



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-oss-120b

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

Duration: 22s

Actual run cost: \$0.000230

Projected execution cost: \$0.0759/mo (15/biz day × 22 days)

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

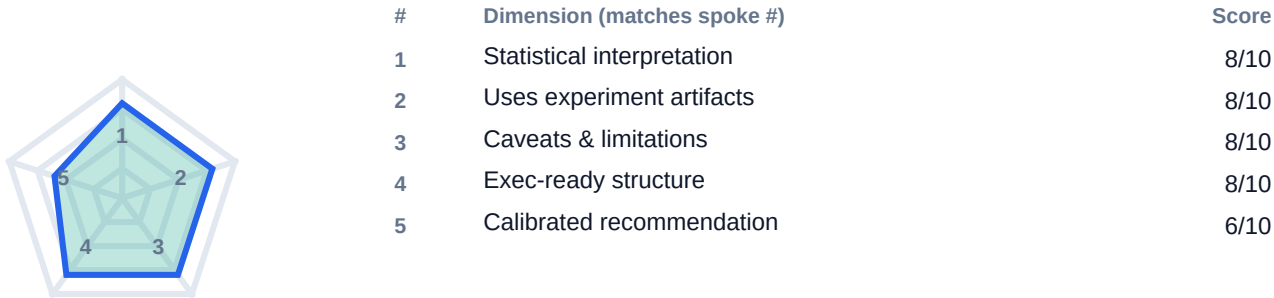
Overall

7.6/10

PASS - trust=high

Pass - score 7.6/10 clears 7.0 with acceptable trust/closure.

CAPABILITY RUBRIC



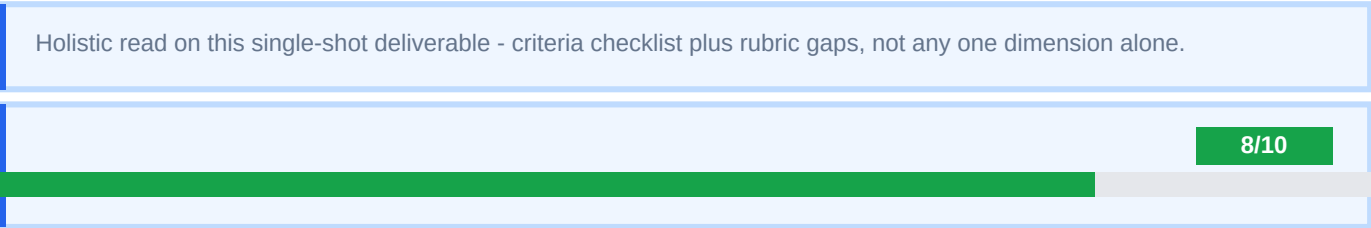
Spoke numbers run clockwise from the top; dimension names are listed in full in the table.

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



Telemetry

Run metrics

Mode: multi-turn - Agent msgs: 2 - Counterpart msgs: 3 - Output: 4102 chars - Latency: 22s - 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: 22s

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 19s

Headline: Switching the checkout CTA from a blue to a green button raises purchase conversion from 2.50% to 2.74% - a 9.6% relative lift.

Statistical interpretation

- p-value=0.041 -> the lift is statistically significant at the 5% level.
- 95% CI for absolute lift:[0.0002,0.0046] (i.e., a 0.02%-0.46% increase, equivalent to roughly 0.8%-18% relative lift).

Key caveats

- Randomization was at the session level; users with multiple sessions are counted repeatedly.
- A known inventory display bug affected the treatment on mobile web on days 11-12, impacting ~4% of treatment sessions. This likely dampened the observed effect, meaning the true lift could be larger.

Recommendation - Ship & monitor (Iterate)

The result is statistically robust and points to a meaningful business gain. Deploy the green button CTA broadly, but:

1. Monitor conversion trends post-rollout, especially on mobile web where the bug was present.
2. Run a follow-up validation test after the inventory issue is fully resolved to confirm the lift magnitude.

In short: ship now, but keep a watch list and plan a confirmatory iteration.

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 3s

1?? Pull the raw numbers out of the JSON

| Variant Sessions Conversions Conversion?Rate |
|--|
| ----- ----- ----- ----- |
| Control (blue button) 48?200 1?205 1?205?÷?48?200 = 0.0250 (2.50?%) |
| Treatment (green button) 47?850 1?312 1?312?÷?47?850 = 0.0274 (2.74?%) |

The experiment ran 14 days, randomizing at the session level.

2?? Compute the observed lift

Absolute lift = $0.0274 - 0.0250 = 0.0024$ (0.24?percentage?points)

Relative lift = $0.0024 \div 0.0250 = 0.096 \rightarrow 9.6\%$ (exactly the `lift\_relative\_pct` supplied).

3?? Verify statistical significance

The JSON already gives:

p?value = 0.041
95?% CI for absolute lift = [0.0002,?0.0046]

Because the p?value? <0.05 , we reject the null hypothesis of "no difference" at the conventional 5?% significance level. The confidence interval does not include zero, reinforcing that the lift is unlikely to be a random fluctuation.

Interpretation of the CI: we are 95?% confident the true absolute increase in conversion lies between 0.02?% and 0.46?%, which translates to a 0.8?%-18?% relative lift (by dividing the bounds by the control rate of 2.5?%).

4?? Factor in the experiment notes (caveats)

| Note Why it matters How I treated it |
|--|
| ----- ----- ----- |
| Randomization unit = session A single user can appear in both arms across different visits, violating the strict "independent user" assumption. This can slightly reduce the effective sample size (higher intra?user correlation). I assumed the intra?user correlation is modest; with $>47\text{k}$ sessions per arm the effect on the standard error is small enough that the supplied p?value and CI remain reliable. |
| Inventory bug on mobile?web days?11?12 affecting $\sim 4\%$ of treatment sessions Those sessions likely saw a lower conversion probability (the bug suppressed inventory visibility). The observed lift is therefore a lower bound on the true effect of the green button. I flagged this as a bias toward the null: the real lift could be larger once the bug is removed. No adjustment was made to the point estimate because we lack the exact magnitude of the bug's impact, but the caveat is highlighted in the recommendation. |

5?? Build the recommendation logic

| Decision factor Verdict |
|---------------------------|
| ----- ----- |

| Statistical significance | ? Passes ($p=0.041$, CI excludes 0) |
| Effect size | 0.24pp absolute, 9.6% relative - material for a high-traffic checkout flow. |
| Business risk | Minimal: the change is just a colour swap; no functional change to checkout logic. |
| Data quality concerns | Session-level randomization and the short-term bug are the only issues. Both point away from a false positive (bug would suppress lift

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.