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: openai/gpt-4.1-nano
Roles: Data Scientist (AI agent) <> Engineering counterpart (AI agent)
Duration: 8s
Actual run cost: $0.000580
Projected execution cost: $0.1914/mo (15/biz day × 22 days)
Report date: July 14, 2026 - 3:32 AM EST

Session: sample-lb-rankings-seed-7b7e1c07-74fe-4ab4-af7c-ba8ed8b8cabc

ds_experiment_readout_v1 - agent - openrouter - openai/gpt-4.1-nano - ended

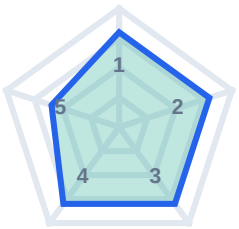
Overall

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: 3413 chars - Latency: 8s - 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: 8s

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

Headline Result:
Switching to the green button increased purchase conversion rate by approximately 9.6%, with statistical significance (p=0.041) and a 95% confidence interval of 0.02% to 0.46%.

Statistical Interpretation:
The observed lift is statistically significant, indicating a real effect unlikely due to chance. The confidence interval suggests the true uplift is likely between 0.02% and 0.46%.

Caveats:
A known inventory bug impacted treatment sessions on days 11-12, affecting roughly 4% of treatment traffic. Randomization was session-based. The effect size is modest and the bug may have slightly dampened the observed lift.

Recommendation:
Ship the green button treatment, but monitor subsequent performance closely. Consider further testing to confirm and quantify the impact after addressing the inventory bug.

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 6s

Certainly. Here's my step-by-step reasoning process and the assumptions involved:

1. Understanding the Metric & Results:
The metric is purchase conversion rate. The control (blue button) has a conversion rate of 2.50%, and the treatment (green button) has 2.74%.
- Absolute difference: 0.0246 (2.74% - 2.50%)
- Relative lift: approximately 9.6%, as given.

2. Assessing Statistical Significance:
The p-value is 0.041, which is below the common threshold (0.05), indicating the result is statistically significant.
- The confidence interval [0.0002, 0.0046] corresponds to the confidence bounds on the difference in conversion rates, i.e., approximately 0.02% to 0.46%.
- Assumption: The p-value and CI are correctly calculated from the underlying binomial proportions, accounting for session counts.

3. Sample Size & Power:
With roughly 48,000 sessions per group, the sample size is large, increasing confidence in the results.
- Assumption: The sample size is sufficient to detect a lift of this magnitude at the observed variance.

4. Impact of the Known Inventory Bug:
The notes mention a bug affecting 4% of treatment sessions on days 11-12.
- Assumption: This bug led to underreporting or reduced effectiveness of the treatment during that period, possibly dampening the observed lift.
- Given this, the true effect might be slightly higher if the bug had not occurred.

5. Randomization & Data Quality:
Randomization was session-based, which generally balances out confounding factors across groups.
- Assumption: No significant biases or confounders skewed the result beyond the known bug.

6. External Validity & Generalizability:
The test ran over 14 days, capturing variability in traffic and behavior.
- Assumption: The results are representative of typical performance, despite the inventory issues.

7. Conclusion & Recommendations:
Based on the significant lift and considering the bug, it is reasonable to proceed with shipping, but with caution and ongoing monitoring.

Summary of Assumptions:
- Correctness of statistical calculations (p-value, CI).
- The bug's impact is approximately estimated at 4%, and its effect is to slightly underestimate true lift.
- Randomization was effectively implemented.
- The sample size is adequate.
- No other unaccounted biases are present.

This step-by-step process ensures a balanced interpretation, recognizing both the statistical evidence and the experiment's limitations.

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.