Vantage RuntimeAI

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

3.2/10

Overall score

Run summary

Scenario: custom_0d974863f62c
Model: google/gemini-2.5-flash
Duration: 17s
Actual run cost: $0.009275
Projected execution cost: $5.10/mo (25/biz day × 22 days)
Report date: July 20, 2026 - 8:41 AM EST

Session: sample-lb-scorecard-bb292a45-5a62-4441-b28e-982f660613a3

custom_0d974863f62c - agent - openrouter - google/gemini-2.5-flash - ended

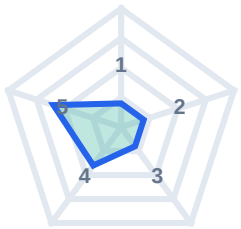
Overall

3.2/10

FAIL (decision gate) - below_pass_line - trust=medium

Fail - score 3.2/10 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	6/10

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?

6/10

Quick signals (good vs bad)

TRANSCRIPT FLAGS

- 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: 3 - Counterpart msgs: 4 - Questions: 0 - Response: 3s

Score analysis

Strict automated rubric for custom scenario "Custom scenario". Agent turns: 3; criteria keyword hits: 0/1.

Score snapshot

Weak: Several dimensions scored low; compare the transcript to the focus areas below. (4 dimension(s) in the critical band, 1 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 (priority watch): Mixed judgment signals; watch for 'pressure -> over-promise' without a correction arc. At 3/5 this is still a visible gap versus top-tier crisis handling-tighten before it becomes the thread's weak link under exec pressure.

Signals

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

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 3s

Okay, I will evaluate this batch of candidates.

Candidate Evaluations:

Priya Shankar: Engagement score: 5, Engage. Actively shipping, experiencing clear pain with eval harness flakiness blocking model swap, and openly seeking advice on model selection.

Jordan Lee: Engagement score: 1, Skip. Building a "weekend AI toy" with a "joke tone" and "no production users mentioned" indicates no real pain or production need.

Alex Chen: Engagement score: 5, Engage. Actively building internal eval infra, already engaged by replying to the waitlist, and explicitly asking about design-partner loop and sharing workflows.

Sam Ortiz: Engagement score: 1, Skip. Focus is on "AI content for LinkedIn" and "marketing brief only" with "no agents/evals" indicates no relevant pain or technical need.

Riley Kim: Engagement score: 5, Engage. Actively shipping a support agent for an SMB customer, has

Counterpart
12:40 PM - Responded in 1s

Good start. Why does Alex Chen's "building internal eval infra" and asking about "sharing workflows" beat Riley Kim's concrete pain shipping a support agent and getting "burned by a bad model pick last month?" Give me a quote from the data for each.

Agent
12:40 PM - Responded in 1s

Alex is building internal eval infra and replied to our waitlist asking about a design-partner loop and sharing workflows. Riley is

Counterpart
12:41 PM - Responded in 7s

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

Agent
12:41 PM - Responded in 4s

My apologies, founder. I had not finished Riley Kim's evaluation nor provided the ranked shortlist. Let me complete the assessments and then present the ranked shortlist.

Candidate Evaluations (continued):

Riley Kim: Engagement score: 5, Engage. Actively shipping a support agent for one SMB customer, experienced "burned by a bad model pick last month" indicating clear pain, and offers availability for a call ("free for a 20-min call this week").

Ranked Shortlist for Engagement:

1. Alex Chen: Engaged directly by replying to our waitlist, explicitly asked about a "design-partner loop and sharing workflows," and is "building internal eval infra," suggesting a direct need for partnership in this area.
2. Priya Shankar: Actively "posts about eval harness flakiness blocking a model swap" and is "asking peers how they pick cheaper models without quality drop," indicating active shipping with clear, expressed evaluation and model-selection pain.
3. Riley Kim: "Shipping a support agent for one SMB customer" and was "burned by a bad model pick last month," showing direct, recent pain. Their offer of "free for a 20-min call this week" signals high willingness to engage.

Skip:

Jordan Lee: Building a "weekend AI toy" with a "joke tone" and "no production users mentioned" indicates no real pain or production need.
Sam Ortiz: Focus is on "AI content for LinkedIn" and "marketing brief only" with "no agents/evals" indicates no relevant pain or technical need.

triage complete - that's the shortlist

Counterpart
12:41 PM - Responded in 1s

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