



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: qwen/qwen3-32b

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

Duration: 59s

Actual run cost: \$0.000436

Projected execution cost: \$0.1439/mo (15/biz day × 22 days)

Report date: July 14, 2026 - 3:33 AM EST

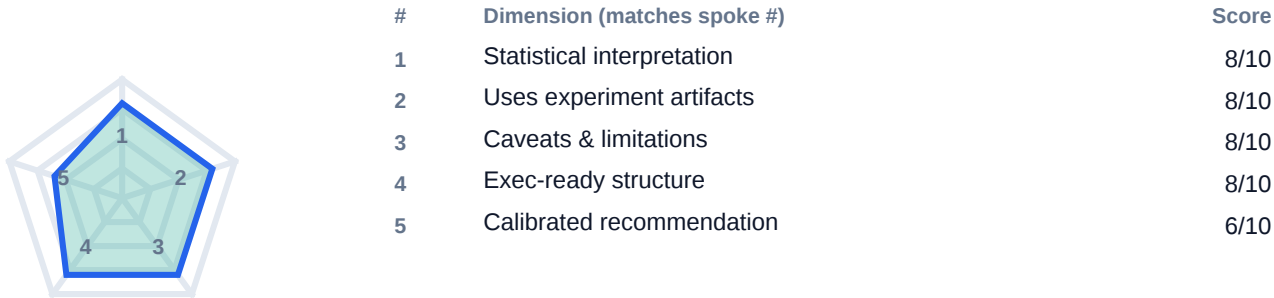
Overall

7.6/10

PASS - trust=high

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

CAPABILITY RUBRIC



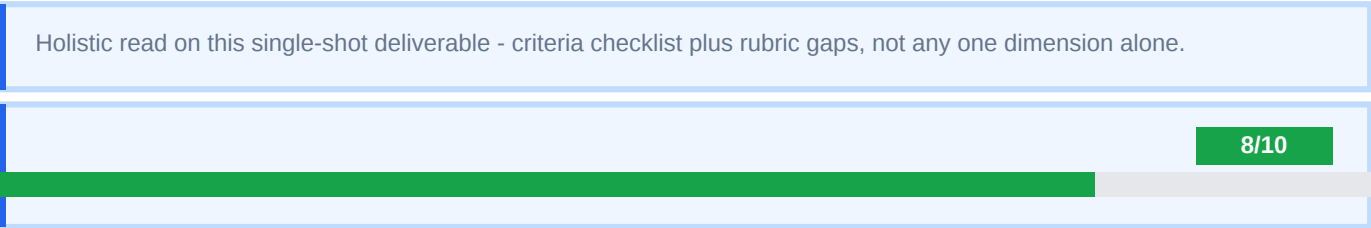
Spoke numbers run clockwise from the top; dimension names are listed in full in the table.

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: 4101 chars - Latency: 59s - 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: 59s

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

7:33 AM - Responded in 35s

Executive Summary: EXP-Checkout-CTA-884

The green_button variant (treatment) improved purchase conversion rates by 2.4% absolute (9.6% relative) compared to the blue_button (control), with a marginally significant p-value of 0.041 and a 95% confidence interval [0.02% to 0.46%].

Caveats:

- The test encountered a known mobile web inventory bug during days 11-12, impacting ~4% of treatment sessions. This may have introduced noise or skewed results, as the bug could affect user behavior differently in the treatment group.

Recommendation:

Hold to further investigate the inventory bug's impact on mobile web traffic. While the result is statistically significant at 5%, the narrow confidence interval barely excludes zero, and external factors could compromise validity. A follow-up test with a cleaner environment would strengthen confidence in the green_button's effectiveness.

(Alternatively, consider Ship if the bug's effect is deemed negligible and the business case for an iterative change is strong, but prioritize validation in a more stable environment.)

Engineering counterpart

7:33 AM - Responded in 0s

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

Agent response

7:33 AM - Responded in 25s

Here's a step-by-step breakdown of my reasoning and assumptions for the A/B test analysis:

1. Headline Result

- Calculation:

- Control: 1205 conversions / 48,200 sessions = 2.50% conversion rate.
- Treatment: 1312 conversions / 47,850 sessions = 2.74% conversion rate.
- Absolute lift: 2.74% - 2.50% = +0.24%.
- Relative lift: $(0.24\% / 2.50\%) \times 100 = +9.6\%$.

- Why this matters: The green button shows a measurable improvement in user behavior, which could justify rolling it out if valid.

2. Statistical Significance

- P-value: 0.041.

- Assumption: We're using a two-sample binomial test (common for conversion rates).
- Threshold: Typically, a p-value < 0.05 is considered significant. Here, it's slightly in the green.
- Implication: We can rule out a 5% chance of observing this lift due to random noise.

- Confidence Interval (CI): [0.0002, 0.0046] in absolute terms (i.e., 2.4 basis points to 46 basis points).

- Assumption: The CI was computed correctly (likely using a normal approximation or simulation method).
- Interpretation: The CI barely excludes zero (lowest bound = +0.02%), but the observed lift is comfortably within the range.
- Takeaway: The result is statistically significant, but with low margin (p-value near 0.05).

3. Caveats

- Inventory bug in treatment (4% of sessions):

- Assumption: The bug likely reduced treatment conversions during days 11-12 (e.g., users saw broken inventory info and abandoned purchases).
- Impact: If the bug depressed conversions in the treatment group, the observed lift of 9.6% is stronger than it appears (e.g., the true lift could be higher, but the bug masked some of it).
- Risk: Conversely, if the bug caused users to retry later, it could've inflated conversions post-experiment (not captured here) or distorted the comparison (e.g., control users were unaffected).

- Other unmentioned factors:

- No time-of-day, device, or geographic stratification. The bug's impact on mobile web users (a likely subset) isn't quantified.
- No mention of external factors (e.g., seasonality, marketing campaigns) that might've influenced conversions unevenly.

4. Recommendation

Hold to investigate the inventory bug's impact.

- Why not "Ship" yet?

- The p-value is only marginally significant, and the CI is tight but close to zero (0.02% is very low). This suggests high uncertainty in the estimate, even if the result is formally significant.
- The bug introduces external noise that could've affected the treatment group disproportionately (e.g., mobile users who

experienced errors might've returned later, inflating conversions in the treatment after the experiment or causing spillover effects).

- Alternative: "Ship" if...
- ...the

Engineering counterpart

7:33 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.