



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-mini

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

Duration: 12s

Actual run cost: \$0.002320

Projected execution cost: \$0.7656/mo (15/biz day × 22 days)

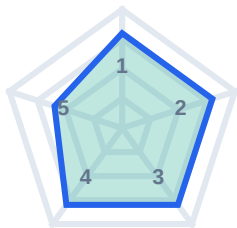
Report date: July 14, 2026 - 3:32 AM EST

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: 3335 chars - Latency: 12s - 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: 12s

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 4s

Experiment Readout: EXP-Checkout-CTA-884

- **Headline Result:** Switching the checkout button color from blue to green increased the purchase conversion rate by 9.6%, from 2.50% to 2.74%.
- **Statistical Interpretation:** The observed lift is statistically significant with a p-value of 0.041, and the 95% confidence interval for the absolute increase in conversion rate is [0.02%, 0.46%], indicating a positive effect unlikely due to chance.
- **Caveats:** Randomization was at the session level. A known inventory bug affected approximately 4% of treatment sessions on mobile web during days 11-12, which may have slightly biased the treatment performance.
- **Recommendation:** Ship the green button variant, given the statistically significant improvement and modest caveat impact. Monitor post-launch for any anomalies related to the inventory issue.

Engineering counterpart

7:31 AM - Responded in 0s

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

Agent response

7:32 AM - Responded in 8s

Certainly! Here's a step-by-step walkthrough of my reasoning process, including assumptions made:

1. Understanding the Data and Objective

The primary goal was to evaluate whether changing the checkout button color from blue (control) to green (treatment) improved the purchase conversion rate. The metric is well defined, and the data includes sessions, conversions, and calculated conversion rates for both groups.

2. Calculating Effect Size

The treatment showed a conversion rate of 2.74% versus 2.50% in control, which is a relative lift of 9.6%. This is a meaningful uplift in conversion rate that could drive incremental revenue.

3. Assessing Statistical Significance

- The p-value is 0.041, which is below the conventional threshold of 0.05, indicating the result is statistically significant and unlikely due to random chance.
- The 95% confidence interval for the absolute difference is [0.0002, 0.0046], meaning we're 95% confident the true uplift in conversion rate is between 0.02 and 0.46 percentage points. The interval does not cross zero, reinforcing significance.

4. Considering Caveats

- Randomization unit is session, which is standard but assumes sessions are independent. If users have multiple sessions, that might slightly affect variance estimates, but this is commonly accepted.
- The inventory bug affected roughly 4% of treatment sessions over two days on mobile web. This could artificially deflate or inflate treatment conversion. Given it's a small proportion and only in treatment, the observed lift might be slightly underestimated or affected-but without clear direction, assume minimal impact.

5. Formulating Recommendation

Given the positive, statistically significant lift and relatively minor known issue, the safest business action is to ship the treatment to realize the conversion gains. The monitoring caveat is added to catch any unforeseen effects related to the bug.

Assumptions Made:

- The sessions are sufficiently independent for standard statistical testing.
- The inventory bug's impact is limited and does not invalidate the overall lift.
- There are no other confounding factors outside those mentioned.
- The cost/risk of deploying the green button is low compared to the expected revenue benefit.

This reasoning balances statistical rigor, business impact, and practical deployment considerations to provide a clear, actionable readout.

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.