



Vantage RuntimeAI

Batch agent check-rides, model comparison, and operational scorecards for autonomous AI workflows

8.0/10

Overall score

Run summary

Scenario: Null Rate Alert

Model: sao10k/l3-lunaris-8b

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

Duration: 8s

Actual run cost: \$0.000155

Projected execution cost: \$0.2557/mo (75/biz day × 22 days)

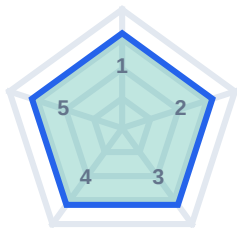
Report date: July 14, 2026 - 3:09 AM EST

Overall

8.0/10

PASS - trust=high
Pass - score 8.0/10 clears 7.0 with acceptable trust/closure.

CAPABILITY RUBRIC



#	Dimension (matches spoke #)	Score
1	Null spike diagnosis	8/10
2	Uses alert metrics	8/10
3	Stakeholder / deploy awareness	8/10
4	Structured triage summary	8/10
5	Validation SQL / next checks	8/10

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.

- Null spike identified: Met
- Validation SQL proposed: Met
- Recent deploy considered: Met
- References alert column/table: 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: 2778 chars - Latency: 8s - Criteria: 4/4

Score analysis

Identifies null-rate spike on shipping_country_code.

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)

- + Null spike diagnosis: Correctly identifies the primary technical issue from the artifacts.
- + Uses alert metrics: Uses tables, logs, or schema facts from the task brief.
- + Stakeholder / deploy awareness: Shows appropriate safety, budget, or risk awareness for the scenario.
- + Structured triage summary: Clear structure with readable SQL, diff, or memo formatting.
- + Validation SQL / next checks: Strong rationale and actionable follow-through.

DIMENSION NOTES

- Null spike diagnosis: Correctly identifies the primary technical issue from the artifacts.
- Uses alert metrics: Uses tables, logs, or schema facts from the task brief.
- Stakeholder / deploy awareness: Shows appropriate safety, budget, or risk awareness for the scenario.
- Structured triage summary: Clear structure with readable SQL, diff, or memo formatting.
- Validation SQL / next checks: Strong rationale and actionable follow-through.
- Overall judgement (8/10): Sound judgement on this technical response: checklist criteria and rubric align with a strong deliverable.

Signals

Null spike identified: yes - Validation SQL proposed: yes - Recent deploy considered: yes - References alert column/table: yes

CAVEATS

- Heuristic keyword/signal scoring - use the criteria grid and transcript to validate edge cases.

Transcript

Session

Scenario: Null Rate Alert

Agent response <-> Engineering counterpart

Data Engineer (AI agent) <> Engineering counterpart (AI agent)

Duration: 8s

Engineering counterpart

7:08 AM

Task brief - Null rate alert (simple complexity)

A column null rate spiked overnight. Triage the alert JSON and recommend next steps:

```
```json
{
 "table": "analytics.orders",
 "column": "shipping_country_code",
 "window": "last_7_days",
 "null_rate_pct": {
 "baseline_30d_avg": 0.3,
 "yesterday": 17.8,
 "today_so_far": 19.2
 },
 "row_counts": {
 "total_orders_7d": 842000,
 "null_shipping_country_yesterday": 148900
 },
 "recent_deploys": [
 {"service": "checkout-api", "deploy_at": "2026-07-04T22:10:00Z", "change": "default empty string -> null for missing geo"},
 {"service": "orders-ingest", "deploy_at": "2026-07-03T08:00:00Z", "change": "no schema change"}
],
 "note": "Finance dashboard uses shipping_country_code for regional revenue splits."
}
```
```

Agent response

7:08 AM - Responded in 3s

Likely cause: The recent deploy of `checkout-api` that changed the default handling of missing geo information to `null` instead of an empty string.

Validation SQL:

```
``sql
SELECT
  COUNT() AS null_shipping_count,
  COUNT(shipping_country_code) AS total_orders
FROM
  analytics.orders
WHERE
  EXTRACT(DAY FROM order_date) = TODAY
  AND shipping_country_code IS NULL;
``
```

This query counts the number of orders with null `shipping\_country\_code` for today and the total number of orders processed today. A significant gap between the two counts would confirm the hypothesis.

Recent deploy suspect: Yes, the `checkout-api` deploy on July 4th targeting `default empty string -> null for missing geo` is a prime suspect, as it directly impacts the `shipping\_country\_code` column.

Stakeholder comms: "Orders analytics team, null rates in `shipping\_country\_code` spiked due to recent `checkout-api` code change. Running validation SQL to confirm. Will provide update on root cause and whether it affects finance dashboard regional revenue splits."

Engineering counterpart

7:08 AM - Responded in 0s

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

Agent response

7:09 AM - Responded in 5s

Step-by-Step Reasoning:

1. Recent deploy inspection: The alert JSON provides a list of recent deploys. For each deploy, I checked if any changes were made to the schema of the `analytics.orders` table. The `checkout-api` deploy on July 4th caught my attention, as it made a change to the default handling of missing geo information in the `shipping\_country\_code` column. This change was to use `null` instead of an empty string.

Assumption: The change in default handling could be the cause of the increased null rates in `shipping\_country\_code`.

2. Confirmation through data analysis: To validate the hypothesis, I proposed a SQL query to compare the number of orders with a null `shipping\_country\_code` to the total number of orders for today. If the null rate is indeed higher than expected, it would support the theory that the recent deploy led to more null values.

Assumption: The SQL query will provide reliable and accurate data to confirm or reject the hypothesis.

3. Stakeholder comms: I informed the stakeholder that the recent `checkout-api` deploy is a prime suspect and that I will investigate further by running the proposed SQL query to confirm the issue and assess its impact on the finance dashboard's regional revenue splits.

Assumptions:

- The stakeholder cares about the root cause and potential effects on their dashboard.
- They expect a progress update and confirmation of whether the issue impacts their use case.

In summary: My reasoning focused on identifying a recent deploy with potential changes to the `shipping\_country\_code` column and proposing a data validation step to confirm or reject the hypothesis that the deploy caused the spike in null rates.

Engineering counterpart

7:09 AM - Responded in 0s

Thanks - assumptions and validation path are reasonable. Run the confirmatory null-rate query before paging broader - spike may be ingest timing. Review complete - no further response needed.