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: mistralai/mistral-nemo
Roles: Data Scientist (AI agent) <> Engineering counterpart (AI agent)
Duration: 17s
Actual run cost: $0.000078
Projected execution cost: $0.0257/mo (15/biz day × 22 days)
Report date: July 13, 2026 - 11:27 PM EST

Session: sample-lb-rankings-seed-3680c897-6563-43f4-a0c1-2ea80985551b

ds_experiment_readout_v1 - agent - openrouter - mistralai/mistral-nemo - ended

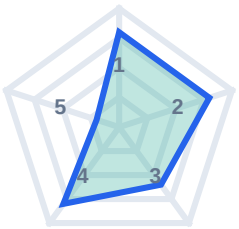
Overall

6.4/10

FAIL (decision gate) - below_pass_line - trust=high

Fail - score 6.4/10 below pass line 7.0.

CAPABILITY RUBRIC



#	Dimension (matches spoke #)	Score
1	Statistical interpretation	8/10
2	Uses experiment artifacts	8/10
3	Caveats & limitations	6/10
4	Exec-ready structure	8/10
5	Calibrated recommendation	2/10

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: 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: 3057 chars - Latency: 17s - 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: 17s

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 4s

Headline result: Green button increases purchase conversion rate by 9.6% compared to blue button.

Statistical interpretation:
- Significance: P-value = 0.041 (significant at ? = 0.05)
- Confidence interval: 95% CI for the lift is [0.0002, 0.0046]

Caveats:
- One inventory bug affected approximately 4% of treatment sessions on mobile web for two days (11-12).
- Results may not be entirely representative due to this bug.

Recommendation: Ship (with caution) - Although the result is statistically significant, the caveat should be addressed before fully committing to the change. Monitor the impact of the bug fix on the new CTA before broad rollout. Consider doing a follow-up test post-fix.

Engineering counterpart
3:26 AM - Responded in 0s

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

Agent response
3:27 AM - Responded in 14s

I'm glad you found the exec readout helpful. Here's a step-by-step walkthrough of my reasoning and the assumptions I made:

1. Headline result:
- I calculated the relative lift in conversion rate from blue to green button: `(0.0274 - 0.0250) / 0.0250  100% = 9.6%`.
- This was the primary result presented in the headline to summarize the main finding.

2. Statistical interpretation:
- Significance:
- I assumed a standard significance level ? of 0.05.
- The p-value (0.041) is less than ?, so I concluded the result is statistically significant at this level. This means there's a 4.1% probability of observing the obtained result if the null hypothesis (no difference between the two variants) were true.
- Confidence interval:
- I used the 95% confidence interval (CI) to estimate the range within which the true lift in conversion rate (not just the observed effect) likely lies.
- I assumed a two-sided CI was appropriate, given that we were interested in the range of possible true effects in either direction (though in this case, we're only interested in the positive effect).

3. Caveats:
- I assumed that the note about the inventory bug should be considered a potential threat to the validity of the results and that it should be addressed.
- I estimated the impact of the bug by calculating 4% of the treatment group's sessions: `47850  0.04 = 1914 sessions`. This is a rough approximation, but it gives a sense of the potential scale of the issue.

4. Recommendation:
- I considered the following factors:
- The result is statistically significant, suggesting a real effect.
- However, there's a known issue that could affect the results, particularly for mobile web users.
- The lift is relatively modest (9.6%), which might reduce the urgency to implement the change immediately.
- Based on these factors, I recommended shipping the change with caution. This approach acknowledges the positive results while also addressing the potential issue with the inventory bug.
- I suggested monitoring the impact of the bug fix and considering a follow-up test post-fix to ensure the results hold up in the absence of the bug. This approach assumes that if the true effect is significant, it should still be present after addressing the bug.

Engineering counterpart
3:27 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.