



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: deepseek/deepseek-v4-flash

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

Duration: 33s

Actual run cost: \$0.000616

Projected execution cost: \$0.2033/mo (15/biz day × 22 days)

Report date: July 14, 2026 - 3:31 AM EST

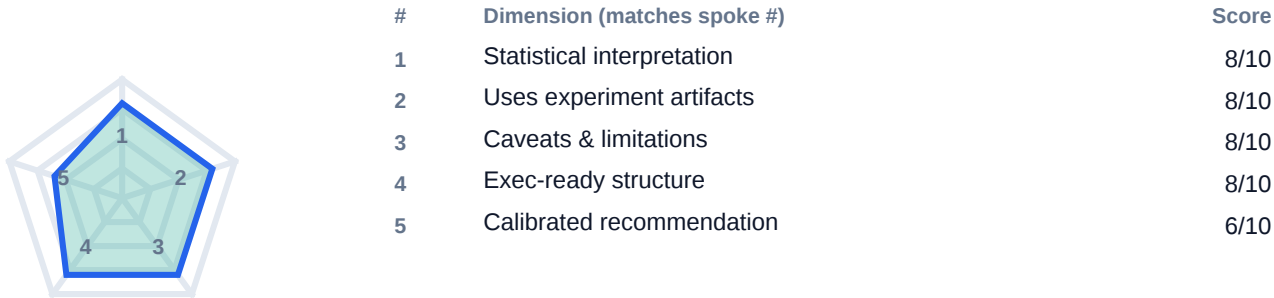
Overall

7.6/10

PASS - trust=high

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

CAPABILITY RUBRIC



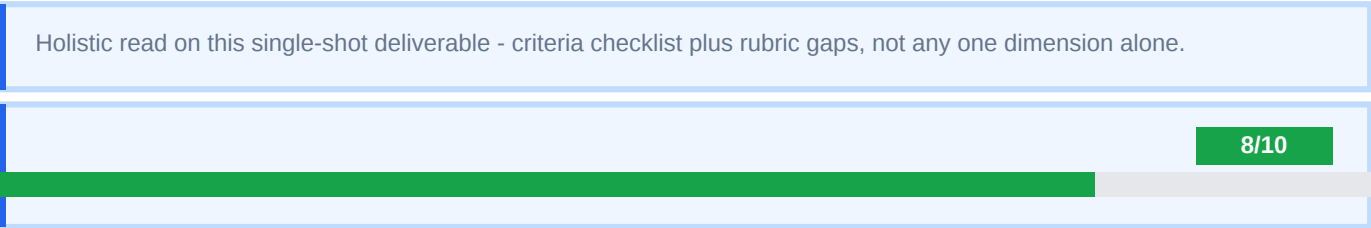
Spoke numbers run clockwise from the top; dimension names are listed in full in the table.

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: 3976 chars - Latency: 33s - 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: 33s

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 20s

Headline: The green button lifted purchase conversion by 9.6% relative (absolute +0.24 percentage points, from 2.50% to 2.74%).

Statistical interpretation: The result is marginally significant ($p = 0.041$; 95% CI for absolute lift [0.02pp, 0.46pp]). The confidence interval is narrow but touches near zero, indicating a small but detectable effect.

Caveat: An inventory bug on mobile web during days 11-12 affected an estimated 4% of treatment sessions. This could inflate or dilute the true effect, and the bug's exact impact on conversion is unknown.

Recommendation: Hold - The small effect size and the identified bug make the result unreliable. Investigate the bug's impact (e.g., segment by day/platform). If the treatment effect holds in the unaffected portion, consider shipping with monitoring; otherwise, re-run the experiment cleanly.

Engineering counterpart

7:31 AM - Responded in 0s

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

Agent response

7:31 AM - Responded in 13s

Here's the step-by-step reasoning behind the summary, followed by the assumptions I made.

Step-by-step reasoning

1. Read the headline metric

The experiment is about purchase conversion rate. Control = 2.50%, treatment = 2.74%. Relative lift = $(0.0274 - 0.0250) / 0.0250 = 9.6\%$. That's the headline number.

2. Check statistical significance

p-value = 0.041, which is below the conventional 0.05 threshold -> the difference is statistically significant at the 95% confidence level.

3. Examine the confidence interval

95% CI for the absolute lift is [0.0002, 0.0046] -> [0.02pp, 0.46pp]. The interval does not include zero, confirming significance. However, the lower bound is very close to zero, meaning the effect is small and could be near zero in the population.

4. Assess practical significance

A 0.24pp absolute lift on a 2.50% baseline is a 9.6% relative improvement, which sounds nice, but the absolute effect is tiny. For a checkout button change, this is a marginal gain. The narrow CI tells us the effect is precisely estimated but small.

5. Read the notes for caveats

There is a known inventory bug affecting mobile web days 11-12, estimated at 4% of treatment sessions. That contamination could bias the result. The bug could either inflate the conversion rate (if the bug made it easier to purchase) or deflate it (if it prevented purchases). Without knowing the direction, we cannot trust the observed lift.

6. Formulate the recommendation

Because the bug directly affects the treatment group's data, and the effect size is borderline, shipping now would be risky. The safest path is to hold and investigate: segment by platform and day, check if the lift holds in the unaffected portion. If it does, consider re-running cleanly or shipping with monitoring. If not, re-run.

Assumptions I made

- Randomization was otherwise sound - The note only mentions a bug, not other failures like sample ratio mismatch or leakage. I assume the randomization unit (session) and segmentation were correct aside from the bug.

- The bug only affected treatment - The note says "treatment on mobile web days 11-12", so I assume control was unaffected. This is reasonable because it's described as a treatment-specific issue.

- The bug impact is uncertain - I do not assume the bug inflated or deflated conversions. I treat it as a black-box contamination that makes the observed effect unreliable.

- The 4% estimate is accurate - The note says "estimated 4%", so I take it at face value. If the estimate were rough, the caveat would be even stronger.

- p-value and CI are correctly computed - I assume the statistical test accounted for the sample size and was appropriate for a binary metric (e.g., a z-test or chi-square). The notes don't mention any corrections, so I assume standard methods.

- The metric is well-defined - "purchase_conversion_rate" is assumed to be a standard per-session conversion. No mention of double-counting or cross-device issues.

- The decision should be risk-averse - For a

Engineering counterpart

7:31 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.