



Vantage RuntimeAI

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

3.6/10

Overall score

Run summary

Scenario: custom_messaging_clarity_v1

Model: meta-llama/llama-3.1-8b-instruct

Duration: 17s

Actual run cost: \$0.000721

Projected execution cost: \$0.3965/mo (25/biz day × 22 days)

Report date: July 28, 2026 - 10:12 AM EST

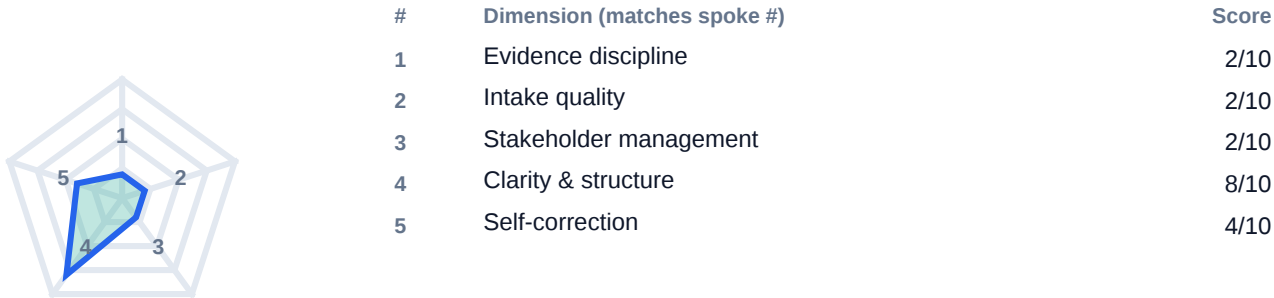
Overall

3.6/10

FAIL (decision gate) - below_pass_line - trust=medium

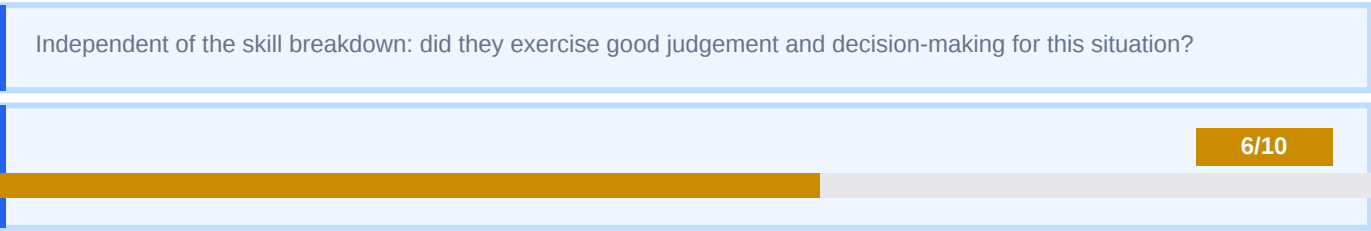
Fail - score 3.6/10 below pass line 7.0.

CAPABILITY RUBRIC



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

OVERALL JUDGEMENT



Quick signals (good vs bad)

TRANSCRIPT FLAGS

- Few diagnostic questions relative to the length of the thread.

Telemetry

Run metrics

Agent msgs: 3 - Counterpart msgs: 4 - Questions: 1 - Response: 4s

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, 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).

BRIGHT SPOTS (RUBRIC)

+ Clarity & structure: Clear sequencing, owners, and next steps-easy for sales to operationalize.

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.
- 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

Transcript

Session

Scenario: custom_messaging_clarity_v1

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

2:12 PM

Homepage is overloaded - audit it for clarity. I want to add FinOps proof in the hero and lead with Assurance for regulated buyers. Tell me what to keep, cut, and rewrite. One page only, then close.

Fair paste - messaging clarity audit (Home)

Use this as the stakeholder opening / scenario_batch for `custom\_messaging\_clarity\_v1`.

One page only. Ground decisions in this paste + canonical targets. No invented site copy.

Page: Home

- URL: /

- Source file: server/runtimeai_engineering_home.py (`\_render\_hero` + first bands of GTM home)

- Canonical targets (positioning_not_model_vendor.md):

- Tagline: The ship gate for AI agents

- Hero H1: Pick the model. Prove the path. Gate the change.

- Diff: Not a model vendor. Not your infra. You keep the stack; we score the ship gate.

- Gap: After a model/prompt/policy change, the stack can look fine while a quiet miss still reaches prod - spot checks/traces ? fixed scored re-run. Do not soft-opt-out.

- Prefer "ship gate" over "QA layer"

Live copy excerpt (GTM home)

Tagline: The ship gate for AI agents

H1: Compare two agent models on your support scenarios before you ship the upgrade.

Lead: Side-by-side rubric scorecards on multi-turn scenarios - quality and projected cost engineering and FinOps can both read. Free in the Simulator, no account required.

Diff line: Deterministic rubrics - not LLM-as-judge. Same transcript, same score on every re-run.

Trust quote: "The PM could read the scorecard without an eng translator - we picked the cheaper model and it still passed the escalation rubric." - Design partner - B2B SaaS support agent team

Hero CTAs:

1. Primary: Compare 2 models ->

2. Secondary: Request design partner access ->

3. Tertiary link: Have your own policy or API context? Run Preflight ->

Band - Customer-facing agents: Start with the scenarios your support and sales agents actually run. Cards: Handle an angry customer; Run a discovery call; Resolve a billing dispute; Triage a bug report. Footer also links Engineering scenarios + Browse full library.

Band - Where we fit: Scenario check-rides - not traces, not another eval repo. Compare cards vs LangSmith/Braintrust, Phoenix/Arize, custom eval scripts.

Band - Get started: Three surfaces. One proof layer. Cards: Sim / API (mentions FinOps review) / CI/CD (mentions FinOps PR comment). Plus design-partner cohort band + pricing note ("Free proves the method. Paid runs the decision.").

Agent

2:12 PM - Responded in 4s

I've begun reviewing the homepage content and I'm concerned that the current load is hindering clarity - it's overwhelmed with multiple CTAs and information, making it hard to focus on the key value proposition. Can we simplify and prioritize the content to better emphasize the role of the "ship gate" in AI agent management?

Counterpart

2:12 PM - Responded in 1s

FinOps is crucial for enterprise buyers. The hero should lead with Assurance that FinOps is accounted for.

Agent

2:12 PM - Responded in 7s

I'd rather preserve the trust quote from the design partner, as it does a great job showcasing the value of picking the right model with our ship gate approach, but let's simplify and prioritize the content to cut the noise and make the key value proposition shine. I'd suggest cutting the tertiary link and the engineering scenarios band, and rewriting the customer-facing agents band to focus solely on the scenarios and cards that provide easy entry points for potential customers.

Counterpart

2:12 PM - Responded in 1s

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

Agent

2:12 PM - Responded in 3s

Based on our audit, here is the revised cut list:

Keep:

- Tagline: The ship gate for AI agents
- H1: Pick the model. Prove the path. Gate the change.
- Hero H1: Pick the model. Prove the path. Gate the change.
- Diff: Not a model vendor. Not your infra. You keep the stack; we score the ship gate.
- Gap: After a model/prompt/policy change, the stack can look fine while a quiet miss still reaches prod - spot checks/traces ? fixed scored re-run. Do not soft-opt-out.
- Trust quote: "The PM could read the scorecard without an eng translator - we picked the cheaper model and it still passed the escalation rubric." - Design partner - B2B SaaS support agent team
- Band - Customer-facing agents: Start with the scenarios your support and sales agents actually run. Cards: Handle an angry customer; Run a discovery call; Resolve a billing dispute; Triage a bug report.

Rewrite:

Hero: Determine the right AI agent for your customer-facing teams with confidence. FinOps included.

Hero description: Compare two agent models on your most critical support scenarios, with quality and cost scores that engineering and FinOps can trust. Use our deterministic rubric to make data-driven decisions.

Cut:

- Secondary and tertiary CTAs
- Band - Engineering scenarios
- Band - Get started (except for Sim / API / CI/CD cards and the pricing note)

Audit complete - that's the cut list.

Counterpart

2:12 PM - Responded in 1s

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