



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: inclusionai/ling-2.6-flash

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

Duration: 5s

Actual run cost: \$0.000051

Projected execution cost: \$0.0842/mo (75/biz day × 22 days)

Report date: July 13, 2026 - 10:53 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: 3457 chars - Latency: 5s - 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: 5s

Engineering counterpart

2:53 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

2:53 AM - Responded in 2s

1. Likely cause of the spike

The `checkout-api` deploy on 2026-07-04 changed the default for missing geo data from empty string `` to `NULL`. Because `shipping\_country\_code` is populated at checkout, this shift directly explains the jump from ~0.3% to ~18% nulls.

2. Validation SQL

```
```sql
```

```
SELECT
```

```
 DATE(created_at) AS dt,
```

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

```
 SUM(CASE WHEN shipping_country_code IS NULL OR shipping_country_code = '' THEN 1 ELSE 0 END) AS
null_plus_empty,
```

```
 100.0 * SUM(CASE WHEN shipping_country_code IS NULL OR shipping_country_code = '' THEN 1 ELSE 0 END) /
COUNT() AS pct_null_plus_empty
```

```
FROM analytics.orders
```

```
WHERE created_at >= CURRENT_DATE - INTERVAL '14' DAY
```

```
GROUP BY 1
```

```
ORDER BY dt DESC;
```

```
```
```

(If your warehouse uses `\_inserted\_at` or similar, swap `created\_at` accordingly.)

3. Recent deploy suspect?

Yes - the `checkout-api` change introducing `NULL` for missing geo data is the prime suspect. Revert or adjust the default/transform to an empty string as a short-term mitigation; add a test assertion for null-rate thresholds.

4. One-line stakeholder comms

The spike in `shipping\_country\_code` nulls is caused by the latest checkout-api deploy, and we're rolling back or tightening defaults; finance dashboards may show gaps until fixed.

Engineering counterpart

2:53 AM - Responded in 0s

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

Agent response

2:53 AM - Responded in 3s

Here's my step-by-step reasoning and the assumptions I made:

1. Alert context and primary symptom

- The alert is about a spike in the null rate of shipping_country_code in analytics.orders.
- The window of concern is "last 7 days," and we see yesterday's null rate jump to ~17.8% from a 30-day baseline of ~0.3%.
- Row counts show ~148,900 nulls out of ~842,000 total orders in 7 days, consistent with the reported rate.

2. Match deploy to change

- Recent_deploys lists two changes:
 - checkout-api (2026-07-04T22:10:00Z): changed the default for missing geo from empty string to NULL.
 - orders-ingest (2026-07-03T08:00:00Z): no schema change.
- The timing of the checkout-api deploy aligns with the overnight spike. I assume deploy ordering is reliable and that the code change meaningfully affects the data written to shipping_country_code.

3. Likely root cause

- The explicit change ("default empty string -> null for missing geo") directly explains the increase in NULLs.
- I assume the service writes to shipping_country_code for a large fraction of orders and that the default applies when geo enrichment fails or is absent.

4. Validation SQL intent

- I proposed a query to compare daily totals, nulls, and the combined null-plus-empty rate to check whether empties remain or were replaced by NULLs.
- I assumed the orders table has a timestamp (created_at) that can be used to slice recent days and that the warehouse syntax is compatible (e.g., CURRENT_DATE, INTERVAL).

5. Mitigation direction

- I assumed a short-term mitigation can be deployed quickly (revert default to empty string or coerce NULLs) and that null-rate thresholds/alerts will be enforced as a test in CI/CD or data quality checks.

Key assumptions stated explicitly:

- The deploy log correctly captures production changes and their timestamps.
- The orders-ingest deploy did not alter the column's behavior (no schema change).
- The shipping_country_code field is populated at order time and the default applies for missing geo data.
- The query time zone and timestamp column align with when orders are written.
- Finance and downstream dashboards depend on this column and will be affected by NULLs.

Engineering counterpart

2:53 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.