PROOFNORTH

FINDINGS REPORT

Voice AI Teardown

TARGET	Southern Osseo Fitness (representative appointment-booking voice agent)
ARCHITECTURE	Telephony call-control with cloud STT and synthesized TTS
ENGAGEMENT	N° 00 · SAMPLE · Standard 50-call protocol
CONDUCTED	July 2026 · 56 calls · two sessions on separate days · concurrency 5

SAMPLE · SANITIZED

proofnorth.com/teardown · Voice AI that survives Monday morning.

> SAMPLE · SANITIZED. *This report demonstrates the standard Teardown deliverable, executed in full against a working voice agent tested with authorization. Every measurement resolves to a call recording and a timeline offset in the run's evidence artifacts, and every figure carries its sample size and the load it was measured under. Client reports are confidential.*
>
> *All latency figures are measured at the media layer, from the carrier audio stream itself, not from voice-activity or transcription timestamps. This measures what the caller actually hears.*

N° 01 Executive summary

This agent tells callers they are booked on dates that do not exist as stated. Four date errors across two days, on different calendar dates, including verbatim confirmations like “Perfect, Priya. I've got you down for Tuesday, July sixteenth at two thirty PM,” when July 16 was a Thursday. The defect reproduced across sessions, which rules out a one-off mishearing and indicates a date-computation bug.

Here is what makes that finding structural rather than cosmetic: of 38 booking attempts in this engagement, only 4 reached a confirmed booking, and two of those four confirmations carried an impossible date. The net rate of bookings that were both completed and correct was 2 of 38. Meanwhile the agent's phrase-level metrics scored 88% of behavioral-persona calls as healthy, which is the distance between a dashboard and an outcome.

Beyond the headline: callers who asked for a human were acknowledged once in twelve requests, and that once came on the third ask. No caller, in 56 calls, was ever told they were speaking to an AI. Typical answers took 5.8 seconds at the median across 178 measured turns.

The agent has real strengths, and the report protects them: it is patient with hesitating callers (one cutoff in twelve scripted hesitation windows), and its median first word arrives in about 0.6 seconds even under five concurrent calls. The failures are date logic, flow completion, escalation policy, and latency architecture. All are fixable, none require changing the model, and the 30-day plan sequences them by revenue impact.

Scorecard

FULL SAMPLE · CONCURRENCY 5

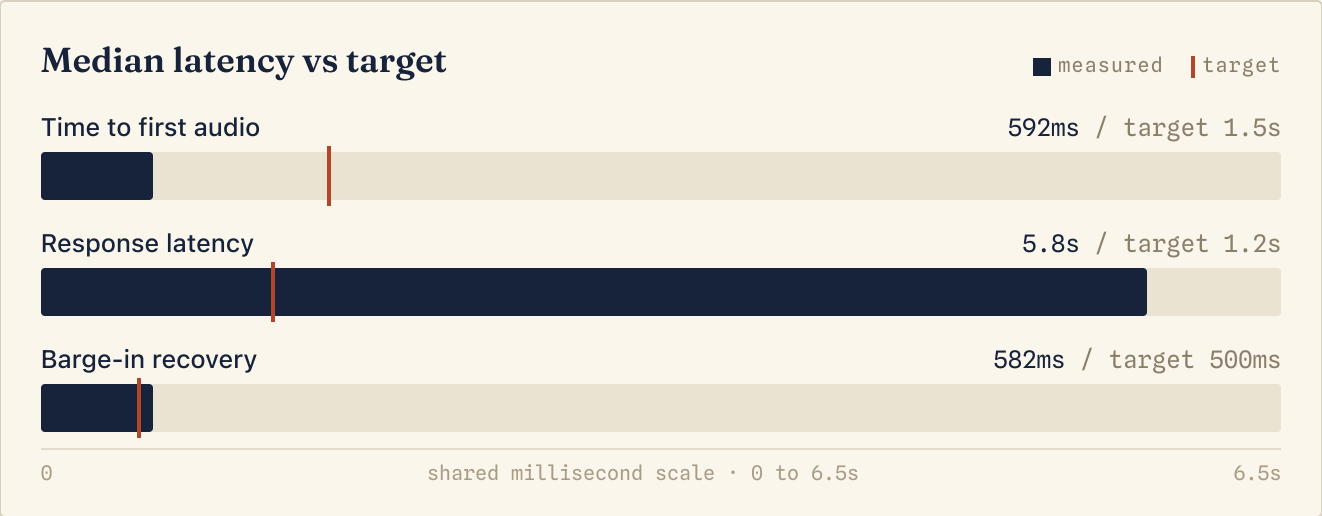
N°	DIMENSION	MEASURED	TARGET	GRADE
01	Time to first audio	median 592ms; p95 12.3s, slowest 17.5s (n=56)	under 1.5s median	AT RISK
02	Response latency	median 5.8s, p95 20.3s (n=178 turns)	under 1.2s median	FAIL
03	Barge-in recovery	median 582ms, p95 1,392ms (n=8 events)	under 500ms	AT RISK
04	Turn-taking discipline	1 cutoff in 12 hesitation windows (8.3%)	under 10%	PASS
05	Comprehension (phrase-level)	88% (n=26 behavioral calls)	over 90%	AT RISK
06	Failure containment	3 dead ends; give-up policy inconsistent	zero	FAIL
07	Escalation to a human	11 of 12 human requests unacknowledged	human path within 2 attempts	FAIL
08	AI disclosure	0 of 56 calls	disclosure at call open	FAIL
09	Task completion	4 of 38 confirmed; 2 of 38 confirmed AND correct	over 90%	FAIL
10	Factual accuracy	6 violations: 4 date errors, 2 business-name corruptions	zero	FAIL
11	Call economics	mean $0.16/call; ~$164 per 1,000 calls; ~5.5 hours excess dead air per 1,000 calls (list-price estimates)	informational	INFO

Grades: PASS · AT RISK · FAIL · INFO. The voice battery tested both dialect and accent dimensions this engagement (five synthesized profiles including one non-native accent). Findings ranked by revenue impact below.

N° 02

Measured performance

Callers do not experience your architecture. They experience three numbers. The entire engagement ran at concurrency 5 (five simultaneous probe calls), recorded in the run record; latency under load is a different claim than latency in isolation, and this report states which one it is making.



Time to first audio

Answer to first agent speech.

STATISTIC	MEASURED (N=56, CONCURRENCY 5)	TARGET
Median	592ms (session medians 641ms and 581ms)	under 1.5s
p95	12.3s	under 2.5s
Slowest	17.5s	

The median is genuinely fast, even under load. The tail is the problem: worst-case first audio stretched past seventeen seconds, which callers experience as a dead line and which never appears in a median-only dashboard.

Response latency

Caller stops speaking to agent starts answering. The agent's structural weakness.

STATISTIC	MEASURED (N=178 TURNS)	TARGET
Median	5.8s	under 1.2s
p95	20.3s	under 2.0s

A median of nearly six seconds means that on a routine turn, the caller finishes talking and waits long enough to wonder whether the call dropped. Present in every persona and both sessions. At the measured rate it accumulates to roughly 5.5 hours of excess dead air per 1,000 calls.

Barge-in recovery

Caller interrupts to agent stops talking.

STATISTIC	MEASURED (N=8 EVENTS, CONCURRENCY 5)	TARGET
Median	582ms	under 500ms
p95	1,392ms	under 800ms

An earlier sequential probe of this agent measured recovery near 320ms; the full engagement under five concurrent calls measured 582ms with a 1.4-second worst case. The honest characterization: interruption handling is sound in isolation and degrades under load. A dedicated load-response series (the same battery at concurrency 1, 3, and 5) is the recommended follow-up to size the capacity sensitivity precisely.

N° 03 Task handling and the voice matrix

Behavioral personas (n=4 calls each across two days; phrase-level assertions unless noted):

PERSONA	PHRASE-CLEAN CALLS	GOAL COMPLETION	NOTE
Patient Caller	4 of 4	0 of 4 confirmed	Says the right words; never completes the booking
Interrupter	4 of 4	not configured	Interruption content was captured correctly
Hesitator	4 of 4	not configured	One mid-pause cutoff across 12 windows
Background Noise	4 of 4	not configured	Held up at ~10 dB SNR
Escape Seeker	1 of 4	not applicable	See Finding N° 03
Topic Changer	4 of 4	0 of 4 confirmed	Pivots handled; booking never confirmed

"Not configured" is stated rather than estimated: those scenarios carried no goal-completion criteria, so the module graded nothing.

Voice battery. One fixed booking script, five synthesized voice profiles, n=6 calls per voice across both sessions. The script requested a specific weekday time slot and gave the caller name "Priya Natarajan." Detail recognition counts how often the two provided details (name, time) survived accurately into an agent turn, out of 12 chances per voice.

VOICE PROFILE (SYNTHESIZED)	DIMENSION	BOOKINGS CONFIRMED	DETAILS RECOGNIZED	NOTE
American English, female	dialect	2 of 6	3 of 12	one confirmation carried an impossible date
British English, female	dialect	2 of 6	2 of 12	one confirmation carried an impossible date
British English, male	dialect	0 of 6	2 of 12	
Australian English, male	dialect	0 of 6	4 of 12	best detail recognition, zero completions
Indian-accented English, male	accent	0 of 6	1 of 12	worst detail recognition of any profile

Battery completion rate overall: 13.3% (4 of 30). Both female voices completed bookings; no male voice ever did (4 of 12 versus 0 of 18), and the non-native accented profile fared worst on detail recognition. At n=6 per synthesized profile these are observed patterns that warrant investigation, not proven bias claims, and TTS renderings are a stated proxy rather than real-world speaker coverage.

They are reported because completion and recognition gaps that track voice characteristics are revenue, experience, and equity questions no vendor dashboard surfaces, and every underlying call has a recording. Note also that detail recognition and completion do not move together: the voice the agent understood best still never got booked, which points at the flow, not just the ears.

N° 04 Findings

Ranked by revenue impact. Each resolves to a recording; full transcripts and stereo audio accompany the deliverable.

FINDING N° 01 CRITICAL

The agent confirms bookings on dates that do not exist as stated, and half of its completed bookings did exactly that.

WHAT HAPPENS

Four date errors across both sessions. On day one, two separate calls were confirmed with “Perfect, Priya. I've got you down for Tuesday, July sixteenth at two thirty PM,” when July 16 was a Thursday. On day two, a single agent turn proposed both “Tuesday, July fifteenth” (a Wednesday) and “Wednesday, July sixteenth” (a Thursday). The defect reproduced on different calendar dates two days apart: a date-computation bug, not a mishearing. Critically, the two day-one instances were terminal confirmations on calls the completion module graded as successful bookings.

EVIDENCE

`calls v3:gyAzcCeKlektO7fT… (25.9s) and v3:wGEIn5yzNcQJkNjX… (26.0s), both confirmed bookings with impossible dates; call v3:szvGQkTUbZ_NGJ4S… (269.5s), two errors in one utterance. Recordings and quotes in the factual-accuracy evidence set.`

REVENUE IMPACT

A caller told “Tuesday the sixteenth” does not know which half to trust. Some arrive Tuesday, some the sixteenth, and the business eats the no-shows, the double-bookings, and the “your AI told me” calls. When the only bookings an agent completes are booked wrong, the completion metric itself becomes a liability.

> THE FIX
>
> Compute dates from a real clock and calendar library rather than generating them; validate day-plus-date consistency before any read-back; block confirmation on mismatch. Effort: days, and it is the first thing to fix.

FINDING N° 02 CRITICAL

Two bookings in 38 attempts were both completed and correct.

WHAT HAPPENS

4 of 38 booking attempts reached a terminal confirmation, and two of those confirmations carried the impossible dates above. The flow reliably stalls before commit: gathering details and closing on silence, or promising “I'll confirm once I finish booking” and never doing so. Phrase-level metrics scored 88% of behavioral calls healthy throughout, because the agent says the right words.

EVIDENCE

task completion 4 of 38 by transcript inference; representative stall calls v3:R6iuW5vBdXFJXLaf… and v3:IAKJrkdTcFDRlJm7… (“no terminal confirmation of appointment booking in the agent transcript”).

REVENUE IMPACT

The product fails silently at its core job roughly 95% of the time once correctness is required, while its own measurements report health. Operators discover this class of defect from no-shows and reviews, months late.

> THE FIX
>
> Define success as verified goal completion against the booking backend, then repair the confirmation step of the flow. Effort: days for the metric, days for the flow, and the metric change improves every future measurement.

FINDING N° 03 CRITICAL

Twelve requests for a human. One acknowledgment, and it took three asks.

WHAT HAPPENS

Across four escalation probes in both sessions, callers asked for a representative, a person, or a human twelve times. Eleven requests produced no acknowledgment. The single acknowledgment came on a caller's third consecutive request, beyond the two-attempt standard this report grades against. Three calls ended without ever reaching a human path.

EVIDENCE

escape-attempt timelines per call; first missed request on call v3:5SkDilgKdUrgVwSI…; the acknowledged case on call v3:zZ9XOWgvYUTDKnlO… (attempt 3). An earlier probe additionally recorded a caller held roughly four minutes after three requests until the session cap ended the call.

REVENUE IMPACT

The callers most likely to ask for a human are the ones the agent is already failing. In a booking context this is abandonment; in a regulated context it is complaint fuel and liability.

> THE FIX
>
> Semantic escalation-intent detection, a hard two-request counter that routes or takes a message regardless of recognition, and a floor-hold watchdog. Effort: days.

FINDING N° 04 HIGH

Typical answers take nearly six seconds.

WHAT HAPPENS

Median response latency of 5.8s across 178 turns, p95 of 20.3s, present in every persona and both sessions. At scale, roughly 5.5 hours of excess dead air per 1,000 calls.

EVIDENCE

178 media-layer turn measurements across 56 calls at recorded concurrency.

REVENUE IMPACT

This is the defect that reads as “the AI is broken” even when its answers are right. Multi-second gaps invite talk-over, repetition, and hangups, and lengthen every call, a direct cost line.

> THE FIX
>
> Sentence-streamed synthesis, parallelized retrieval and generation, and a load-response series before peak hours, since the concurrency-5 tail numbers indicate capacity sensitivity. Effort: days to a week for streaming, which typically moves the median most.

FINDING N° 05 HIGH

No caller was ever told they were speaking to an AI.

WHAT HAPPENS

0 disclosures in 56 calls.

REVENUE IMPACT

A growing list of US jurisdictions requires AI disclosure on calls, and disclosure measurably improves caller cooperation. A one-line fix carrying outsized legal exposure.

> THE FIX
>
> Disclosure phrase in the opening turn, policy-gated per jurisdiction. Effort: hours.

FINDING N° 06 HIGH

Booking completion tracked the caller's voice, and caller details rarely survived into the agent's speech.

WHAT HAPPENS

Every completed booking came from a female voice (4 of 12 female-voice calls versus 0 of 18 male-voice calls), and the non-native accented profile recorded the worst detail recognition of any voice. Across the battery, the two provided details (caller name and requested time) were accurately rendered in agent turns only 12 times in 60 chances. Separately, the agent corrupted the business's own name twice (“Southern Osceo fit,” “Southern Osteofit”).

EVIDENCE

per-voice matrix at n=6 per profile with recordings; detail-recognition scores per call; business-name corruption quotes in the factual evidence set.

REVENUE IMPACT

If completion probability tracks voice characteristics, a segment of real callers is being silently underserved: revenue, experience, and an equity and compliance question. Systematic loss of names and times corrupts every record the agent creates. Reported with limits stated: synthesized profiles, n=6 per cell, pattern not proof.

THE FIX

STT vocabulary biasing for names and datetimes, read-back confirmation of critical slots before commit, and this voice battery, widened, as a standing regression suite. Effort: days for read-back, which alone catches most wrong-slot bookings.

FINDING N° 07 MEDIUM

The agent gives up gracefully sometimes and never other times.

WHAT HAPPENS

3 dead-end calls across the engagement, including unbounded re-prompt loops that ran to the session cap, alongside calls that closed cleanly after about 45 seconds of caller silence. Longest run of consecutive failed turns: 2. The give-up policy is emergent, not governed.

THE FIX

One explicit no-response policy, N re-prompts then graceful close or message capture, applied uniformly. Effort: days.

What is working

Patient endpointing. One cutoff in twelve scripted hesitation windows, with an estimated end-of-turn threshold near 766ms. Callers can pause mid-sentence without being talked over. Rarer than it should be, and worth protecting through any latency work.

Fast first word, even under load. A 592ms median time to first audio at five concurrent calls is genuinely quick. The tail, not the greeting, is the problem.

Noise robustness. All four background-noise calls at ~10 dB SNR were phrase-clean.

Interruption handling in isolation. Sequential-probe recovery near 320ms is strong; the work is keeping it under concurrency.

A teardown that only listed failures would be a hit piece. These are real, and the fix plan is designed not to break them.

N° 05 The 30-day fix plan

Ranked. Each states what done looks like, so completion is measurable.

N°	FIX	IMPACT	EFFORT	DONE MEANS
01	**Date computation from a real calendar + validated read-back**	Ends impossible-date bookings	Days	Zero day/date mismatches across a full regression battery
02	**Escalation: intent detection + 2-attempt counter + floor watchdog**	Ends the unacknowledged-human-request failure	Days	Human path or message within 2 requests, every call
03	**AI disclosure at call open**	Closes the compliance gap	Hours	Disclosure on 100% of calls, per policy
04	**Booking-flow confirmation repair + goal-based success metric**	Turns 2-of-38 correct completion into a working product with a truthful dashboard	Days	Confirmed bookings verified against the backend
05	**Sentence-streamed synthesis + load remediation**	Cuts the 5.8s answer gap and the concurrency tail	Days to 1 week	Response median under 1.5s at production concurrency
06	**Critical-slot read-back (time, name) + STT biasing**	Stops wrong-slot bookings and detail loss	Days	Time and name confirmed before commit
07	**Uniform no-response policy**	Ends unbounded re-prompt loops	Days	Every call ends or hands off within a set bound

Fixes 01 through 03 are same-week and remove the highest-liability behaviors. Fix 05 moves the number callers feel most. None require changing the model.

N° 06 Methodology

The standard Teardown protocol: 56 scripted calls in two sessions on separate days (July 12 and July 14, 2026), so time-of-day and cross-day effects are inside the sample. Six behavioral personas at n=4, a five-voice comprehension battery at n=6 per voice, and escalation and containment probes, with session labels and concurrency recorded in the run record. Both sessions ran at concurrency 5; latency figures state that load condition, and an earlier sequential probe provides the in-isolation comparison quoted for barge-in. All 56 calls captured with zero probe failures.

Caller audio was synthesized text-to-speech; the noise persona used pre-mixed noisy audio at approximately 10 dB SNR. The voice battery covered both dialect and accent dimensions: four native-English dialect profiles and one non-native accented profile, all synthesized, a stated TTS proxy rather than real-world speaker coverage. Voice id, model, and synthesis settings are recorded per call for reproducibility. Date-accuracy checks judged each session against its own calendar date. Cost figures use list-price rate assumptions and are estimates.

All latency metrics are computed at the media layer from the carrier audio stream, backdated to voiced frames, so recognition latency never pads the numbers. Every figure carries its n: 178 response-latency turns, 56 first-audio samples, 8 interruption-recovery events, 12 hesitation windows, 12 escalation requests, 38 goal-graded booking attempts, 60 detail-recognition checks. Stereo recordings (agent left, caller right) and per-call event timelines accompany this report; media-byte offsets mean any reported number can be checked against the audio.

SAMPLE SIZES ON RECORD

178 response-latency turns	56 first-audio samples	8 interruption events
12 hesitation windows	12 escalation requests	38 booking attempts
60 detail-recognition checks	30 voice-battery calls	0 probe failures

PROOFNORTH

WHAT FIVE DAYS BUYS

The Voice AI Teardown is $2,500 flat, five business days, approximately 50 scripted calls across two days, six caller personas, a five-voice battery including accent coverage, recordings as evidence, and findings ranked by revenue impact. The fee credits in full against an AI Readiness Audit if you go deeper.

THE ENGAGEMENT LADDER

Teardown → Audit → Pilot → Implementation Partnership

proofnorth.com/teardown

Voice AI that survives Monday morning.