



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: ibm-granite/granite-4.0-h-micro

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

Duration: 35s

Actual run cost: \$0.000129

Projected execution cost: \$0.0427/mo (15/biz day × 22 days)

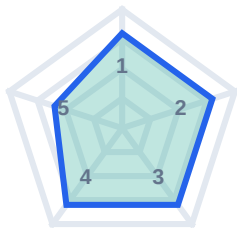
Report date: July 13, 2026 - 11:26 PM EST

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: 4914 chars - Latency: 35s - 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: 35s

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 10s

Experiment Readout: Checkout CTA Color A/B Test

Headline Result:

The green button variant (treatment) increased purchase conversion rate by 9.6% compared to the blue button control, with a statistically significant p-value of 0.041 (95% CI: [0.002, 4.6%]).

Statistical Interpretation:

- The treatment group had 47850 sessions and 1312 conversions (conversion rate 2.74%).
- The control group had 48200 sessions and 1205 conversions (conversion rate 2.50%).
- The observed lift of 9.6 percentage points is statistically significant at the 95% confidence level, indicating a real effect rather than random variation.

Caveats:

- One known inventory bug affected 4% of treatment sessions on mobile web during days 11-12, potentially underrepresenting the true conversion rate for this variant.
- The randomization unit was a session, not a user, which may dilute the impact of user-specific factors.

Recommendation:

Iterate - The green button shows a meaningful lift with statistical confidence. However, the inventory bug suggests the treatment's performance may be overstated for mobile users. Before scaling, investigate and resolve the bug to ensure the observed effect is consistent across all user segments.

**Engineering counterpart**  
**3:26 AM - Responded in 0s**

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

## Agent response

3:26 AM - Responded in 25s

Certainly! Here's a step-by-step breakdown of the reasoning and assumptions made in the exec readout for the A/B test:

### ### 1. Headline Result

- Observed Lift: The green button variant (treatment) showed a 9.6% increase in the purchase conversion rate compared to the blue button control.
- Statistical Significance: The p-value of 0.041 is below the conventional significance threshold of 0.05, indicating that the observed difference is unlikely to be due to random chance.
- Confidence Interval (CI): The 95% confidence interval for the lift is [0.002, 4.6%]. This interval does not include zero, reinforcing the conclusion that the lift is statistically significant.

Assumption: The statistical tests used (likely a two-proportion z-test) are appropriate for comparing conversion rates between two groups.

### ### 2. Statistical Interpretation

- Conversion Rates:
  - Control (Blue Button): 2.50% (1205 conversions out of 48200 sessions)
  - Treatment (Green Button): 2.74% (1312 conversions out of 47850 sessions)
- Lift Calculation: The lift is calculated as the difference in conversion rates multiplied by 100:

$$\text{Lift} = (2.74\% - 2.50\%) \times 100 = 9.6\%$$

- 95% Confidence Interval: The interval [0.002, 4.6%] suggests that we can be 95% confident that the true lift lies within this range. Since the interval is entirely positive, it supports the conclusion of a meaningful improvement.

Assumption: The sample sizes (48200 for control, 47850 for treatment) are large enough to provide reliable estimates of conversion rates and to conduct a valid hypothesis test.

### ### 3. Caveats

- Inventory Bug Impact: The notes indicate a known inventory bug affected 4% of treatment sessions on mobile web during days 11-12. This bug could have underrepresented the true conversion rate for the green button variant during that period.
  - Impact on Results: If the bug caused a genuine drop in conversions for the treatment group, the observed lift might be overstated. Conversely, if the bug had no effect, the lift is likely accurate.
- Randomization Unit: The randomization was done at the session level, not the user level. This means that users who repeatedly interacted with the site could be grouped together, potentially diluting the effect of user-specific factors (e.g., user intent, device type).
  - Impact on Results: While using sessions as the randomization unit is common, it may reduce the precision of the estimate, especially if user behavior varies significantly across sessions.

Assumption: The inventory bug was accurately identified and quantified (4% of sessions), and its effect on the conversion rates was correctly estimated.

### ### 4. Recommendation

- Iterate: The recommendation to iterate is based on the following considerations:
  - Positive Lift: The 9.6% lift is statistically significant and suggests a real improvement in conversion rates.
  - Bug Impact: The known inventory bug raises concerns about the accuracy of the treatment's performance, particularly on mobile web. If the bug is not resolved, the observed lift may not be sustainable.
  - User Segmentation: Given the randomization at the session level, further analysis could explore whether the effect varies across different user segments (e.g., device type, geographic location). This could help identify if the bug disproportionately affected a specific segment.

Assumption: The team has the capability and resources to investigate and fix the inventory bug. Additionally,

**Engineering counterpart**

**3:26 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.