



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-4-maverick

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

Duration: 17s

Actual run cost: \$0.001160

Projected execution cost: \$1.91/mo (75/biz day × 22 days)

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

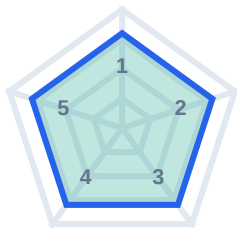
Overall

8.0/10

PASS - trust=high

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

CAPABILITY RUBRIC



#	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: 4942 chars - Latency: 17s - 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: 17s

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 5s

Triage Report: Null Rate Alert for `shipping\_country\_code` in `analytics.orders`

1. Likely Cause of the Spike: The recent deploy to `checkout-api` on `2026-07-04T22:10:00Z` changed the default for missing geo data from an empty string to null. This change likely caused the spike in null rates for `shipping\_country\_code` as previously non-null empty strings are now being recorded as null.
2. Validation SQL to Confirm: To validate the cause, you can run the following SQL query to compare the count of rows with empty strings or null values in `shipping\_country\_code` before and after the deploy:

```
```sql
SELECT
 COUNT(CASE WHEN shipping_country_code = "" THEN 1 END) AS empty_string_count,
 COUNT(CASE WHEN shipping_country_code IS NULL THEN 1 END) AS null_count,
 created_at::date AS date
FROM
 analytics.orders
WHERE
 created_at >= '2026-07-03' AND created_at < '2026-07-05'
GROUP BY
 created_at::date
ORDER BY
 date;
```
```

This query will help identify if there's a significant shift from empty strings to null values around the time of the deploy.

3. Whether a Recent Deploy is Suspect: Yes, the deploy to `checkout-api` on `2026-07-04T22:10:00Z` is suspect. The change from defaulting to an empty string to defaulting to null for missing geo data directly correlates with the observed spike in null rates.
4. One-line Stakeholder Comms: "The spike in null rates for `shipping\_country\_code` is likely due to a recent change in `checkout-api` where missing geo data now defaults to null instead of an empty string, affecting regional revenue splits in the Finance dashboard."

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 12s

Step-by-Step Reasoning for Triage Report

1. Understanding the Alert: The alert indicates a significant spike in the null rate for `shipping\_country\_code` in the `analytics.orders` table. The null rate has increased from a 30-day average of 0.3% to 17.8% yesterday and 19.2% so far today.

2. Identifying Key Information:

- The table and column affected: `analytics.orders` and `shipping\_country\_code`.
- The time window for the alert: `last\_7\_days`.
- Null rates: Baseline average (0.3%), yesterday (17.8%), and today so far (19.2%).
- Row counts: Total orders in the last 7 days (842,000) and null `shipping\_country\_code` counts for yesterday (148,900).
- Recent deploys: Two deploys are listed, one to `checkout-api` and one to `orders-ingest`.

3. Analyzing Recent Deploys:

- The deploy to `orders-ingest` on `2026-07-03T08:00:00Z` had "no schema change." This makes it less likely to be the direct cause of the null rate spike since schema changes are more likely to affect data ingestion or processing directly.
- The deploy to `checkout-api` on `2026-07-04T22:10:00Z` changed the default for missing geo data from an empty string to null. This is a significant change that could directly impact the null rate of `shipping\_country\_code` if the data ingested after this change is processed differently.

4. Correlating the Deploy with the Null Rate Spike:

- The null rate spike started after the `checkout-api` deploy. The deploy occurred on `2026-07-04T22:10:00Z`, which is likely after the data for `2026-07-04` was ingested or processed, suggesting that the data for `2026-07-05` (yesterday, considering the alert time) would be the first to fully reflect this change.
- The change from empty string to null for missing geo data would directly increase the null count, as previously non-null (but empty) values are now counted as null.

5. Assumptions Made:

- The `created\_at` timestamp or a similar field exists in `analytics.orders` to track when orders were created or ingested. This assumption is used in the validation SQL query.
- The data for `shipping\_country\_code` is populated or affected by the `checkout-api`. This is a reasonable assumption given that `checkout-api` is related to the processing or ingestion of order data.
- The `recent\_deploys` list is comprehensive for the services that could impact `shipping\_country\_code` in `analytics.orders`. It's assumed that there are no other significant changes or deploys not listed that could have caused the spike.

6. Formulating the Validation SQL:

- The SQL query is designed to check the count of empty strings and null values before and after the deploy. This helps confirm if the change in `checkout-api` is indeed the cause of the spike by showing a shift from empty strings to null values.

7. Stakeholder Communication:

- The one-line summary is designed to be clear and concise, directly linking the cause (the `checkout-api` deploy) to the effect (the spike in null rates) and mentioning the impact (regional revenue splits in the Finance dashboard).

By breaking down the information provided in the alert and correlating it with the recent deploys, we identified the likely cause of the null rate spike and formulated a plan to validate this hypothesis.

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.