



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: mistralai/mistral-nemo

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

Duration: 30s

Actual run cost: \$0.000078

Projected execution cost: \$0.1287/mo (75/biz day × 22 days)

Report date: July 13, 2026 - 10:54 PM 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: 4093 chars - Latency: 30s - 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: 30s

Engineering counterpart

2:54 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

2:54 AM - Responded in 12s

(1) Likely cause of the spike: The recent deploy on the "checkout-api" service on 2026-07-04T22:10:00Z, where the default for missing geographic data was changed from an empty string to null, is likely the cause of the spike in null rates for the `shipping\_country\_code` column.

(2) Validation SQL to confirm: To confirm this, we can check if the spike in nulls happened around the same time as the deploy. Here's an SQL query that can help validate this:

```
``sql
SELECT
  DATE(timestamp) AS date,
  COUNT() AS total_orders,
  SUM(CASE WHEN shipping_country_code IS NULL THEN 1 ELSE 0 END) AS null_country_orders
FROM
  analytics.orders
WHERE
  timestamp >= DATE_SUB(CURDATE(), INTERVAL 14 DAY)
GROUP BY
  date
ORDER BY
  date;
``
```

(3) Recent deploy suspect: Yes, the recent deploy on the "checkout-api" service is suspect. The change introduced by this deploy directly affects the `shipping\_country\_code` column, and the timing of the null rate increase aligns with the deploy.

(4) One-line stakeholder comms: "...The recent 'checkout-api' deploy on 2026-07-04T22:10:00Z, changing default missing geo data to null, seems to be causing the overnight spike in null rates for `shipping\_country\_code`. Let's validate and address."

Engineering counterpart

2:54 AM - Responded in 0s

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

Agent response

2:54 AM - Responded in 18s

You're welcome! I'd be happy to walk you through my reasoning step by step and explain the assumptions I made.

1. Correlation between deploy and null rate spike:

- Assumption: The null rate spike is likely caused by a recent change in the data processing logic or schema.
- Reasoning: In data engineering, changes in null rates often correlate with changes in how data is ingested, processed, or stored. In this case, the most recent deploy on the "checkout-api" service (2026-07-04T22:10:00Z) could have introduced such a change.
- Evidence: The note in the JSON states that the Finance dashboard uses `shipping\_country\_code` for regional revenue splits, indicating that null values in this column can impact business operations.

2. Identifying the suspect deploy:

- Assumption: The deploy that changed the default for missing geographic data from an empty string to null is suspect.
- Reasoning: This change directly affects the `shipping\_country\_code` column by introducing null values where previously empty strings were present.
- Evidence: The "checkout-api" deploy occurred just before the null rate spike, and its description ("default empty string -> null for missing geo") directly implicates the `shipping\_country\_code` column.

3. Confirming the deploy's impact:

- Assumption: To strengthen the evidence, we should validate that the null rate spike indeed occurred around the same time as the suspect deploy.
- Reasoning: If the null rate spike coincides with the deploy's timestamp, it provides further confirmation that the deploy is the cause of the issue.
- Action: Propose an SQL query to plot the daily total orders and null country orders over the past two weeks to visualize this correlation.

4. Communicating the findings:

- Assumption: Stakeholders will benefit from a clear, concise, and actionable summary of the findings.
- Reasoning: Communicating the suspected cause, the need for validation, and the recommended next steps will help stakeholders understand the issue and collaborate on resolving it.
- One-line stakeholder comms: "...The recent 'checkout-api' deploy on 2026-07-04T22:10:00Z, changing default missing geo data to null, seems to be causing the overnight spike in null rates for `shipping\_country\_code`. Let's validate and address." This statement summarizes the suspected cause, refers to the relevant deploy, and encourages further investigation and action.

Throughout this reasoning process, I assumed that the provided JSON data is accurate and up-to-date, and that the described deploys and changes reflect the actual state of the systems. Additionally, I presumed that the spike in null rates is indeed an issue that needs to be addressed, as indicated by the alert being raised.

Engineering counterpart

2:54 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.