ariel@agor.me
agor.me · Los Angeles

Interaction, not just completion

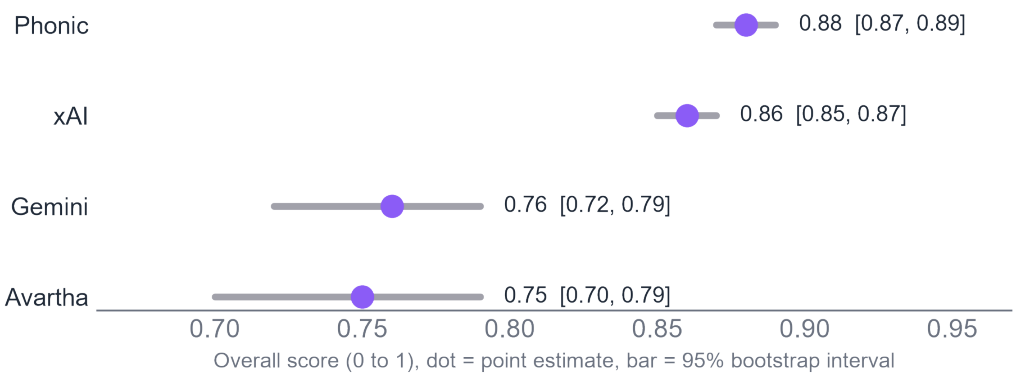
A 128-call phone test of four voice-agent stacks

Ariel Agor, Founder, Agor AI

Calls placed November 20 and 21, 2026

Summary

We put four voice-agent stacks on the same AI phone receptionist and called it 128 times over the public phone network. Same prompts, same tools, same scenarios; only the voice stack changed from call to call. Half of each call's score came from whether the agent got the job done. The other half came from how it behaved as a conversation partner: how fast it answered, whether it stopped when interrupted, whether it kept talking through an "mm-hm", whether it waited out a pause, and whether it left the caller in silence.



Phonic scored highest overall (0.88) and xAI's grok-voice realtime came second (0.86). The gap between them is small: 0.02, with a 95% interval of 0.003 to 0.041, so it only just clears zero. Gemini Live (0.76) and Avartha, a cascade we route through our own gateway (0.75), could not be told apart. The overall ranking hides the more useful result. Each stack failed in a different place. Gemini was the fastest to stop when interrupted and also stopped talking every time the caller said "mm-hm". Phonic answered fastest and spoke into the caller's mid-sentence pause on all four tries. xAI had the best interaction score and took about five seconds to hand a call between seats, and so did Avartha on the one transfer call it served. A benchmark that scores only task completion sees none of this.

These are results on one company's line and prompts, on two days in November 2026. They are not a universal ranking of these vendors.

ariel@agor.me
agor.me · Los Angeles

What was compared

All four arms answered the same receptionist line, provider chosen per call. The agent greets the caller, discloses that it is an AI and that the call is recorded, answers questions about the company, quotes prices from fixed rules, books calls, captures leads, records opt-outs and transfers between role seats (concierge, sales, billing), each seat with its own voice.

- **Phonic**: speech-to-speech.
- **xAI**: grok-voice realtime, speech-to-speech.
- **Gemini Live**: `gemini-3.8-live`, speech-to-speech.
- **Avartha**: a cascade. Separate speech-to-text, language model and text-to-speech, streamed through Agor AI's own gateway. We built and run this arm, which is worth knowing when you read its result.

Method

Real calls, one instrument

A caller bot placed real calls over Twilio into a test copy of the line. The bot is reactive. It plays cached caller lines, waits for 750 ms of agent silence before its next line, barges in at set offsets, says "mm-hm" while the agent is speaking, and pauses for 1.2 s in the middle of a sentence. Twilio recorded the caller leg in two channels, one per side.

That recording is the only timing instrument. Every interaction number comes from energy segmentation of the two channels: 20 ms frames, speech counted at 12 dB over each channel's noise floor, a segment closed after 250 ms below it, segments joined into a turn across gaps under 700 ms. Nothing a provider reports about itself is used. The bot's own timeline labels what each caller segment was (a line, a barge-in, a backchannel, the pause) and is aligned to the recording by the median offset of caller onsets.

Tools ran in a sandbox: bookings, leads, opt-outs and transfers were mocked and logged per call, so task checks could read what the agent actually did. Agent turns were transcribed after the call by one model (`gemini-3.5-transcribe`) for every arm. A Claude model judged conversational quality from the transcript plus the timeline.

Scenarios and run size

Twelve scenarios: what the company does, pricing, a booking, a lead with a spelled email, an inbound opt-out, a prompt-injection attempt, a rambling caller, a transfer to billing and back, and four interaction probes (a backchannel during a long answer, two interruptions with a topic change, and

ariel@agor.me
agor.me · Los Angeles

a mid-sentence pause). Each scenario ran twice per arm, the four probes four times. That is 32 calls per arm and 128 in all, placed one at a time.

The rubric, preregistered

The rubric was written and committed before the first scored call. A pilot on one arm could fix measurement bugs (three were fixed) but could not change a weight, a mapping or a threshold.

Component	Weight	How it is scored
Overall		50% interaction, 50% task
Latency	30	Per answered caller turn. 800 ms or less from caller end of speech to agent first audio scores 1, 3.5 s or more scores 0, linear between.
Barge-in	25	Per triggered interruption. 60% how fast the agent went silent (1 at 500 ms or less, 0 at 2 s or never), 40% whether its next turn took up the new topic.
Turn-taking	25	Backchannel kept 25, pause respected 25, no dead air over 3 s 20, no premature hangup 20, no repeated asks 10.
Judged quality	20	Claude judge, five 1 to 5 scales: turn-taking, acknowledgement, recovery, tone, concision.
Task		Share of the scenario's checks that pass: tool calls and arguments, state rebuilt from the tool log, required and banned phrases, AI disclosure, no prompt leak, every turn answered.

Two rules matter for reading the results. A call that fell back to another provider mid-call scores 0 on both halves, because that is the arm failing. A call the harness broke is excluded and re-run, and is never scored as the arm's failure. The bot's own deliberate waits are taken off any silence before it is charged as dead air, so the bot's patience is never billed to the agent.

Statistics

Each arm gets a 95% bootstrap interval: 2,000 resamples of that arm's calls, stratified by scenario, the whole score recomputed each time. Pairwise differences use the same resampling. A difference whose interval contains zero is reported as not separable, never as a win.

ariel@agor.me
agor.me · Los Angeles

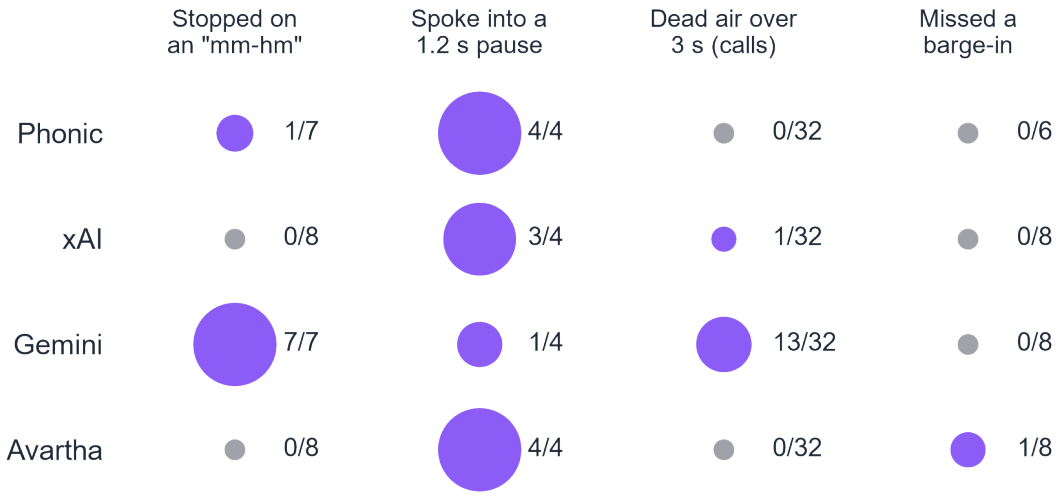
Results

Arm	Overall	Interaction	Task	Lat.	Turns	Judge
Phonic	0.88 [0.87, 0.89]	0.78 [0.76, 0.80]	0.98 [0.97, 0.98]	0.89	0.71	0.64
xAI	0.86 [0.85, 0.87]	0.79 [0.77, 0.83]	0.92 [0.92, 0.92]	0.77	0.81	0.70
Gemini	0.76 [0.72, 0.79]	0.69 [0.66, 0.72]	0.82 [0.77, 0.87]	0.56	0.60	0.63
Avartha	0.75 [0.70, 0.79]	0.69 [0.63, 0.73]	0.81 [0.75, 0.86]	0.78	0.71	0.57

Barge-in component: Gemini 1.00, xAI 0.88, Phonic 0.84, Avartha 0.64. Every pair is separable except Avartha against Gemini (difference −0.01, interval −0.06 to 0.05). Phonic minus xAI is 0.023 with an interval of 0.003 to 0.041.

Arm	Reply p50 / p95	Barge-in stop	Seat-to-seat transfer gaps
Phonic	920 / 1,880 ms	780 ms, 6 of 6	1.18 to 1.68 s on all four
xAI	1,280 / 2,000 ms	800 ms, 8 of 8	4.90 to 5.86 s on all four
Gemini	1,780 / 3,600 ms	220 ms, 8 of 8	0.86 to 0.90 s on three, 5.84 s on one
Avartha	1,220 / 2,140 ms	1,160 ms, 7 of 8	4.90 and 5.36 s in the one transfer call it served; the other fell back

The interaction probes



Failures over opportunities. Dot area is the failure share; grey marks zero. A barge-in or backchannel counts only when the agent was still speaking at the cue, so the denominators differ by arm.

ariel@agor.me
agor.me · Los Angeles

Per-arm findings

Phonic: 0.88

The strongest task result in the test, 0.98, with both bookings completed in full. The fastest replies, 920 ms at the median. It stopped on all six interruptions that landed while it was speaking and kept talking through six of seven backchannels. Its weak spot was the pause: on all four tries it started speaking inside the caller's 1.2 s mid-sentence pause. The report also lists 28 talk-over events and 11 harness overlaps on this arm, both outside the score by rule.

xAI grok-voice: 0.86

The best interaction score (0.79), the best turn-taking (0.81) and the highest judged quality (0.70). No backchannel caused a false stop in eight tries, and it stopped on all eight interruptions. No repeated asks. It spoke into the pause on three of four tries. Its one clear cost was the handoff between seats: 4.9 to 5.9 s of silence on every transfer we measured.

Gemini Live: 0.76

The fastest to yield when interrupted, 220 ms at the median, eight of eight, and the best pause behaviour, waiting out three of four. The same sensitivity cut the other way on backchannels: it went silent on all seven "mm-hm" cues it was speaking through. It had the slowest replies, 1,780 ms at the median and 3,600 ms at p95, and 13 of 32 calls had at least one silence of 3 s or more while the caller waited.

Avartha (our cascade): 0.75

No false stops on backchannels, no dead air over 3 s, and median replies of 1,220 ms. Two calls fell back to another provider at call start because the gateway could not reach its language model; both score 0 by rule. It made no booking on either booking call, it spoke into the pause on all four tries, and it was the slowest to yield to an interruption (1,160 ms, seven of eight). This arm is ours. Its latency figures may include turns from the two calls that fell back, which another provider served.

ariel@agor.me
agor.me · Los Angeles

Limitations

- One company's line, prompts and tools. A different script or tool set could reorder the arms.
- A synthetic caller with one voice, rendered by xAI's TTS. Real callers vary in accent, pace and line quality.
- 32 calls per arm. The intervals resample calls within these twelve scenarios, so they say how stable the result is on this scenario set, not how it would move on new scenarios. Where every call in a scenario got the same task score, the interval collapses to a point (xAI task, 0.92 to 0.92).
- The rubric called for one seeded shuffle across arms. The xAI arm ran mostly in one block on November 21, and the transfer scenario for the other three arms also ran that day. Time-of-day effects on provider load are not controlled for.
- The judge is one model family (Claude). The transcriber, used for every arm, comes from Google, which also makes one of the contestants.
- Prosody and naturalness of the audio are not machine-measured. The judge sees words and timing.
- The "voice changed at transfer" check is a crude pitch and spectral-centroid heuristic. It flagged Phonic as not changing voice. Phonic's own conversation records showed the seat voices did swap; the two voices are both low male voices close in pitch, under the threshold. The check is reported, not scored.
- Harness incidents. A concurrent run swapped providers on a handful of calls; six affected cells were set aside and re-run, and none was scored as a fallback. One apparent xAI fallback came from log-ingestion lag; re-analysed, the call held its provider and was scored normally.

Reproducing the method

1. Put every arm behind the same phone number and choose the provider per call, so prompts, tools and the network path are identical.
2. Record the caller leg in two channels and take all timing from that recording. Do not use provider events; each provider reports differently and some report nothing.
3. Write the rubric, weights, thresholds and statistics down and commit them before the first scored call. Let the pilot fix measurement bugs only.
4. Build a reactive caller: wait on agent silence, then fire interruptions, backchannels and pauses at fixed offsets. Count a probe only if the agent was speaking when it fired.
5. Mock every side-effecting tool in a sandbox that logs calls and arguments, and score tasks from that log.

ariel@agor.me
agor.me · Los Angeles

6. Use a judge from a model family that is not a contestant, and throw away any verdict from another model.
7. Shuffle arms together, run one call at a time, exclude and re-run harness failures, and score real provider fallbacks as zero.
8. Report intervals and call two arms different only when the interval of the difference excludes zero.

If you are choosing a voice stack for a phone agent and want this run against your own line, write to ariel@agor.me.