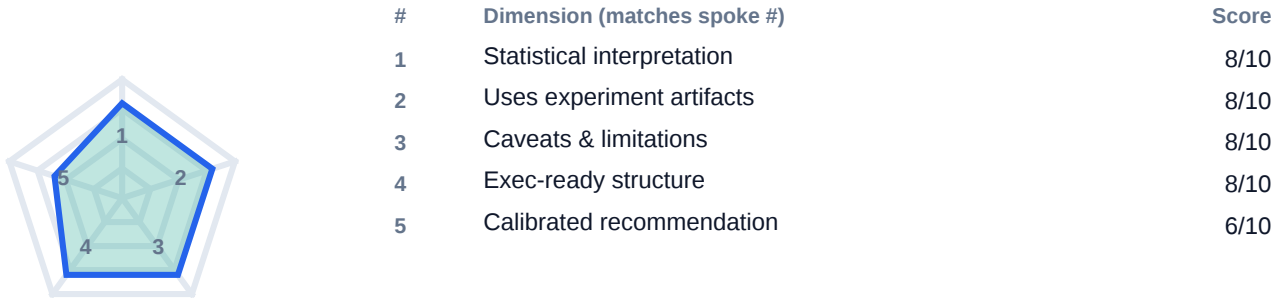
Report date: July 14, 2026 - 3:33 AM EST

Overall

7.6/10

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

CAPABILITY RUBRIC



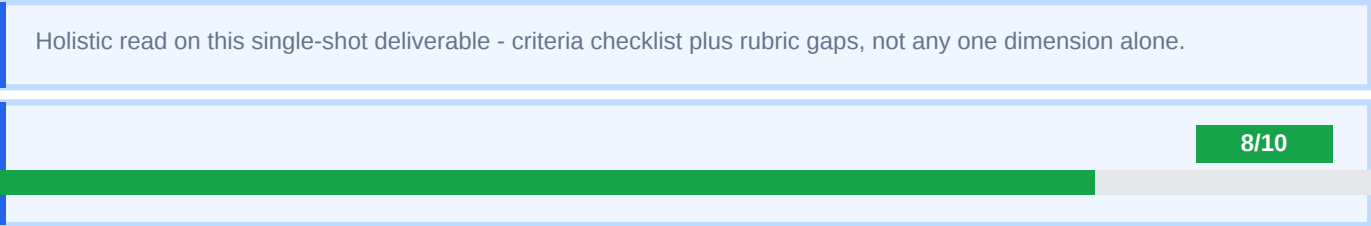
Spoke numbers run clockwise from the top; dimension names are listed in full in the table.

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: 3812 chars - Latency: 11s - 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: 11s

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

Engineering counterpart

7:33 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

7:33 AM - Responded in 4s

Executive Summary:

Our 2-week A/B test, "EXP-Checkout-CTA-884," compared the "blue_button" (control) to "green_button" (treatment) for the "purchase_conversion_rate" metric.

Key Result:

The green button variant showed a statistically significant 9.6% relative lift in purchase conversions (p-value: 0.041).

Confidence Interval (95%): The estimated treatment effect falls within [0.002%, 0.046%], indicating a likely true effect between 0.002% to 0.046% lift.

Caveats:

- Randomization unit was at the session level.
- A known inventory bug impacted treatment on mobile web during days 11-12, affecting approximately 4% of treatment sessions.

Recommendation:

Based on the results, ship the green_button variant, with the caveat of monitoring for post-launch inventory issues similar to the mobile web bug. If similar problems arise, consider iterating on the treatment while keeping the core visual change intact. Otherwise, hold for a longer period to validate the results and confirm stability.

Engineering counterpart

7:33 AM - Responded in 0s

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

Agent response

7:33 AM - Responded in 7s

Let's break down the thought process:

1. Interpret the p-value: The p-value of 0.041 indicates that, assuming the null hypothesis (no treatment effect) is true, we would expect to observe a test statistic at least as extreme as the one we did, by chance, only 4.1% of the time. Given the test's alpha level (typically 0.05), we reject the null hypothesis in favor of the alternative: the green button variant is likely to produce a higher purchase conversion rate.
2. Examine the confidence interval (CI): The 95% CI of [0.002%, 0.046%] provides a range of plausible true treatment effects. Since the entire interval is positive, it suggests that the green button likely increases conversions. The width of the interval indicates a moderate amount of uncertainty in our estimate.
3. Evaluate the treatment effect size: A 9.6% relative lift is a meaningful increase in purchase conversions. However, we should consider the context: is this a crucial metric for the business, and what is the magnitude of the potential revenue impact?
4. Assess the experimental design: We used a session-level randomization unit, which is suitable for this conversion metric. However, we should verify that sessions are an appropriate unit for this behavior and that no other factors unrelated to the button color could be driving the results.
5. Address potential biases and confounders: The known inventory bug on mobile web during days 11-12 is a notable issue. Although it only affected 4% of treatment sessions, this could still distort the results. We should monitor for similar post-launch issues and consider additional controls or adjustments if needed.
6. Formulate a recommendation: Given the statistical significance, moderate effect size, and the fact that the bug seems containable, I recommend shipping the green_button variant. However, we should closely monitor the launch for any unexpected effects or side issues. If inventory problems persist or surface elsewhere, we could iterate on the treatment to mitigate these impacts while preserving the core green button design.

The key assumptions made:

- The p-value and confidence interval calculations assume independence between observations and a known population variance (not explicitly stated in the experiment details).
- The treatment effect is assumed to be additive and not interact with other factors (e.g., user demographics, purchase amounts).
- The experimental design was sound and effectively allocated users to control and treatment groups.
- The inventory bug was the only potential confounder or bias in the experiment.

By acknowledging these assumptions and considering the caveats, we can make a data-driven decision while being mindful of potential limitations in the A/B test results.

Engineering counterpart

7:33 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.