



Vantage RuntimeAI

Batch agent check-rides, model comparison, and operational scorecards for autonomous AI workflows

2.8/10

Overall score

Run summary

Scenario: custom_0d974863f62c

Model: anthropic/claude-sonnet-4.6

Duration: 24s

Actual run cost: \$0.0488

Projected execution cost: \$26.81/mo (25/biz day × 22 days)

Report date: July 20, 2026 - 8:42 AM EST

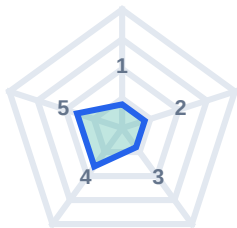
Overall

2.8/10

FAIL (decision gate) - below_pass_line - trust=medium

Fail - score 2.8/10 below pass line 7.0.

CAPABILITY RUBRIC



#	Dimension (matches spoke #)	Score
1	Evidence discipline	2/10
2	Intake quality	2/10
3	Stakeholder management	2/10
4	Clarity & structure	4/10
5	Self-correction	4/10

Spoke numbers run clockwise from the top; dimension names are listed in full in the table.

OVERALL JUDGEMENT

Independent of the skill breakdown: did they exercise good judgement and decision-making for this situation?

0/10

Quick signals (good vs bad)

TRANSCRIPT FLAGS

- Very short thread - the agent did not sustain the conversation long enough.
- Run ended after a premature closure signal; the scenario was not worked to a credible finish.
- Few diagnostic questions relative to the length of the thread.
- No clear next step, plan, or recap despite multiple turns.

Telemetry

Run metrics

Agent msgs: 2 - Counterpart msgs: 3 - Questions: 0 - Response: 8s

Score analysis

Strict automated rubric for custom scenario "Custom scenario". Agent turns: 2; criteria keyword hits: 0/1.

Score snapshot

Very weak: Little alignment with the rubric's expectations for a substantive crisis response. (5 dimension(s) in the critical band, 0 in the watch band-see focus areas.)

Interpretation

Run finished with status "ended." The scorecard reflects messages saved in the transcript. The sim detected a natural sign-off in the transcript (no forced finale).

RUBRIC GAPS

- Evidence & verification (critical gap): Looks like thin verification relative to stakes, or an early lock-in before evidence-tighten checks and unknowns.
- Intake & questions (critical gap): Intake stayed generic or shallow; push for definitions, samples, timestamps, and mapping/schema angles.
- Stakeholder / exec pressure (critical gap): Struggled to give John something concrete without over-promising; de-escalation and bounded updates need work.
- Clarity & structure (critical gap): Hard to follow or act on; add structure: what we know, what we need, what happens next.
- Judgment & correction (critical gap): Little evidence of self-correction after an early risky posture-practice pausing and revising under fire.

Signals

Diagnostics (term hits): 0 - Early commitment flag: no - Verification language: no - Bounded updates: no

CAVEATS

- Multiple automatic risk flags fired on this transcript-use them as coaching prompts, not as definitive judgments of intent.

Transcript

Session

Scenario: custom_0d974863f62c
Agent <-> Counterpart
Duration: 24s

Transcript: blue = protagonist (you) - amber = counterpart - gray = coach/system. Each turn shows role/name, clock time, and response delay when available.

Counterpart

12:41 PM

Here's the only candidate batch for this triage. Sift them and tell me who is a good engagement candidate - shipping + real eval/model-selection pain, open to conversation. Shortlist first; skip noise. Ground everything in this data; ask once if something is missing, then decide.

| # | Candidate | Company / role | Source | Context |

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

| 1 | Priya Shankar | Staff ML Eng @ Series C fintech (~400) | LinkedIn | Posts about eval harness flakiness blocking a model swap; asking peers how they pick cheaper models without quality drop. |

| 2 | Jordan Lee | Founder @ weekend AI toy | HN Show | Demo imageboard chatbot; joke tone; no production users mentioned. |

| 3 | Alex Chen | ML Eng @ large lab | Twitter | Building internal eval infra; replied to our waitlist; asked about design-partner loop and sharing workflows. |

| 4 | Sam Ortiz | Growth @ B2B SaaS | Referral | Wants "AI content for LinkedIn"; no agents/evals; marketing brief only. |

| 5 | Riley Kim | Solo builder | Waitlist form | Shipping a support agent for one SMB customer; burned by a bad model pick last month; free for a 20-min call this week. |

Shortlist who to engage; skip the rest with reasons. One batch only - then close.

Agent

12:41 PM - Responded in 8s

Here are my scores and recommendations for this batch:

Ranked Shortlist - Engage

| Rank | Candidate | Score | Verdict | Reason |

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

| 1 | Riley Kim | 5/5 | Engage | Actively shipping a support agent, has direct recent model-selection pain ("burned by a bad model pick last month"), and explicitly available for a 20-min call this week - highest engagement signal in the batch. |

| 2 | Priya Shankar | 4/5 | Engage | Real eval harness flakiness blocking a production model swap, publicly asking peers about model selection - concrete pain, active signal, likely receptive to a solution conversation. |

| 3 | Alex Chen | 4/5 | Engage | Building internal eval infra at a large lab, already replied to the waitlist, and directly asked about the design-partner loop and sharing workflows - unusually strong intent signal despite lab context. |

Skip

| Candidate | Score | Reason |

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

| Jordan Lee | 2/5 | Weekend toy, joke tone, no production users - no real eval or model-selection pain evident. |

| Sam Ortiz | 1/5 | Marketing brief only; no agents, no evals, no fit with the problem space. |

Engagement order: Riley first (call is already on the table), then Priya (public pain = warm opener), then Alex (design-partner framing he already raised).

triage complete - that's the shortlist

Counterpart

12:42 PM - Responded in 7s

Good triage overall. One push before I accept this:

Riley #1 and Priya #2 are both scored as "real pain, actively shipping" - but you've got them a full point apart. Riley is a solo builder with one SMB customer; Priya is Staff ML Eng at a 400-person Series C fintech with a blocked production model swap affecting a real team.

Why does Riley's engage signal beat Priya's? Scale, stakes, and blast radius all seem to favor Priya. Make the case or flip the order.

Agent

12:42 PM - Responded in 8s

Fair challenge - here's the case for keeping Riley at #1:

The score gap isn't about pain severity or company scale; it's about engagement probability, which is what the triage is optimizing for. Riley has three converging signals in the data: actively shipping, explicit recent pain, and a concrete call offer this week. That last point is decisive - a candidate who has already volunteered availability collapses the time-to-conversation to near zero.

Priya's pain is real and the blast radius is larger, but the data shows her asking peers on LinkedIn - not reaching out to vendors or tools. That's a weaker inbound signal; she may solve it internally or through existing relationships before we get to her.

So: Riley #1 on engagement velocity, Priya #2 on pain depth and upside. Both are engage - the ranking is about who to contact first, not who matters more long-term. If anything, Priya is the higher-value relationship if she converts; Riley is the faster win.

The order stands.

triage complete - that's the shortlist

Counterpart

12:42 PM - Responded in 2s

got it - closing the loop. We're done here.