



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: qwen/qwen-2.5-7b-instruct

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

Duration: 8s

Actual run cost: \$0.000440

Projected execution cost: \$0.7260/mo (75/biz day × 22 days)

Report date: July 14, 2026 - 3:08 AM EST

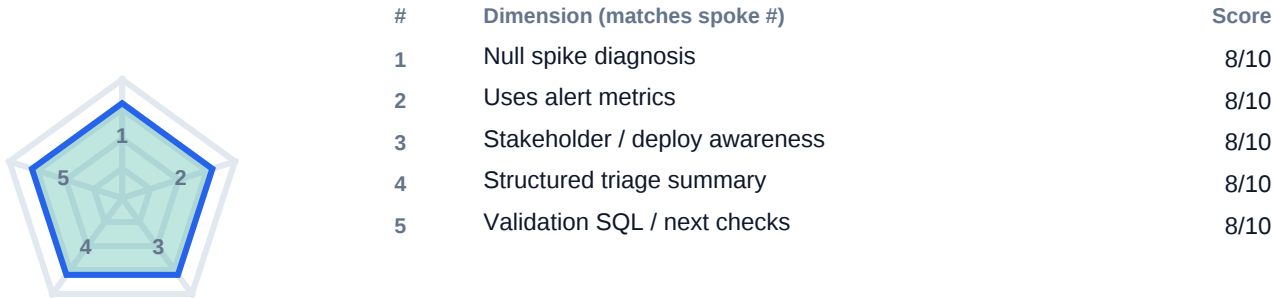
Overall

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: 3833 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

Triage Summary

1. Likely Cause of the Spike:

The spike in null rates is likely due to the change made in the `orders-ingest` service on July 3, 2026, where the default empty string for missing geolocation data was changed to `null`. This change would cause previously non-null but empty strings to become `null`, thus increasing the null rates.

2. Validation SQL to Confirm:

To verify this, run the following SQL to check the change in null counts before and after the deployment:

```
```sql
SELECT
 COUNT() AS total_orders,
 SUM(CASE WHEN shipping_country_code = " " THEN 1 ELSE 0 END) AS null_orders,
 SUM(CASE WHEN shipping_country_code IS NULL THEN 1 ELSE 0 END) AS null_shipping_country
FROM analytics.orders
WHERE order_date >= '2026-07-02' AND order_date < '2026-07-05';
```
```

3. Whether a Recent Deploy is Suspect:

Yes, the recent deploy in the `orders-ingest` service is likely responsible for the spike. The change introduced a shift from empty strings to `null` for missing geolocation data.

4. One-Line Stakeholder Communication:

"We've identified the recent change in the `orders-ingest` service as the likely cause of the null rate spike. Further investigation is underway to validate and correct the issue."

Engineering counterpart

7:08 AM - Responded in 0s

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

Agent response

7:08 AM - Responded in 5s

Certainly! Let's walk through the reasoning step by step and outline the assumptions made.

Step-by-Step Reasoning

1. Identify the Spike in Null Rates:

- The `null\_rate\_pct` shows a significant increase from 17.8% yesterday to 19.2% today.
- The baseline average null rate over the last 30 days is 0.3%.

2. Review the Recent Deploys:

- The `orders-ingest` service was deployed on July 3, 2026, with a change from using an empty string (`""`) to `null` for missing geolocation data.
- The `checkout-api` service was deployed on July 4, 2026, but no schema changes were made.

3. Assumption:

- The recent change in the `orders-ingest` service is likely causing the spike because it involves altering the handling of missing geolocation data.

4. Verification:

- To confirm, we can run a SQL query to check the number of orders with `null` shipping country codes around the deployment date.

5. SQL Query:

- The SQL query checks the number of orders with `null` shipping country codes around the deployment date:

```
```sql
SELECT
 COUNT() AS total_orders,
 SUM(CASE WHEN shipping_country_code = "" THEN 1 ELSE 0 END) AS null_orders,
 SUM(CASE WHEN shipping_country_code IS NULL THEN 1 ELSE 0 END) AS null_shipping_country
FROM analytics.orders
WHERE order_date >= '2026-07-02' AND order_date < '2026-07-05';
```
```

- This query will help us confirm if there was a significant increase in null values around the deployment date.

Assumptions Made

1. Assumption 1: The change from empty strings (`""`) to `null` in the `orders-ingest` service directly affects the `shipping\_country\_code` column.

- This assumption is based on the provided change notes and the recent date of the deployment.

2. Assumption 2: The `shipping\_country\_code` column values are checked by comparing `""` (empty string) and `NULL`.

- This assumption helps in verifying the exact nature of the change and its impact.

3. Assumption 3: The spike in null rates is recent and not due to other factors.

- This assumption is based on the recent deployment and the significant change in behavior.

Conclusion

The recent change in the `orders-ingest` service seems to be the likely cause of the null rate spike. Running the SQL query will help validate this hypothesis.

One-Line Stakeholder Communication

"We've identified the recent change in the `orders-ingest` service as the likely cause of the null rate spike. Further

investigation is underway to validate and correct the issue."

Engineering counterpart**7:08 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.