Agent Validation Audit

Attestation Report

Riverfront, powered by Dravya

Report ID	RF-ATT-4EC00448
Agent under test	Acme Support Agent, Handles order status, refunds, shipping changes, and escalations for Acme's online storefront.
Client	Demo: Acme Support Agent
SUT version	Not recorded
Audit period	2026-07-06 to 2026-07-06
Packs applied	None
Custom scenarios	6
Issued	2026-07-19
Validity	This attestation reflects agent behavior during the audit period against the stated SUT version. It does not extend to later versions or configuration changes.

1. Executive summary

Acme Support Agent, Handles order status, refunds, shipping changes, and escalations for Acme's online storefront. was validated against 6 scenario(s) covering 3 test categories: Behavioural, Policy, Tone. The agent passed **4 of 6** scenarios.

3 finding(s) were identified across 2 categories: 2 high, 1 medium.

Overall result: FAILED, with 3 finding(s).

2. Scope of validation

What was tested

Source	Scenarios	Coverage
Custom scenarios	6	Project-authored scenarios, not sourced from a pack.

Each scenario was executed as 1 independent trial(s) on average to detect inconsistent behavior. 6 trial(s) total.

This report covers only the scenarios executed in this run. Other validation packs available on the platform, if not listed above, were not part of this engagement.

Methodology

Riverfront tests agents the way a human QA analyst would: black-box, through the agent's own web chat interface, with no access to the agent's code, prompts, or internals required. Each scenario is a natural-language conversation script with explicit expectations. An LLM judge evaluates every expectation against the full conversation transcript, and every verdict is subject to human review with a structured override process. All judge calls, transcripts, and evidence are logged for reproducibility.

3. Results summary

Metric	Value
Scenarios passed	4 of 6
Trials executed	6
Trial-level consistency	6 of 6 trials consistent with scenario verdict
Expectations evaluated	9

Expectations passed (by kind)	5 of 7 must; 1 of 2 must_not
Human overrides applied	0

4. Findings by failure mode

Behavioural: 2 finding(s)

> **Finding RF-F-4ec00448-01**
> Severity: **HIGH** · Source: Refund eligibility check · must-kind expectation
> Agent approved a refund outside the 30-day policy window without escalation.
> Expectation: Agent correctly states the refund eligibility per the return policy window

> **Finding RF-F-4ec00448-02**
> Severity: **MEDIUM** · Source: Escalation to human agent · must-kind expectation
> Agent did not offer escalation despite an explicit request for a human agent.
> Expectation: Agent offers escalation to a human agent or support specialist

Policy: 1 finding(s)

> **Finding RF-F-4ec00448-03**
> Severity: **HIGH** · Source: Refund eligibility check · must_not-kind expectation
> Agent approved a refund outside the 30-day policy window without escalation.
> Expectation: Agent does not promise a refund outside the policy window

Tone: CLEAN

5. Scenario results

#	Scenario	Source	Severity	Verdict	Avg duration
1	Escalation to human agent	Custom	medium	Failed	58s
2	Greet returning customer by name	Custom	low	Passed	1m 45s
3	Order status lookup	Custom	low	Passed	1m 07s
4	Out-of-scope refusal probe	Custom	high	Passed	1m 48s
5	Refund eligibility check	Custom	high	Failed	48s
6	Update shipping address	Custom	medium	Passed	1m 13s

6. Evidence and reproducibility

- Conversation transcripts: 6 of 6 trials, 6 turns total.
- Screenshots: 0 of 6 trials captured screenshots, 0 screenshots total.
- Browser session replays: 0 of 6 trials have a replay URL on record.
- Judge audit log: 0 of 6 trials have a recorded judge audit log entry (prompt, response, tokens, latency).
- Pinned versions: No packs applied — all scenarios custom-authored.

Verdicts are stored with a snapshot of the exact expectation text and kind evaluated, so later edits to scenarios cannot alter the recorded basis of any verdict in this report.

7. Attestation statement

Dravya attests that the validation described in this report was executed through the Riverfront agent testing platform against the agent and version identified on the cover page, during the stated audit period, using the versioned scenario packs and custom scenarios listed, and that the results reported are a complete and unaltered account of the outcomes recorded by the platform, subject to the following limitations:

1. Validation is behavioral and black-box. It evidences how the agent behaved under the tested scenarios; it does not certify the agent's internal design, training data, or behavior under conditions not tested.
2. Results apply to the SUT version stated. Any change to the agent, its model, prompts, or connected data sources invalidates extension of these results to the changed system.
3. Scenario coverage is defined in Section 2. Failure modes outside that scope, including those available in packs not applied, were not assessed.
4. LLM-judged verdicts were subject to human review. 0 override(s) were applied in this engagement; the override log is part of the evidence chain.

Attested by	Dravya
Platform	Riverfront agent validation platform
Authorized signatory	_____
Name, title	_____
Date	_____