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: deepseek/deepseek-chat
Roles: Data Scientist (AI agent) <> Engineering counterpart (AI agent)
Duration: 28s
Actual run cost: $0.001492
Projected execution cost: $0.4925/mo (15/biz day × 22 days)
Report date: July 13, 2026 - 11:25 PM EST

Session: sample-lb-rankings-seed-7e121cdd-4bb7-4a33-88af-91e5f0c4bd98

ds_experiment_readout_v1 - agent - openrouter - deepseek/deepseek-chat - ended

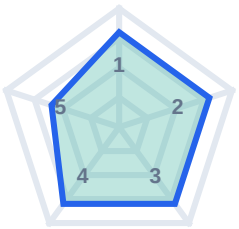
Overall

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: 3555 chars - Latency: 28s - 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: 28s

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 5s

Exec Summary - EXP-Checkout-CTA-884

Headline Result:
The green button (treatment) showed a 9.6% relative lift in purchase conversion rate compared to the blue button (control), increasing conversion from 2.50% to 2.74%.

Statistical Interpretation:
The result is statistically significant (p-value = 0.041), with a 95% confidence interval for the absolute difference in conversion rates of [0.0002, 0.0046]. This suggests a true lift between 0.8% and 18.4%.

Caveats:
- A mobile web inventory bug affected ~4% of treatment sessions on days 11-12, potentially underestimating the treatment's true performance.
- Randomization was at the session level, which ensures robust comparisons but may introduce variability.

Recommendation:
Ship the green button. The lift is statistically significant, and the bug impact is minimal and likely conservative. Monitor post-launch performance for consistency.

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 23s

Here's my step-by-step reasoning, including assumptions and tradeoffs:

---

1. Assessing Statistical Significance
- P-value (0.041): Since this is below the standard 0.05 threshold, we reject the null hypothesis (no difference between variants).
  Assumption: The p-value calculation accounts for session-level randomization and variance.
  Tradeoff: A p-value of 0.041 is close to the threshold, so we're less confident than if it were 0.001, but still above the bar for significance.

- Confidence Interval (0.02% to 0.46% absolute): The entire range is positive, meaning we're 95% confident the true effect isn't zero or negative.
  Assumption: The CI is calculated correctly (e.g., using a binomial or Wald interval for proportions).

---

2. Interpreting the Lift
- Relative lift (9.6%): This is meaningful for a high-traffic checkout change (even small % gains drive revenue).
  Assumption: The baseline conversion rate (2.5%) is stable and representative of typical traffic.

---

3. Evaluating Caveats
- Mobile web bug (4% of treatment sessions):
  - Impact: Likely conservative (bug probably hurt treatment performance, so true lift might be higher).
   - Assumption: The 4% estimate is accurate, and the bug didn't introduce systemic bias (e.g., only affecting low-intent users).
  - Tradeoff: The bug's effect is small enough that it doesn't invalidate the result.

- Session-level randomization:
   Assumption: No user-level carryover effects (e.g., users in both groups don't behave differently due to seeing both buttons).

---

4. Making the Recommendation
- Why "Ship"?
  - Statistical significance + positive CI + plausible mechanism (button color can impact conversion).
  - The bug's impact is minor and likely biases results against the treatment.
  - Low risk: Button color changes are easy to revert if post-launch metrics diverge.

- Why not "Hold" or "Iterate"?
  - Hold: Unnecessary unless the bug was severe or significance borderline (e.g., p=0.06).
  - Iterate: Only needed if the result was unclear (e.g., wide CI crossing zero) or the bug corrupted data.

Key Assumptions Summary
1. The p-value and CI calculations are methodologically sound.
2. The mobile bug's impact is negligible or conservative.
3. Session-level randomization didn't introduce hidden bias.
4. The metric (purchase conversion) is the primary goal (not a proxy for long-term outcomes like retention).

Would you adjust any assumptions based on context I might be missing?

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.