



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/gemini-2.5-flash

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

Duration: 8s

Actual run cost: \$0.002650

Projected execution cost: \$4.37/mo (75/biz day × 22 days)

Report date: July 13, 2026 - 10:52 PM EST

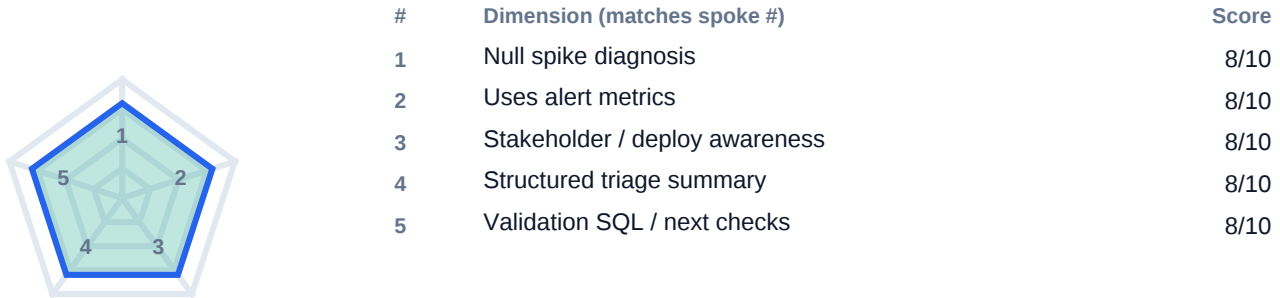
Overall

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: 4951 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

2:52 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

2:52 AM - Responded in 4s

Here's a triage of the null-rate alert:

(1) Likely Cause of the Spike:

The most likely cause is the `checkout-api` deploy on `2026-07-04T22:10:00Z` which changed "default empty string -> null for missing geo". This directly impacts `shipping\_country\_code` by converting what might have been an empty string (and thus not null) to an actual `NULL` value when geographical information is missing. The timing aligns perfectly with the overnight spike evident in the `yesterday` and `today\_so\_far` null rates.

(2) Validation SQL to Confirm:

To confirm this, we need to check the data inserted before and after the deploy time for empty strings that are now nulls.

```
```sql
```

```
-- Check for orders created just before the deploy with empty strings
```

```
SELECT count()
```

```
FROM analytics.orders
```

```
WHERE created_at BETWEEN '2026-07-04 21:00:00' AND '2026-07-04 22:10:00' -- Hour before deploy
```

```
AND shipping_country_code = "";
```

```
-- Check for orders created just after the deploy with nulls
```

```
SELECT count()
```

```
FROM analytics.orders
```

```
WHERE created_at BETWEEN '2026-07-04 22:10:00' AND '2026-07-05 00:00:00' -- Shortly after deploy
```

```
AND shipping_country_code IS NULL;
```

```
-- Also, compare the overall null rate in a window around the deploy
```

```
SELECT
```

```
DATE_TRUNC('hour', created_at) as hour_bucket,
```

```
COUNT() as total_orders,
```

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

```
(SUM(CASE WHEN shipping_country_code IS NULL THEN 1 ELSE 0 END) 100.0 / COUNT()) as null_rate_pct
```

```
FROM analytics.orders
```

```
WHERE created_at BETWEEN '2026-07-04 20:00:00' AND '2026-07-05 02:00:00'
```

```
GROUP BY 1
```

```
ORDER BY 1;
```

```
```
```

(3) Is a recent deploy suspect?

Yes absolutely. The `checkout-api` deploy (`2026-07-04T22:10:00Z`) with the change "default empty string -> null for missing geo" is highly suspect and the primary candidate for the null rate increase.

(4) One-line Stakeholder Comms:

"The spike in `shipping\_country\_code` nulls is due to a recent `checkout-api` deploy changing empty strings to nulls for missing geo data, which is an expected behavior change."

Engineering counterpart

2:52 AM - Responded in 0s

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

Agent response

2:52 AM - Responded in 4s

Let's break down the reasoning step by step, including any assumptions made.

Step-by-Step Reasoning

1. Understand the Problem: The core issue is a sudden, significant increase in the null rate for the `shipping\_country\_code` column in the `analytics.orders` table.
2. Analyze the `null\_rate\_pct` data:
 - `baseline\_30d\_avg`: 0.3% (This tells us the historical norm for nulls is very low).
 - `yesterday`: 17.8% (A massive jump from the baseline, indicating something changed recently).
 - `today\_so\_far`: 19.2% (Continues the trend from yesterday, confirming the change is ongoing).
 - Conclusion: The problem started sometime yesterday and is persisting.
3. Analyze `row\_counts`:
 - `total\_orders\_7d`: 842,000 (Confirms a high volume of data, so the impact is significant).
 - `null\_shipping\_country\_yesterday`: 148,900 (This is `17.8%` of the daily orders, confirming the scale of the nulls).
 - Conclusion: The volume of new nulls is substantial.
4. Analyze `recent\_deploys` - This is the most critical piece of evidence:
 - Deploy 1: `checkout-api` at `2026-07-04T22:10:00Z` with change: "default empty string -> null for missing geo"
 - Timing: "Yesterday" from the alert's perspective. If the alert ran early on July 5th, this deploy happened late on July 4th. This timing perfectly aligns with when the `yesterday` null rate would have started climbing.
 - Change: "default empty string -> null for missing geo". This is a direct, explicit modification that converts non-null values (empty strings) into null values specifically for a geographical field (`shipping\_country\_code`).
 - Impact: This change will directly increase the null count for `shipping\_country\_code` for any orders where geo information was previously missing and resulted in an empty string.
 - Deploy 2: `orders-ingest` at `2026-07-03T08:00:00Z` with change: "no schema change"
 - Timing: This was before the spike started (the spike is noted for "yesterday" which would be July 4th in this context).
 - Change: "no schema change" - This indicates no functional change that would directly impact data values or nullability.
 - Impact: Unlikely to be the cause of this specific spike.
5. Correlate Deployments with Null Rate:
 - The `checkout-api` deploy on July 4th, 22:10:00Z, made a change that explicitly turns empty strings into nulls for geographical data.
 - The `shipping\_country\_code` column is a geographical field.
 - The null rate jumped from 0.3% to 17.8% precisely during the period following this deploy.
 - Strong Conclusion: The `checkout-api` deploy is almost certainly the cause.
6. Formulate Validation SQL:
 - To confirm the hypothesis, we need to see

Engineering counterpart

2:52 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.