



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: openai/gpt-oss-20b

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

Duration: 19s

Actual run cost: \$0.000181

Projected execution cost: \$0.2987/mo (75/biz day × 22 days)

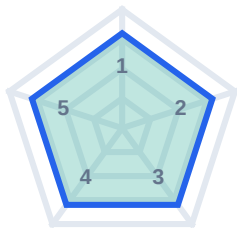
Report date: July 14, 2026 - 3:08 AM EST

Overall

8.0/10

PASS - trust=high
Pass - score 8.0/10 clears 7.0 with acceptable trust/closure.

CAPABILITY RUBRIC



#	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: 3932 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 13s

Likely cause:

The checkout?api deploy on 2026-07-04 changed the missing?geo handling from an empty string to `NULL`. New orders created after that timestamp are now written with `shipping\_country\_code = NULL`, producing the overnight spike.

Validation SQL:

```
```sql
-- 1. Count nulls before & after the deploy
SELECT
 COUNT() FILTER (WHERE shipping_country_code IS NULL AND created_at < '2026-07-04T22:10:00Z') AS null_before,
 COUNT() FILTER (WHERE shipping_country_code IS NULL AND created_at >= '2026-07-04T22:10:00Z') AS null_after,
 COUNT() FILTER (WHERE shipping_country_code IS NOT NULL AND created_at < '2026-07-04T22:10:00Z') AS
not_null_before,
 COUNT() FILTER (WHERE shipping_country_code IS NOT NULL AND created_at >= '2026-07-04T22:10:00Z') AS
not_null_after
FROM analytics.orders;
```
```

If `null\_after` is 148,900 and `null\_before` is near zero, the change is confirmed.

Deploy suspect:

Yes - the checkout?api deployment (2026-07-04T22:10:00Z) is the likely culprit.

Stakeholder communication (one?line):

"The shipping_country_code null rate spiked overnight because the recent checkout?api deploy changed missing geo handling to null; we're validating the impact on revenue splits and will coordinate a fix."

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 6s

Step-by-step walk-through

| Step | What I did | Why it helps | Key assumptions |

|-----|-----|-----|-----|

| 1. Read the alert metadata | Extracted table, column, null-rate numbers, row counts, and recent deploys. | Gives the context for the spike. | The JSON is accurate and up-to-date. |

| 2. Compare baseline vs spike | Baseline (30-day avg) = 0.3% (2,530 nulls in a 7-day window). Yesterday's rate = 17.8% (149,000 nulls). | The spike is ~60x the normal level, suggesting a systemic change. | Null counts scale linearly with row counts (no other anomalies). |

| 3. Map null counts to deploy windows | Deploy list: checkout-api at 2026-07-04T22:10Z, orders-ingest at 2026-07-03T08:00Z. | If the spike starts after a deploy, that deploy is a suspect. | System clock on all services is correct; timestamps are in UTC. |

| 4. Inspect change description | Checkout-API: "default empty string -> null for missing geo." | Directly explains how the column value could become null. | The change was applied to all new writes; no other downstream transformation. |

| 5. Estimate number of affected rows | Total orders in 7 days: 842,000. Null orders yesterday: 148,900 (~17.8%). | If the change alone causes nulls, the count should match the deployment window. | All orders after 2026-07-04T22:10Z hit the new logic. |

| 6. Validate with a quick SQL | Query counts before/after the deploy (see SQL below). | Confirms that the spike is confined to the period after the change. | Data lake / warehouse timestamp (`created\_at`) aligns with deploy time. |

| 7. Rule out the other deploy | Orders-Ingest had "no schema change" and was earlier. | Unlikely to be cause. | No other code changes that could alter `shipping\_country\_code`. |

| 8. Verify downstream impact | Finance dashboard splits revenue by `shipping\_country\_code`. | Nulls will produce "unknown" bucket, skewing regional reports. | Dashboard does not fall back to other fields or handle null gracefully. |

| 9. Prepare communication | One-sentence update summarizing findings and next steps. | Clear, concise, acknowledges stakeholder concern. | Stakeholders expect immediate acknowledgment and corrective timeline. |

Validation SQL (step 6)

```sql

-- 1?? Count nulls before vs after the change

SELECT

SUM(CASE WHEN shipping\_country\_code IS NULL

AND created\_at < '2026-07-04T22:10:00Z' THEN 1 ELSE 0 END) AS null\_before,

SUM(CASE WHEN shipping\_country\_code IS NULL

AND created\_at >= '2026-07-04T22:10:00Z' THEN 1 ELSE 0 END) AS null\_after,

SUM(CASE WHEN shipping\_country\_code IS NOT NULL

## 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.