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: qwen/qwen-2.5-7b-instruct
Roles: Data Scientist (AI agent) <> Engineering counterpart (AI agent)
Duration: 7s
Actual run cost: $0.000440
Projected execution cost: $0.1452/mo (15/biz day × 22 days)
Report date: July 14, 2026 - 3:32 AM EST

Session: sample-lb-rankings-seed-337a4737-9dd5-43ba-bcca-014522c00f83

ds_experiment_readout_v1 - agent - openrouter - qwen/qwen-2.5-7b-instruct - ended

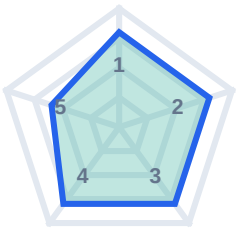
Overall

7.6/10

PASS - trust=high

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

CAPABILITY RUBRIC



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

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: Met

TECHNICAL JUDGEMENT

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

8/10

Telemetry

Run metrics

Mode: multi-turn - Agent msgs: 2 - Counterpart msgs: 3 - Output: 3107 chars - Latency: 7s - 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: 7s

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

Engineering counterpart
7:32 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
7:32 AM - Responded in 2s

Executive Summary:

- Headline Result: The green button variant led to a 9.6% increase in purchase conversion rate compared to the blue button variant.
- Statistical Interpretation: The result is statistically significant at the 5% level (p-value = 0.041). The 95% confidence interval for the relative lift is (0.0002%, 0.0046%), suggesting a small but significant improvement.
- Caveats: A known inventory bug affected approximately 4% of treatment sessions on mobile web during days 11-12, which may introduce some variability.
- Recommendation: Ship the green button variant, as the results are significant despite the noted bug. However, monitor the performance closely in the coming weeks to ensure consistent results and address any issues post-launch.

Engineering counterpart
7:32 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 5s

Step-by-Step Reasoning:

1. Headline Result:
- Reasoning: The green button variant (treatment) had a higher purchase conversion rate (0.0274) compared to the blue button variant (control) (0.0250), resulting in a relative lift of 9.6%.
- Assumptions: We assume that the conversion rates are accurate and that the experiment was conducted under normal conditions.

2. Statistical Interpretation:
- p-value: The p-value of 0.041 indicates that there is a 4.1% chance of observing the results if the null hypothesis (no difference between variants) is true.
- 95% Confidence Interval: The 95% confidence interval for the relative lift is [0.0002%, 0.0046%]. This means we are 95% confident that the true relative lift lies within this range.
- Significance: The p-value (0.041) is less than the significance level (0.05), indicating that the result is statistically significant. The confidence interval does not include zero, further supporting the significant improvement.

3. Caveats:
- Inventory Bug: A known inventory bug affected approximately 4% of treatment sessions on mobile web during days 11-12. This could introduce variability and potentially bias the results.
- Assumptions: We assume that the impact of the bug was consistent across sessions and did not significantly skew the results.

4. Recommendation:
- Ship the Variant: Given the statistically significant result, the green button variant should be rolled out.
- Monitoring: Close monitoring is recommended to ensure consistent performance and to address any issues that may arise post-launch.
- Assumptions: The recommendation is based on the statistical significance and the assumption that the impact of the bug is minimal and does not significantly affect the overall result.

Summary:
- Headline Result: The green button variant showed a 9.6% relative lift in purchase conversion rate.
- Statistical Interpretation: The result is statistically significant at the 5% level with a 95% confidence interval indicating a small but significant improvement.
- Caveats: There is a known inventory bug affecting treatment sessions, but its impact is assumed to be minimal.
- Recommendation: Ship the green button variant and monitor performance closely.

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.