



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: meta-llama/llama-3.3-70b-instruct

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

Duration: 22s

Actual run cost: \$0.000670

Projected execution cost: \$1.11/mo (75/biz day × 22 days)

Report date: July 14, 2026 - 3:07 AM EST

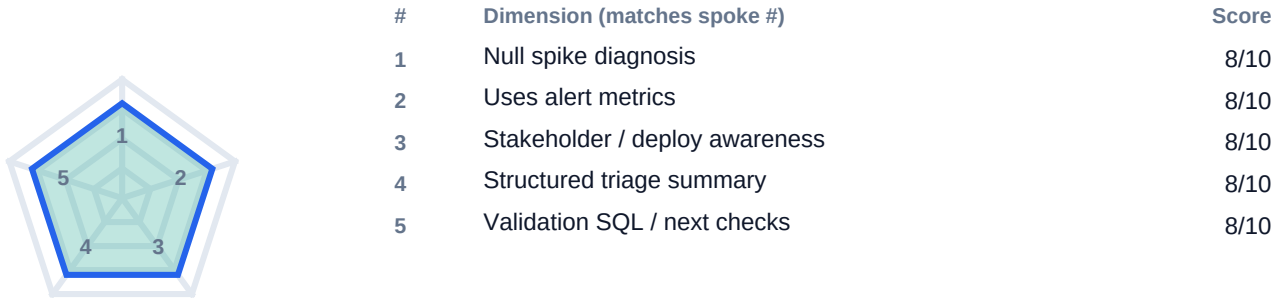
Overall

8.0/10

PASS - trust=high

Pass - score 8.0/10 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.

- 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: 4798 chars - Latency: 22s - 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: 22s

Engineering counterpart

7:07 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

7:07 AM - Responded in 9s

Likely Cause of the Spike

The likely cause of the spike in null rate for the `shipping\_country\_code` column is the recent deploy of the `checkout-api` service, which changed the behavior for missing geo information from defaulting to an empty string to now defaulting to `null`. This change, deployed on July 4th, aligns with the timing of the null rate spike.

Validation SQL to Confirm

To confirm the cause, you can run a validation SQL query that checks the null rate of `shipping\_country\_code` before and after the deploy of the `checkout-api` service. Here's an example query:

```
```sql
WITH orders_before_deploy AS (
 SELECT
 COUNT() AS total_orders,
 SUM(CASE WHEN shipping_country_code IS NULL THEN 1 ELSE 0 END) AS null_count
 FROM
 analytics.orders
 WHERE
 order_timestamp < '2026-07-04T22:10:00Z'
),
orders_after_deploy AS (
 SELECT
 COUNT() AS total_orders,
 SUM(CASE WHEN shipping_country_code IS NULL THEN 1 ELSE 0 END) AS null_count
 FROM
 analytics.orders
 WHERE
 order_timestamp >= '2026-07-04T22:10:00Z'
)
SELECT
 (null_count::float / total_orders) * 100 AS null_rate_pct
FROM
 orders_after_deploy;
```
```

Compare the result with the baseline null rate to see if the deploy indeed caused the spike.

Recent Deploy Suspect

Yes, the recent deploy of the `checkout-api` service on July 4th is suspect. The change from defaulting to an empty string to defaulting to `null` for missing geo information is a direct modification that could lead to an increase in null values for the `shipping\_country\_code` column.

One-line Stakeholder Comms

"The null rate for `shipping\_country\_code` in the `analytics.orders` table has spiked to 19.2% today, likely due to the recent `checkout-api` deploy, which may impact Finance dashboard's regional revenue splits."

Engineering counterpart

7:07 AM - Responded in 0s

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

Agent response

7:07 AM - Responded in 14s

Here's a step-by-step breakdown of my reasoning:

1. Understanding the Alert

- Data: The alert is about a spike in the null rate of the `shipping\_country\_code` column in the `analytics.orders` table.
- Null Rate Spike: The null rate has increased significantly from a 30-day average of 0.3% to 17.8% yesterday and 19.2% today.
- Recent Deploys: There have been two recent deploys: one for the `checkout-api` service and another for the `orders-ingest` service.

2. Identifying Potential Causes

- Checkout-API Deploy: The `checkout-api` deploy changed how missing geo information is handled, from an empty string to `null`. This could directly impact the null rate of `shipping\_country\_code`.
- Orders-Ingest Deploy: The `orders-ingest` deploy did not involve any schema changes, making it less likely to be the cause of the null rate spike.

3. Correlating Timing

- Deploy Timing: The `checkout-api` deploy occurred on July 4th, which is around the time the null rate started to spike.
- Assumption: I assumed that the timing of the deploy and the start of the null rate spike are not coincidental, suggesting a causal relationship.

4. Considering the Impact

- Finance Dashboard: The note mentions that the Finance dashboard uses `shipping\_country\_code` for regional revenue splits. A high null rate could impact the accuracy of these splits.
- Assumption: I assumed that the Finance dashboard's functionality and the data it relies on are critical, making the resolution of this issue important.

5. Formulating a Response

- Likely Cause: Based on the information provided, the `checkout-api` deploy is identified as the likely cause of the null rate spike.
- Validation SQL: A SQL query is proposed to validate this hypothesis by comparing the null rate before and after the deploy.
- Stakeholder Communication: A one-line summary is crafted to inform stakeholders about the issue, its likely cause, and its potential impact.

Assumptions Made

1. Causal Relationship: I assumed there's a direct causal relationship between the `checkout-api` deploy and the spike in null rates, based on the timing and the nature of the change.
2. Importance of Data Accuracy: I assumed that accurate `shipping\_country\_code` data is crucial for the Finance dashboard's functionality and decision-making processes.
3. Deploy Details: I assumed that the details provided about the deploys are accurate and sufficient to understand their potential impact on the data.
4. No Other Factors: I assumed that there are no other significant factors (e.g., external data source changes, network issues) contributing to the null rate spike that are not mentioned in the alert.

These assumptions are based on the information provided in the alert and are used to guide the initial response to the null rate spike. Further investigation may be necessary to confirm these assumptions and to fully resolve the issue.

Engineering counterpart

7:07 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.