



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/qwen3-32b

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

Duration: 19s

Actual run cost: \$0.000436

Projected execution cost: \$0.7194/mo (75/biz day × 22 days)

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

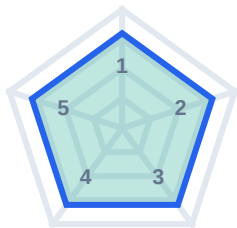
Overall

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: 2357 chars - Latency: 19s - 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: 19s

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 9s

1. Likely cause of the spike

The deploy to `checkout-api` (July 4) changed the default value for missing geography data from an empty string (`""`) to `NULL`. Previously, missing values were stored as non-null empty strings; now they are stored as null, inflating the null rate by ~17.5%. This likely broke assumptions in systems expecting empty strings instead of `NULL`.

2. Validation SQL

```
```sql
```

```
-- Check historical empty string vs NULL distribution
```

```
SELECT
```

```
 COUNT() AS total,
```

```
 SUM(CASE WHEN shipping_country_code = "" THEN 1 ELSE 0 END) AS empty_count,
```

```
 SUM(CASE WHEN shipping_country_code IS NULL THEN 1 ELSE 0 END) AS null_count
```

```
FROM analytics.orders
```

```
WHERE order_date BETWEEN '2026-06-25' AND '2026-07-04' -- 1 day pre-deploy
```

```
UNION ALL
```

```
SELECT
```

```
 COUNT(),
```

```
 SUM(CASE WHEN shipping_country_code = "" THEN 1 ELSE 0 END),
```

```
 SUM(CASE WHEN shipping_country_code IS NULL THEN 1 ELSE 0 END)
```

```
FROM analytics.orders
```

```
WHERE order_date BETWEEN '2026-07-05' AND '2026-07-06'; -- 1 day post-deploy
```

```
```
```

3. Suspect deploy

Yes, the `checkout-api` deploy on July 4 is highly suspect. The change from empty string to `NULL` directly explains the null rate jump. The `orders-ingest` deploy (July 3) had no schema change and is unlikely to be responsible.

4. Stakeholder comms

"Shipping country null rate spiked due to a `checkout-api` change defaulting missing values to NULL (previously empty strings); finance splits may have shifted-validate data integrity and consider rolling back the default logic."

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 10s

Here's my step-by-step reasoning and assumptions:

1. Identifying the Null Rate Spike

Observation: Null rate jumped from 0.3% (baseline) to 17.8% in one day.

Assumption: This is not a data ingestion issue or a one-off anomaly. The spike is consistent and large, suggesting a systematic schema or logic change.

2. Correlating with Recent Deploys

Key clue: The `checkout-api` deploy on July 4 changed the default value of missing geography entries from:

- `` (empty string) -> `NULL`

Assumption: This change directly affects the `shipping\_country\_code` column, which stores these values. If the column previously accepted empty strings (non-null), changing defaults to `NULL` would inflated null counts overnight.

Calculation:

- 17.8% null rate on

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.