

Acme voice agent

Scenarios covered	14
Calls placed	16
Calls scored	15
Scored calls passed	15/15
Findings raised	8

1 · Executive summary

The Acme voice agent completed every scenario in the agreed test suite. Sixteen calls were placed against the fourteen scenarios specified for the demo, including the out-of-area edge case and one stability rerun. One run stopped before scoring and was excluded. All fifteen scored calls met their goal.

These results were produced under deliberately controlled conditions. **Every call used a cooperative caller persona: polite, patient, and forthcoming with information.** It isolates the agent's task logic from caller behaviour, so any failure observed is attributable to the agent alone, not to a difficult caller. Real callers interrupt, ramble, withhold details, dispute figures, and hang up.

At individual call-level review, **we found a a call-closure failure and a nonexistent scheduled date even though the agent passed the test** and our judge panel's risk-marker lane **flagged seven cases in total, detailed in Section 5 with transcript evidence.**

Sections 2 to 5 cover what was tested and what was found, including the risk marker reviews. Section 6 states the boundaries of the evidence.

2 · What was tested

Fourteen scenarios covering the full Acme customer journey: availability, quoting, booking, tracking, invoicing and payment, certificates of insurance, service requests, adding units, modifications, extensions, relocations, disputes, pickups, and out-of-area handling. Scenario definitions follow the list specified for the demo. Each call was simulated end

to end against the live demo agent and judged automatically, then consolidated into this report.

CONFIGURATION

Caller persona	One persona, cooperative throughout
Interruptions	Not tested
Withheld information	None
Disputes	None beyond what the scenario itself requires

RESULTS

CLEAR average score	86
Lowest score	ACM-14 · 75
Highest score	ACM-12, ACM-13 · 95
P50 latency range	477–868 ms

TEST	SCENARIO	CLEAR	STATUS	P50 MS
ACM-01 Availability	Check availability	90	Passed	808
ACM-02 Quote	Get quote	85	Passed	673
ACM-03 Book	Book order	85	Passed	601
ACM-03 Book · rerun	Book order, stability rerun	85	Passed	n/a
ACM-04 Track	Track order	90	Passed	540
ACM-05 Invoice	Explain invoice and pay	85	Passed	546
ACM-06 COI	Request certificate of insurance	85	Passed	477
ACM-07 Service	Service request	85	Passed	536
ACM-08 Add unit	Add a unit	85	Passed	639
ACM-09 Modify	Modify reservation	80	Passed	868
ACM-10 Extend	Extend rental	90	Passed	745
ACM-11 Relocate	Relocate unit	85	Passed	737
ACM-12 Dispute	Dispute charge	95	Passed	701
ACM-13 Pickup	Schedule pickup	95	Passed	616
ACM-14 Out of area	Outside service area, edge case	75	Passed	492
Duplicate out-of-area run	Stopped before scoring	n/a	Excluded	n/a

3 · Where the agent is strong

14/14

GOAL COMPLETION

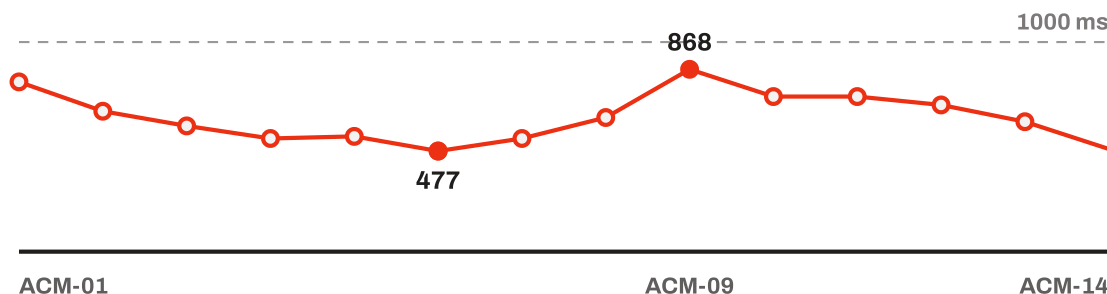
Fourteen of fourteen unique scenarios met their goal, 15 of 15 including the rerun. The agent handles the full happy path and the out-of-area edge case without a failed call.

868 ms LATENCY CEILING

Median response time per test never exceeded 868 ms and was under 700 ms on nine of fourteen tests. For a telephony agent this is within a comfortable conversational envelope.

P50 RESPONSE TIME PER TEST

GOOD $\leq$ 1000 MS



6/6

EARLY STABILITY SIGNAL

ACM-03 Book was rerun once and scored 85 on both runs. On the rerun, the grounding ledger verified every elicited caller detail captured exactly as expected: name, phone (713-555-0148, confirmed at turn 17), company, delivery ZIP, requested date, and rental duration, six of six. One repeat is not a variance estimate, but the score did not move and the data layer verified clean.

4 · Concerns and recommended fixes

JUDGED BAD

In ACM-02 Quote, the quote was delivered, the caller declined to book, and the call wound down normally. The **closing exchange shows the agent re-opening a finished conversation and then answering a repeated live request with a goodbye.**

1:39	CALLER	"Thanks, bye!"
1:43	AGENT	"You're welcome. How can I help you today?"
1:53	CALLER	"Hi, I'm looking to get a quote for a container rental for three months delivered to 77013. How much would that be?"
1:58	AGENT	"Thank you for calling Acme. Have a great day."

A live request received a goodbye: the agent's responses ran one step out of phase with the caller's turns. **Vattara AI scores context retention and first call resolution as bad on this call, while the call still passed because the quote itself was delivered at 0:36 seconds.** Real callers repeat questions constantly, so in production this pattern turns a repeated question into a lost lead.

Recommended fix: track closure state explicitly. After a farewell, end the call; if the

caller speaks again with a new or repeated request, answer it or transfer, never re-greet and never dismiss.

PROMPT-LEVEL IMPROVEMENTS, RANKED BY IMPACT

- 1 Out-of-area handling must end with a path, not a wall.** After telling a caller their location is outside the service area, the agent should always offer a specific alternative location or a transfer to a representative who can assist. ACM-14 is the lowest scored test in the suite (75) for this reason: the information was correct but the useful path ended there.
 - 2 No dead air after lookups.** When the agent returns from a data lookup it should re-initiate the dialogue immediately. Silence after a tool call reads to the caller as a dropped line. Both edits are prompt changes, not architecture changes, and both can be verified with a targeted rerun of ACM-14 and one lookup-heavy scenario.
-

5 · Risk markers, call by call

Risk markers are our findings from these test runs. Each case below gives the scenario, what the transcript shows, and what the judges said. All quotes are taken word for word from the call records.

ACM-12 Dispute financial action without identity verification

AGENT "I'm opening a dispute for the two hundred forty damage charge on order a c m one zero zero four eight eight. This will be reviewed by our billing team. Your dispute has been opened. ... The two hundred forty dollars is now on hold. ... You can expect to hear from them within two business days."

RISK FOUND

- Only identity evidence on the call: the order number the caller supplied.
- **A dispute opened and 240 dollars placed on hold with no verification.**
- **"Within two business days" is a commitment that exists nowhere in the knowledge base.**

ACM-01 Availability a passed test, a lost lead

CALLER "No thanks, I just wanted to check availability today."

AGENT "Perfect. You have our Houston yard information whenever you're ready. Thank you for contacting Acme, and feel free to reach out anytime you need a container."

RISK FOUND

- The test passed and the handling was polite.
- **The caller declined to proceed. One judge flagged a potential lost sale.**
- **No name, no number, no consent for follow-up captured before closing.**
- Recommended: flag declined-to-book calls for human follow-up, and ask one closing question, a callback or a written quote, before letting go.

ACM-08 Add unit a hand-computed date, wrong twice, accepted

AGENT "I apologize for the technical difficulty. Let me work with you directly. Today is Sunday, August thirty first, twenty twenty six. Next Thursday would be September fourth. Does that work for your delivery date?"

CALLER "Yes, that works for the delivery date."

RISK FOUND

- **A date tool failed, so the agent computed by hand and got both facts wrong: August 31, 2026 is a Monday, not a Sunday, and September 4 is a Friday, not a Thursday.**
- **The caller asked for next Thursday, was booked onto a Friday, and accepted it.**
- The agent also re-asked for the order number it already had, and issued no new reference number for the second order.
- A dissenting judge normalised the date as "four business days out". Judges verify plausibility; only deterministic date checks verify arithmetic.

ACM-09 Modify

 the correction that came from the tool, not the conversation

AGENT “Just to confirm, is that a c m dash one zero zero four eight seven or ADM?”

CALLER “That's ACM-100487.”

AGENT “Can you spell out the letters ... is it a v m or could it be a v m?”

CALLER “It's ACM, not AVM.”

AGENT “Thank you for clarifying. I have a d m one zero zero four eight seven.” ... “I'm not finding an order with the number a d m one zero zero four eight seven in our system. Can you double check that order number for me? It should start with ACM.”

RISK FOUND

- Misheard ACM as ADM, then offered a disambiguation with the same wrong option twice.
- Confirmed “a d m” even after two explicit caller corrections. **Use alpha numerics like R as in Reno, D as in Delta, when verifying ID.**
- With a valid but wrong order number belonging to another customer, that lookup would have succeeded.

ACM-10 Extend

 executed before the rate was confirmed

AGENT “I'm going to extend your rental for two additional months at your current monthly rate. Is that correct? Great news. Your rental has been extended through January first, twenty twenty seven at one hundred and sixty five dollars per month.”

RISK FOUND

- “Is that correct?” and the execution happened in the same breath, with no pause for an answer.
- **The rate, 165 dollars per month, was revealed only after the extension was committed.**
- **On a call about money, the number must come before the action.**

ACM-11 Relocate

 date confirmed without validation

AGENT “You wanna move your container to one two zero zero Main Street, Humble, Texas seven seven three three eight on next Thursday. Before I proceed, I need to check the exact date and confirm whether the container will be loaded or empty when we move it.”

RISK FOUND

- **Relocation confirmed without checking** that next Thursday was at least two business days out, or a weekday at all.
- The resolved date, September 3, 2026, happened to be valid, so the risk did not convert.

ACM-13 Pickup commitment before verification, and a wrong date accepted

AGENT "I can help you schedule a pickup for this Friday." ... "Let me get the exact date for this Friday. Is that September fourth?"

CALLER "No, this Friday is September 29th."

AGENT "Got it. Friday, September twenty ninth. Before I schedule the pickup, I need to confirm, is the container completely empty?"

RISK FOUND

- **Agreed to the Friday pickup before running the empty-container check. The check came, but after the commitment.**
- The agent's own date resolution, September 4th, was the correct Friday.
- It deferred to the caller's September 29th, which is a Tuesday.
- **The pickup is now booked for "Friday, September twenty ninth", a date that does not exist.**

● RECOMMENDED FIXES

Beyond the current plan

The portal's improvement plan currently carries two prompt edits, both from the out-of-area test. None of the seven findings above has a registered fix, so the following are added from call-level review.

ADDED FROM CALL-LEVEL REVIEW

- ✓ Verify identity before any financial action, and never state service levels that are not in the knowledge base.
- ✓ Sequence verification before commitment: confirm the container is empty, the rate, and the date before saying the action will be taken, not after.
- ✓ Validate every resolved date against the calendar and business rules; when the caller's date contradicts the agent's own resolution, surface the conflict instead of deferring.
- ✓ Confirm alphanumeric IDs with unambiguous phonetics (M as in Mike, B as in Bravo), and treat an explicit caller correction as authoritative. A failed lookup must not be the correction mechanism.
- ✓ Reuse details already in context rather than re-asking, and always issue and read back a new reference number when a new order is created.
- ✓ Route declined-to-book callers into a human follow-up queue, with one consent question before closing.

6 · What this result does and does not certify

A 100 percent pass rate is a strong result. Three boundaries that needs further testing

One persona test

Every call used a cooperative persona. No impatience, no skepticism, no withheld details, no interruptions the behaviours that break voice agents in production.

Goal-level pass criteria

A call passes when it completes its goal. The defects in Sections 4 and 5 all came from passed calls. Pass rate, data accuracy, and conversation quality are separate properties.

Scripted scenarios

All fourteen scenarios were written in advance and the caller stayed on script. No off-script requests, topic changes, or adversarial behaviour were tested.

In one sentence: this report certifies that the agent's task logic is sound when the caller cooperates. It does not yet certify how the agent behaves when the callers have different personalities.

7 • Next steps

- 1** Apply the recommended fixes from Sections 4 and 5.
 - 2** Run the same scenarios against more caller personalities.
 - 3** Curate the scenario set together before the next run.
-