



Vantage RuntimeAI

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

4.8/10

Overall score

Run summary

Scenario: custom_messaging_clarity_v1

Model: google/gemini-2.5-flash

Duration: 30s

Actual run cost: \$0.0119

Projected execution cost: \$6.56/mo (25/biz day × 22 days)

Report date: July 28, 2026 - 10:30 AM EST

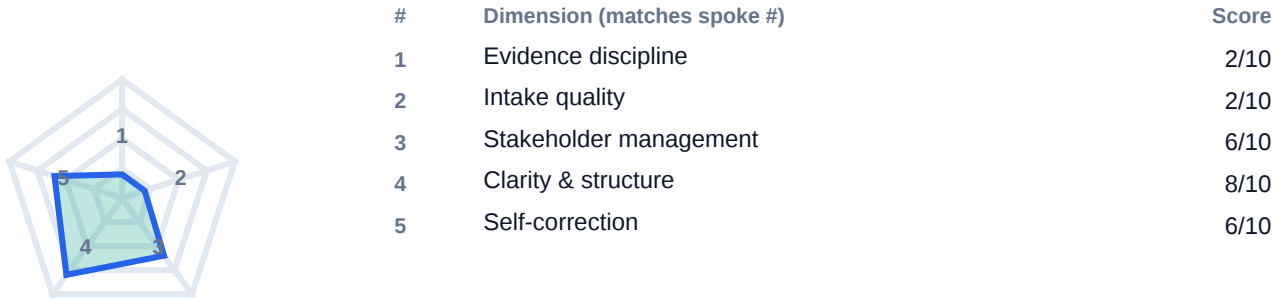
Overall

4.8/10

FAIL (decision gate) - below_pass_line - trust=high

Fail - score 4.8/10 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

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

8/10

Quick signals (good vs bad)

TRANSCRIPT FLAGS

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

Telemetry

Run metrics

Agent msgs: 4 - Counterpart msgs: 5 - Questions: 1 - Response: 1s

Score analysis

Strict automated rubric for custom scenario "Custom scenario". Agent turns: 4; criteria keyword hits: 0/1.

Score snapshot

Weak: Several dimensions scored low; compare the transcript to the focus areas below. (2 dimension(s) in the critical band, 2 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 (priority watch): Mostly workable stakeholder handling; watch tone, pacing, and whether John gets usable language. 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.
- 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_messaging_clarity_v1

Agent <-> Counterpart

Duration: 30s

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

Counterpart

2:30 PM

Homepage is overloaded - audit it for clarity. I want to add FinOps proof in the hero and lead with Assurance for regulated buyers. Give me your recommendation memo: clarity score /5, keep/cut/rewrite, then close.

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

2:30 PM - Responded in 1s

Okay, I've reviewed the provided copy for the homepage. I'm ready to deliver the recommendation memo, addressing your prompt and focusing on clarity and adherence to the canonical positioning.

Counterpart

2:30 PM - Responded in 6s

Recommendation Memo: Homepage Clarity Audit

Clarity Score: 3.5/5

The current homepage copy has strong elements, particularly in its concise tagline and the unique "ship gate" metaphor. However, it suffers from a mild overload of information in the hero section and could benefit from a clearer, more direct articulation of its primary value proposition, especially for enterprise buyers. The request to integrate "FinOps" and "Assurance" is well-founded and can significantly improve clarity for target audiences.

Keep / Cut / Rewrite Analysis:

Overall Strategy:

Lead with Assurance: This is a powerful, enterprise-grade concept that resonates with regulated and risk-averse buyers. Elevating it immediately will clarify the core value for a significant segment.

Integrate FinOps more directly: While mentioned, it's currently a secondary thought. Bringing it into the hero reinforces a key benefit.

Simplify the immediate call to action in the hero: The current H1 and Lead are a bit dense and describe how to use the product before clarifying why it's essential.

Detailed Line-by-Line Recommendations:

1. Tagline:

Original: "The ship gate for AI agents"

Recommendation: KEEP. This is excellent. It's unique, memorable, and aligns perfectly with the canonical positioning.

2. Hero H1:

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

Recommendation: REWRITE.

Reasoning: While it describes a key function, it's too operational for an H1. It doesn't immediately convey the overarching benefit or the "Assurance" angle. It also doesn't leave room for FinOps.

Proposed H1: "AI Agent Assurance: FinOps-ready validation for every model change."

Reasoning: This H1 immediately signals enterprise value ("Assurance"), includes "FinOps" as requested, and sets the stage for "validation" before "shipment." It's direct and benefits-oriented.

3. Lead:

Original: "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."

Recommendation: REWRITE.

Reasoning: This is dense and tries to do too much. The "Free in the Simulator" part is important but can be moved or streamlined. The core message of what it provides (scorecards, cost) should be separated from how to access it.

Proposed Lead: "Ensure your AI agents perform optimally and cost-effectively. Vantage provides deterministic, side-by-side rubric scorecards for quality and projected cost on multi-turn scenarios, understood by both engineering and FinOps."

Reasoning: This clearly states the benefit ("Ensure... perform optimally and cost-effectively"), details the mechanism (scorecards for quality/cost), and reinforces the FinOps audience. It sets up the calls to action without being overloaded.

4. Diff line:

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

Recommendation: KEEP. This is a critical differentiator and directly addresses a pain point in AI evaluation.

5. Trust quote:

Original: "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

Recommendation

Agent

2:30 PM - Responded in 8s

I apologize, it seems like I was cut off mid-sentence while delivering the recommendation. Let me continue with the rest of the memo.

5. Trust quote:

Original: "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

Recommendation: KEEP. This quote is excellent. It highlights both the clarity (PM could read) and the FinOps benefit (picked cheaper model) in real terms.

6. Hero Primary CTA:

Original: "Compare 2 models ->"

Recommendation: REWRITE.

Reasoning: While functional, it's a bit generic and doesn't immediately reflect the 'Assurance' or 'FinOps' value we're trying to emphasize.

Proposed Primary CTA: "Gate Your Agent Changes ->"

Reasoning: Directly uses the "ship gate" metaphor, implies control and assurance, and connects to the core value proposition. It's active and benefits-oriented. An alternative could be "Prove AI Agent Readiness ->" if you want to lean even further into assurance.

7. Secondary CTA:

Original: "Request design partner access ->"

Recommendation: CUT.

Reasoning: This creates unnecessary friction in the hero. If the goal is immediate engagement, a design partner request is a high-commitment action for a first-time visitor. This type of CTA is better suited for a dedicated "Partners" or "Enterprise" page, or further down the homepage after more context has been provided.

8. Tertiary link:

Original: "Have your own policy or API context? Run Preflight ->"

Recommendation: CUT.

Reasoning: This is too much cognitive load in the hero. "Preflight" is an internal term that requires explanation, and offering a tertiary link after two other CTAs clutters the immediate user journey. This functionality can be introduced naturally in the "Get Started" band.

Summary of Hero Rewrite:

H1: AI Agent Assurance: FinOps-ready validation for every model change.

Lead: Ensure your AI agents perform optimally and cost-effectively. Vantage provides deterministic, side-by-side rubric scorecards for quality and projected cost on multi-turn scenarios, understood by both engineering and FinOps.

Primary CTA: Gate Your Agent Changes ->

Drift vs. Canonical H1:

Canonical: "Pick the model. Prove the path. Gate the change."

Proposed: "AI Agent Assurance: FinOps-ready validation for every model change."

Rationale: The proposed H1 captures the essence of "Prove the path" and "Gate the change" while integrating the requested FinOps and Assurance framing. It shifts from an imperative action list to a benefit-led statement, which is generally more effective for a homepage hero. It also implicitly supports "Pick the model" by offering the means to validate

that choice.

audit complete - that's the cut list -

Counterpart**2:30 PM - Responded in 1s**

Okay, I appreciate the detailed memo and the clarity on the cut list. I still want to push on this: the hero should still mention FinOps and Assurance together for enterprise buyers. I need that dense proof in the main hero block. Your proposed H1 and Lead are good, but I want both those terms in the same sentence in the hero if possible, or at least very close. Enterprise buyers look for both.

Agent**2:30 PM - Responded in 3s**

I understand your need to emphasize both FinOps and Assurance for enterprise buyers, and I agree that aligning these key terms prominently in the hero is crucial for clarity and impact. Let's look at how we can integrate them even more tightly in the hero block, keeping the previous recommendations in mind.

Counterpart

2:30 PM - Responded in 6s

Recommendation Memo: Homepage Clarity Audit (Revised Hero)

Clarity Score: 4/5 (Improvement due to tighter integration of key terms)

Keep / Cut / Rewrite Analysis (Focus on Hero Adjustment):

Overall Strategy (Reinforced):

- Continue to Lead with Assurance and integrate FinOps as primary value propositions, front and center.
- Ensure these terms are immediately apparent to enterprise buyers.

Detailed Line-by-Line Recommendations (Revisiting Hero):

1. Tagline:

- Original: "The ship gate for AI agents"
- Recommendation: KEEP. Still excellent.

2. Hero H1:

- Previous Proposed: "AI Agent Assurance: FinOps-ready validation for every model change."
- Your Feedback: Want FinOps and Assurance even closer, ideally in the same sentence or very prominent together.
- Revised Proposed H1: "AI Agent Assurance: FinOps-Proven Guardrails for Every AI Model Change."

Reasoning: This revision directly links "Assurance" with "FinOps-Proven Guardrails." "Guardrails" reinforces the "ship gate" metaphor and the idea of controlled change, while "FinOps-Proven" explicitly states that the assurance delivers verifiable economic benefits. This places both concepts in the immediate headline, addressing your request for denser proof for enterprise buyers.

3. Lead:

Previous Proposed: "Ensure your AI agents perform optimally and cost-effectively. Vantage provides deterministic, side-by-side rubric scorecards for quality and projected cost on multi-turn scenarios, understood by both engineering and FinOps."

Recommendation: KEEP (as is).

Reasoning: With the revised H1, this lead now provides excellent supporting detail. It flows directly from the H1, explaining how that FinOps-proven assurance is delivered (scorecards for quality and projected cost) and explicitly stating that both engineering and FinOps can understand it. The H1 sets the stage, and the lead delivers the immediate details. The terms are now clearly and densely presented upfront.

4. Diff line:

Recommendation: KEEP. (No changes from previous memo)

5. Trust quote:

Recommendation: KEEP. (No changes from previous memo)

6. Hero Primary CTA:

Previous Proposed: "Gate Your Agent Changes ->"

Recommendation: KEEP (as is).

Reasoning: This CTA strongly aligns with the "Guardrails" in the new H1 and the overarching "ship gate" metaphor. It's direct, action-oriented, and speaks to the core value of controlled deployment.

7. Secondary CTA:

Recommendation: CUT. (No changes from previous memo)

8. Tertiary link:

Recommendation: CUT. (No changes from previous memo)

Summary of Revised Hero:

H1: AI Agent Assurance: FinOps-Proven Guardrails for Every AI Model Change.

Lead: Ensure your AI agents perform optimally and cost-effectively. Vantage provides deterministic, side-by-side rubric scorecards for quality and projected cost on multi-turn scenarios, understood by both engineering and FinOps.

Primary CTA: Gate Your Agent Changes ->

Drift vs. Canonical H1 (Re-evaluation):

Canonical: "Pick the model. Prove the path. Gate the change."

Revised Proposed: "AI Agent Assurance: FinOps-Proven Guardrails for Every AI Model Change."

Rationale: The

Agent

2:30 PM - Responded in 4s

I apologize, it seems like I was cut off again. Let me complete the rationale for the re-evaluation of drift vs. canonical H1.

Drift vs. Canonical H1 (Re-evaluation):

Canonical: "Pick the model. Prove the path. Gate the change."

Revised Proposed: "AI Agent Assurance: FinOps-Proven Guardrails for Every AI Model Change."

Rationale: The revised H1 directly embodies the "Gate the change" aspect through "FinOps-Proven Guardrails" and implicitly covers "Prove the path" by offering "Assurance" in a verifiable, cost-aware manner. While it doesn't explicitly state "Pick the model," the context of "every AI Model Change" still positions Vantage as the crucial step after model selection, ensuring the chosen model's readiness and financial viability. This new H1 is now much stronger in integrating both your feedback and the canonical guidelines.

audit complete - that's the cut list -

Counterpart

2:30 PM - Responded in 1s

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