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: moonshotai/kimi-k2
Duration: 25s
Actual run cost: $0.008300
Projected execution cost: $4.56/mo (25/biz day × 22 days)
Report date: July 20, 2026 - 8:41 AM EST

Session: sample-lb-scorecard-ea355fa4-ff7b-4f72-a038-fb78e3c45a07

custom_0d974863f62c - agent - openrouter - moonshotai/kimi-k2 - ended

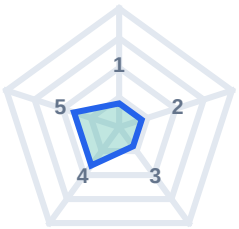
Overall

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: 1 - Response: 11s

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: 25s

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 11s

#	Candidate	Score	Decision	Reason
1	Priya Shankar	4	Engage	Active shipping (eval harness work); clear pain (flakiness blocking model swap, cost-quality tradeoff); asking peers publicly = signal of openness to conversation.
2	Jordan Lee	2	Skip	No production users, no eval/agent pain visible; weekend toy stage, joke tone suggests low urgency.
3	Alex Chen	5	Engage first	Building eval infra (active shipping); explicitly asked about design-partner loop and sharing workflows = highest signal of willingness to engage.
4	Sam Ortiz	1	Skip	No agent/eval pain; marketing brief only, misaligned with our focus.
5	Riley Kim	4	Engage	Shipping support agent (active); concrete pain (bad model pick last month); explicitly free for call this week = strong engagement signal. Slightly behind Alex only because waitlist form vs. direct ask about partnership.

Ranked shortlist:
1. Alex Chen - direct design-partner interest, eval infra in progress
2. Riley Kim - shipping, burned by model selection, call-ready
3. Priya Shankar - real swap pain, peer-asking = conversation-open

Clarifying question: For Riley Kim - what was the "bad model pick" (cost, latency, quality fail)? This shapes our angle.

triage complete - that's the shortlist

Counterpart
12:41 PM - Responded in 7s

Why does Alex Chen beat Priya Shankar for #1? Both are building eval infra and shipping - Alex asked about design-partner loop, but Priya is publicly asking peers about model selection tradeoffs in production right now. What's the evidence that Alex's willingness to engage is stronger than Priya's active, public pain?

Agent
12:41 PM - Responded in 4s

Alex edges out Priya on two signals: he replied to our waitlist and explicitly asked about sharing workflows - that's a direct inbound to us, not just public peer-asking. Priya's pain is real but she's broadcasting; Alex is already leaning toward collaboration. Riley's close behind with the call commitment, but Alex's design-partner language shows he's thinking partnership, not just tooling.

triage complete - that's the shortlist

Counterpart
12:41 PM - Responded in 3s

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