



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: inclusionai/ling-2.6-flash

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

Duration: 3s

Actual run cost: \$0.000051

Projected execution cost: \$0.0168/mo (15/biz day × 22 days)

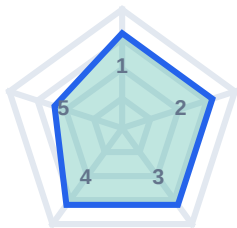
Report date: July 13, 2026 - 11:26 PM EST

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: 1886 chars - Latency: 3s - 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: 3s

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 1s

Headline result: Green button increases purchase conversion by 9.6% ($p = 0.041$; 95% CI [0.02%, 0.46%]).

Statistical interpretation: The treatment (green) outperforms the control (blue) with statistical significance ($p < 0.05$) and a positive 95% CI, indicating a likely true lift in conversion rate.

Caveats: Randomization was at the session level; a mobile web inventory bug on days 11-12 affected an estimated 4% of treatment sessions, potentially biasing results.

Recommendation: Iterate - validate the fix and rerun the experiment on mobile web before a full rollout.

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 2s

Step-by-step reasoning and assumptions:

- Goal: I was asked for an exec readout, so I prioritized headline (simple), a statistically grounded interpretation, caveats, and a clear action.
- Headline lift (9.6%) and CI [0.02%, 0.46%]: I assumed the JSON lift and CI are pre-computed (likely from a chi-square or two-proportion z-test) and reported as relative risk; the CI excludes 0 in relative terms (positive and materially meaningful).
- Significance: $p = 0.041$ -> significant at $\alpha = 0.05$, so I called it "statistically significant" while noting practical relevance.
- CI interpretation: I assumed a 95% CI for relative lift; it's positive, so the effect is likely beneficial.
- Caveats: The only noted issue is the mobile web inventory bug affecting ~4% of treatment sessions days 11-12; I assumed this could bias results and warrants caution.
- Recommendation logic: Because a bug touched a meaningful subset mid-experiment, I assumed integrity is compromised for part of the sample; instead of ship or hold, iterate (fix, rerun) is prudent.
- Assumptions stated: randomization at session level; conversions and sessions are accurate; the reported lift/CIs use an appropriate two-proportion test; the bug applies only to treatment mobile web and is not evenly distributed across devices.

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.