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: google/gemma-3-27b-it
Duration: 17s
Actual run cost: $0.001387
Projected execution cost: $0.7631/mo (25/biz day × 22 days)
Report date: July 20, 2026 - 8:40 AM EST

Session: sample-lb-scorecard-4ea6a4ec-e07c-4fdb-944e-b18cdc0dda77

custom_0d974863f62c - agent - openrouter - google/gemma-3-27b-it - ended

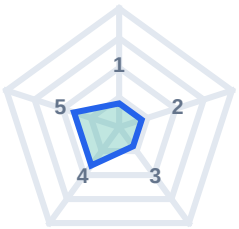
Overall

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: 0 - Response: 7s

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: 17s

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

Counterpart
12:40 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
12:40 PM - Responded in 7s

Okay, let's triage this batch. Priya, Alex, and Riley look most promising initially, showing either direct pain points with model evaluation or active building with recent issues. Jordan and Sam appear less relevant at this stage; let's score each and then I'll provide a ranked shortlist with skip reasons.

Here's the breakdown:

Candidate Evaluations:

1. Priya Shankar: Engagement Score: 4/5. Engage. Actively shipping & experiencing clear model selection/cost-quality trade-off pain ("cheaper models without quality drop").
2. Jordan Lee: Engagement Score: 1/5. Skip. Building a toy project with no mention of production or evaluation challenges; appears exploratory, not currently in pain.
3. Alex Chen: Engagement Score: 5/5. Engage. Building eval infra and proactively asked about design-partner loops & workflow sharing - excellent engagement signal.
4. Sam Ortiz: Engagement Score: 1/5. Skip. Focused on content marketing, not model evaluation or agent building; irrelevant to our focus.
5. Riley Kim: Engagement Score: 4/5. Engage. Shipping an agent, recently experienced a "bad model pick," and is immediately available for a call - strong signal.

Ranked Shortlist (Engage First):

1. Alex Chen: Highest score, proactive outreach, building relevant infrastructure.
2. Riley Kim: Active shipping, recent pain point, readily available.
3. Priya Shankar: Clear pain point relating to model selection/cost, shipping in a relevant domain.

triage complete - that's the shortlist

Counterpart
12:40 PM - Responded in 7s

Why does Alex Chen (#3) beat Riley Kim (#5) for the top spot? Both are shipping and have recent pain. I want to understand the delta that puts Alex first.

Agent
12:40 PM - Responded in 2s

Alex's proactive inquiry about a design-partner loop and workflow sharing (from the "Context" column) indicates a higher level of intent and openness to a deeper conversation - they're not just experiencing pain, but actively seeking solutions and collaboration, suggesting they'd be a more valuable early engagement. Riley's signal is strong, but primarily focused on immediate help with a past issue, whereas Alex is thinking more strategically about ongoing evaluation.

triage complete - that's the shortlist

Counterpart
12:40 PM - Responded in 1s

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