



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: google/gemini-2.5-flash

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

Duration: 6s

Actual run cost: \$0.002650

Projected execution cost: \$0.8745/mo (15/biz day × 22 days)

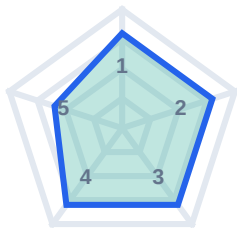
Report date: July 13, 2026 - 11:25 PM EST

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: 3688 chars - Latency: 6s - 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: 6s

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

Engineering counterpart

3:25 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

3:25 AM - Responded in 2s

## Experiment Readout: EXP-Checkout-CTA-884

Headline Result: The "Green Button" CTA increased purchase conversion rate by 9.6% compared to the "Blue Button."

Statistical Interpretation: This result is statistically significant ( $p=0.041$ ), meaning we are 95% confident the green button drives a real uplift. The 95% confidence interval for the absolute difference is [0.02%, 0.46%].

Caveats: A known inventory bug on mobile web days 11-12 affected an estimated 4% of treatment sessions. This could slightly understate the true impact of the green button by a small margin.

Recommendation: Ship. The positive and statistically significant uplift outweighs the minor, likely understating, impact of the bug. We can continue to monitor performance post-launch.

## Engineering counterpart

3:25 AM - Responded in 0s

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

## Agent response

3:25 AM - Responded in 4s

Okay, let's break down the reasoning process for generating that exec readout.

### ### Step-by-Step Reasoning

1. Understand the Goal: The request is to provide an "exec readout" - this means concise, clear, actionable, and focused on key business metrics and decisions. It needs to cover a headline, statistical interpretation, caveats, and a recommendation.

2. Deconstruct the JSON Input:

`experiment\_id`: "EXP-Checkout-CTA-884" - Good for identification.

`metric`: "purchase\_conversion\_rate" - This is the primary business metric we're optimizing for. Crucial.

`runtime\_days`: 14 - Sufficient for many tests, good context.

`control`: Details on the baseline.

    `variant`: "blue\_button"

    `sessions`: 48200

    `conversions`: 1205

    `conversion\_rate`: 0.0250 (2.50%) - This is the baseline.

`treatment`: Details on the new variant.

    `variant`: "green\_button"

    `sessions`: 47850

    `conversions`: 1312

    `conversion\_rate`: 0.0274 (2.74%) - This is the new performance.

`lift\_relative\_pct`: 9.6 - This is the immediate business impact.

`p\_value`: 0.041 - Critical for statistical significance.

    `confidence\_interval\_95`: [0.0002, 0.0046] - Also critical for statistical significance and understanding the range of effect. Note these are absolute differences, not relative.

    `notes`: "Randomization unit = session. One known inventory bug affected treatment on mobile web days 11-12 (estimated 4% of treatment sessions)." - EXTREMELY IMPORTANT for caveats and recommendation.

3. Draft the Headline Result:

Focus on the `lift\_relative\_pct` and the `metric`.

Identify the winner: "Green Button" vs "Blue Button."

Initial thought: "Green button performed better." - Too weak.

Better: "Green Button increased purchase conversion rate by 9.6%." - Good, specific, impactful.

4. Formulate Statistical Interpretation:

Significance: Look at `p\_value`. Is it below the standard alpha (0.05)? Yes,  $0.041 < 0.05$ . So, "statistically significant."

Confidence: State the confidence level (95%).

Meaning of Significance: Explain what "statistically significant" means in plain language (e.g., "we are 95% confident the green button drives a real uplift").

Confidence Interval: Present the `confidence\_interval\_95`. It's given as an absolute difference. Convert to percentage points for clarity. "[0.0002, 0.0046]" becomes "[0.02%, 0.46%]".

Self-correction: Ensure the CI is presented appropriately. It represents the range of the true difference between the treatment and control conversion rates. Since both ends are positive, it reinforces the positive lift. If it included zero or went negative, it would change the interpretation significantly.

## Engineering counterpart

3:25 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.