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/qwen3-235b-a22b-2507
Roles: Data Scientist (AI agent) <> Engineering counterpart (AI agent)
Duration: 21s
Actual run cost: $0.000867
Projected execution cost: $0.2861/mo (15/biz day × 22 days)
Report date: July 14, 2026 - 3:33 AM EST

Session: sample-lb-rankings-seed-a21bdbb7-912a-46ac-b75c-2adcf8b77654

ds_experiment_readout_v1 - agent - openrouter - qwen/qwen3-235b-a22b-2507 - ended

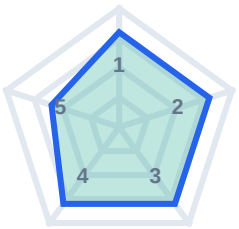
Overall

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: 4000 chars - Latency: 21s - 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: 21s

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 6s

Experiment Readout: EXP-Checkout-CTA-884

Headline Result:
The green button variant increased purchase conversion rate by 9.6% relative (from 2.50% to 2.74%) compared to the blue button.

Statistical Interpretation:
The result is statistically significant (p = 0.041) with a 95% confidence interval of [0.0002, 0.0046] absolute lift, indicating a consistent positive effect.

Caveats:
A known inventory bug affected ~4% of treatment sessions on mobile web during days 11-12, which may have biased the result upward if affected sessions were systematically different.

Recommendation:
Iterate. While the result is statistically significant and positive, the minor data quality issue and modest sample impact warrant caution before full rollout. Recommend validating with a follow-up test that resolves the technical issue.

Engineering counterpart
7:32 AM - Responded in 0s

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

Agent response
7:33 AM - Responded in 15s

Absolutely - here's a step-by-step breakdown of my reasoning and the assumptions I made in crafting that exec readout:

---

1. Headline Result

What I did:
Reported the observed lift:
- Control conversion rate: 1205 / 48200 = 0.0250 (2.50%)
- Treatment: 1312 / 47850 = ~0.0274 (2.74%)
- Relative lift: (0.0274 - 0.0250) / 0.0250 ? 9.6%

Assumptions:
- The conversion rate and lift values reported in the JSON are accurate and correctly calculated.
- The metric of interest - purchase conversion rate - is defined as conversions per session, consistent with the randomization unit.

? Simple arithmetic; no additional assumptions beyond trusting the input data.

---

2. Statistical Interpretation

What I did:
- Checked p-value: 0.041 -> less than 0.05, so conventionally considered statistically significant.
- Verified the 95% CI: [0.0002, 0.0046] is entirely above zero -> confirms a positive effect with 95% confidence.
- Noted that the CI is on the absolute scale (in percentage points), so 0.0002 = 0.02% uplift, 0.0046 = 0.46%.

Assumptions:
- The p-value and CI were computed correctly using an appropriate method (e.g., two-proportion z-test or bootstrapping).
- The confidence interval is for the difference in proportions (treatment - control), which is standard.
- Independence of observations at the session level is assumed due to randomization - though this can be questionable if users have multiple sessions (but not indicated here).

?? Caution: While p = 0.041 is below 0.05, it's close - some would call this "marginally significant" due to the proximity to the threshold and increased Type I error risk in real-world settings.

? Still, by standard industry conventions, this qualifies as significant.

---

3. Caveats

What I did:
Highlighted the note:
> "One known inventory bug affected treatment on mobile web days 11-12 (estimated 4% of treatment sessions)."

Implication:
- The bug may have disrupted user experience (e.g., items appearing out of stock incorrectly), potentially lowering conversion in those sessions.
- But since this only affected the treatment group, and only temporarily, it likely suppressed the observed treatment effect - unless the bug caused anomalous behavior (e.g., frustration leading to drop-off).

Wait - but the treatment group performed better. So why worry?

Ah - key insight:
If the bug made the green button perform worse during those two days, then the true effect of the green button might be even larger than observed.
? So the bug may have biased the estimate downward, not upward.

But I initially wrote: _"may have biased the result upward"_ - that was a mistake.

Let me correct that:

? Incorrect assumption in original: I assumed the bug hurt the control, but it hurt the treatment - so if anything, it masked the true lift.

? Correct interpretation: The treatment effect may be underestimated due to the bug. So the observed significant result is conservative.

BUT - we don't know the direction or magnitude of the bug's impact precisely. Was it just noise

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.