



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: google/gemma-4-31b-it

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

Duration: 21s

Actual run cost: \$0.000538

Projected execution cost: \$0.8877/mo (75/biz day × 22 days)

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

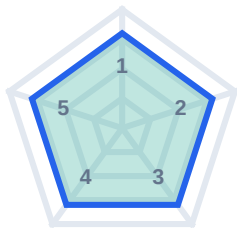
Overall

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: 3633 chars - Latency: 21s - 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: 21s

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 6s

1. Likely Cause

The spike was caused by the `checkout-api` deployment on 2026-07-04. The change explicitly shifted the handling of missing geographic data from empty strings (`""`) to `NULL`. Since the null rate jumped from 0.3% to ~18% immediately following this deploy, the "missing geo" cases are now being recorded as nulls.

2. Validation SQL

```
```sql
```

```
SELECT
```

```
 DATE(order_timestamp) as date,
```

```
 COUNT() as total_rows,
```

```
 COUNT() FILTER (WHERE shipping_country_code IS NULL) as null_count,
```

```
 (COUNT() FILTER (WHERE shipping_country_code IS NULL) 100.0 / COUNT()) as null_rate
```

```
FROM analytics.orders
```

```
WHERE order_timestamp >= '2026-07-03'
```

```
GROUP BY 1
```

```
ORDER BY 1;
```

```
```
```

3. Recent Deploy Suspect

Yes. The `checkout-api` deploy at `2026-07-04T22:10:00Z` is the direct cause.

4. Stakeholder Comms

"We are investigating a spike in null shipping country codes caused by a recent API change; regional revenue splits in Finance dashboards may be under-reported until a fix is deployed."

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 15s

To arrive at that conclusion, I followed a standard data engineering triage pattern: Correlation $\rightarrow$ Causation $\rightarrow$ Impact.

Here is the step-by-step breakdown:

1. Quantitative Analysis (The "What")

First, I looked at the magnitude of the shift.

Baseline: 0.3% (negligible, likely natural noise).

Current: ~18-19% (a massive, systemic jump).

Scale: ~149k rows in a single day.

This tells me it isn't a "glitch" or a few bad records; it is a change in how the system is behaving at scale.

2. Temporal Correlation (The "When")

I mapped the spike to the timeline provided in the `recent\_deploys` list.

The `orders-ingest` deploy happened on July 3rd, but the spike is noted as "yesterday/today" relative to the alert.

The `checkout-api` deploy happened on July 4th at 22:10Z.

Since the null rate jumped "overnight" and stayed high ("today so far"), the timing aligns perfectly with the `checkout-api` release.

3. Qualitative Matching (The "Why")

I looked for a "smoking gun" in the deploy notes.

The `checkout-api` change note explicitly says: "default empty string $\rightarrow$ null for missing geo."

The affected column is `shipping\_country\_code` (a "geo" field).

Conclusion: The code was intentionally changed to produce `NULL`s where it previously produced empty strings. The "spike" isn't necessarily a bug in the sense of broken code, but a change in data contract that the downstream analytics table wasn't expecting.

4. Impact Assessment (The "So What")

Finally, I checked the `note` regarding the Finance dashboard.

Because Finance uses this column for regional revenue splits, any `NULL` value will likely be excluded from those splits (falling into an "Unknown" bucket or disappearing entirely).

This elevates the priority from a "data quality annoyance" to a "financial reporting risk."

Assumptions Made

During this process, I made three key assumptions:

1. Direct Pipeline Flow: I assumed that `checkout-api` is the primary source of truth that feeds into `analytics.orders`. If there were an intermediate transformation layer that filled these nulls, the alert wouldn't have triggered.
2. Consistent Timezones: I assumed the `deploy\_at` timestamp (UTC) aligns with the "yesterday/today" window of the alert.
3. Data Distribution: I assumed that roughly 18% of orders historically had "missing geo" data (previously represented as empty strings). The spike isn't necessarily an increase in missing data, but an increase in how that missing data is represented.

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.