Field latency of voice agents

A benchmark of four stacks at three public events, with a stage-by-stage breakdown

Cris Olmedo and Kurt Sommer
Divina Machina Research Group
Working paper · Benchmark
DM-2026-03, version 1.0 · 30 October 2026

Abstract. Latency figures for voice agents are almost always published from the lab, with different metrics and without uncertainty. This work contributes a field benchmark. The same character, GAIA 2147, answered calls in a telephone booth at three public events on four voice stacks, two cascades and two speech-to-speech models. We analyse 3,357 response turns in 530 conversations with the platform's per-turn metric, the time to first audio since detected silence.

The median was 2.68 s with a Qwen 3.5 cascade, 3.92 s with a Gemini 3.7 Flash cascade, 0.96 s with Eleven Realtime v1 Max and 0.81 s with v1 Mini, with conversation-bootstrap intervals. The fastest stack was 3.3× faster than the main cascade. The breakdown attributes almost all of the difference to the language model's first token and to the configured end-of-turn wait. In three of the four stacks, the first response of each call was slower than the ones that followed. With the cascade, waiting took up 16 % of the median call.

We propose a reference table that separates what each published figure measures, and a minimum reporting standard for field latency. The data are observational and the stacks change together with the event.

Keywords: voice agents; latency; turn-taking; field benchmark; speech-to-speech; cluster bootstrap.

1. Introduction

On the telephone, silence is not neutral. A pause sounds like a broken line or like nobody is there. In conversation between people the typical answer arrives within a few hundred milliseconds [1, 2], and a silence of around one second starts to mean something in itself [3]. No voice agent answers that fast in a noisy room today. The design question is how far a system departs from that reference, where the time is lost and which part can be recovered.

The available figures do not answer that question. Vendors publish inference or synthesis times [4, 5]; independent evaluators measure time to the first audio token on a question bank [6]; technical guides propose design targets [7]. These are different metrics, with different start and end events, and they almost never carry uncertainty. None comes from real calls with the public.

1.1. Contributions

1. A field benchmark of four voice stacks under the same character, with medians, 90th percentile and tails, all with conversation-bootstrap confidence intervals.
2. A stage-by-stage breakdown that locates where time is gained.
3. Two design effects the lab does not see: the first-turn penalty and the share of the call spent waiting.
4. A comparison table that separates what each published figure measures, and a minimum reporting standard.

2. Background

2.1. The human gap

Stivers and colleagues measured answers to yes/no questions in ten languages and 101 conversations, with a pooled mean of 208 ms, a median of 100 ms and a mode of 0 ms. The answer could be vocal or gestural and measurements were rounded to 100 ms [1]. A reading of those data places the typical answer between 0 and 300 ms [2]. Jefferson proposed a standard maximum silence of about one second [3]. Levinson and Torreira explain that this speed coexists with language production times of 600 ms or more, which requires anticipating the end of the other's turn [8]. Templeton and colleagues observed that faster responses are associated with greater perceived connection [9]. None of these works sets a target for machines; they describe a human distribution under their conditions [EMP].

2.2. Vendor and evaluator figures

A cascade chains speech recognition, a language model and synthesis. A speech-to-speech model integrates those stages in a jointly tuned system [10]. Artificial Analysis publishes mean time to the first audio token (TTFA) on Big Bench Audio for different configurations [6, 11] [EVA]. ElevenLabs states inference latencies for its synthesis models [4, 5] [FAB]. Twilio proposes targets for a simple cascade [7] [FAB]. The companion paper DM-2026-01 explains why audio availability, release and audible onset are distinct events [12].

3. Data and method

3.1. Setting

GAIA 2147 is a character who answers in a telephone booth. It always runs on ElevenLabs and Vidiv; this analysis uses the ElevenLabs agent logs. The gaia-2147

Table 1. Voice stacks by event. The character is the same; model, event and audience change.

Event (2026)	Stack
Sónar+D, Barcelona, June	Eleven v3 Conversational; no end-to-end metric
IFA, Berlin, September	Cascade: Qwen 3.5 397B or Gemini 3.7 Flash, with Eleven v3 Conversational
Barcelona Design Week, October	Speech-to-speech: Eleven Realtime v1 Max or Mini, with Eleven v4 Turbo

dataset holds 1,527 conversations and 8,151 turns, described in DM-2026-09 [13]. Table 1 summarises the stacks.

3.2. Variable

The main variable is the per-turn metric `convai_ttf_audio_since_silence`: the time from the silence detected after the visitor's speech to the first audio of the response, measured on the server [DEF]. It excludes the network to the booth, the playback buffer and the loudspeaker, and it is not the perceived audible onset. Turn 0, the greeting, is excluded because it plays before anyone speaks. The recorded stages are the end-of-turn wait t_{fin}, recognition t_{asr}, the model's first token t_{llm} and the first voice byte t_{tts}:

$$T_i = t_{\text{fin},i} + t_{\text{asr},i} + t_{\text{llm},i} + t_{\text{tts},i} + \varepsilon_i. \tag{1}$$

ε_i collects what the log does not separate. Stages may overlap when streaming and medians are not additive, so stages are compared one by one and never summed.

3.3. Statistics

Turns within the same call are not independent. 95 % intervals for medians, percentiles and turn proportions use a cluster bootstrap [14, 15]: whole conversations are resampled $B = 2000$ times with a fixed seed, taking

$$\hat{\theta}^{(b)} = \theta\Big(\bigcup_{j \in S^{(b)}} \{T_{jk}\}_k\Big), \quad S^{(b)} \sim \mathrm{U}\{1..J\}^J, \tag{2}$$

with the 2.5th and 97.5th percentiles of $\hat{\theta}^{(b)}$. Contrasts between stacks use the median of each conversation with at least two turns, which is the independent unit: an overall Kruskal-Wallis test and pairwise Mann-Whitney tests with Holm correction [16], reported with Cliff's delta as effect size [17]. The dead time of a call is the share of its duration spent waiting for responses:

$$\delta_j = \min\Big(1,\ \tfrac{1}{D_j}\textstyle\sum_k T_{jk}\Big), \tag{3}$$

computed on calls of at least 20 s with three or more measured responses.

4. Results

4.1. Latency by stack

Table 2 and figure 1 summarise the four distributions [MED]. The fastest speech-to-speech stack, Realtime v1 Mini, had a median of 0.81 s (95 % CI 0.78–0.84 s) and 71 % of its turns arrived in under a second. The Qwen cascade had a median of 2.68 s (95 % CI 2.45–2.94 s); only 3 % of its turns came in under a second and 41 % took more than three. The improvement in the median between the two was 3.3× (95 % CI 3.1–3.6×).

Gemini 3.7 Flash was the slowest stack, with a median of 3.92 s (95 % CI 3.67–4.43 s). It served only 20 conversations, so its interval is the widest; its slowness comes from the model's first token (table 4).

4.2. Contrasts

Per-conversation medians differ between stacks (Kruskal-Wallis, $H = 201.1$, $p < 0.001$). Every pair differs after Holm correction (table 3). The smallest effect separates Max from Mini, with a Cliff's delta of 0.43: at the same event and with the same voice, Mini answered sooner in most comparisons between calls.

4.3. Where the time goes

Table 4 and figure 2 break the wait down. Recognition was never the bottleneck. The model's first token went from 1,515 ms with Qwen to 220 ms within Realtime Max. The end-of-turn wait is not a random variable but a configuration parameter: 320 ms at IFA and 160 ms in Barcelona [MED]. Subtracting that wait, the cascade's median is still 2.08 s against 0.59 s for Mini.

4.4. The first turn is slower

Latency is not constant along the call (figure 3). With Realtime Max, the first response of each call had a median of 1.10 s and the seventh and later ones 0.84 s (Spearman between position and latency, $\rho = -0.21$). With the Qwen cascade the first response took 3.15 s against 2.54 s for the following ones. With Realtime Mini the difference almost disappears, 0.83 s against 0.80 s. The first response is the visitor's first real impression, and in three of the four stacks it was the worst.

4.5. Dead time

Following equation (3), the Qwen cascade spent 16 % (95 % CI 15–18 %) of the median call waiting; Realtime Mini, 7 % (95 % CI 6–7 %) (table 5). One in every six seconds of the cascade was machine silence.

4.6. Sónar+D and cost

The Sónar+D logs do not include the end-to-end metric. Across 2,908 turns, the model's first token took 1,011 ms and the first voice byte 284 ms, in median. Cloud cost was 0.12 USD per minute with the IFA cascade and 0.10 USD with Realtime: the faster stack was not more expensive.

5. Comparative benchmark

Table 6 sets our figures beside the published ones. The decisive column is comparability. No published figure measures the same thing as ours, and ours does not measure the audible onset either.

Two readings hold. First, in the field Realtime Max and Mini ran a few hundred milliseconds above their

Table 2. Field benchmark. Time to first audio since silence, per response turn. 95 % intervals by conversation bootstrap.

Stack		Conv.	Turns	Median (s) [95 % CI]	P90 (s)	< 1 s	> 3 s
Qwen 3.5 397B + Eleven v3 Conv.	IFA	98	374	2.68 [2.45; 2.94]	5.0	3 %	41 %
Gemini 3.7 Flash + Eleven v3 Conv.	IFA	20	48	3.92 [3.67; 4.43]	5.8	0 %	88 %
Eleven Realtime v1 Max + v4 Turbo	BDW	331	2,336	0.96 [0.93; 1.00]	2.3	53 %	5 %
Eleven Realtime v1 Mini + v4 Turbo	BDW	81	599	0.81 [0.78; 0.84]	1.5	71 %	2 %

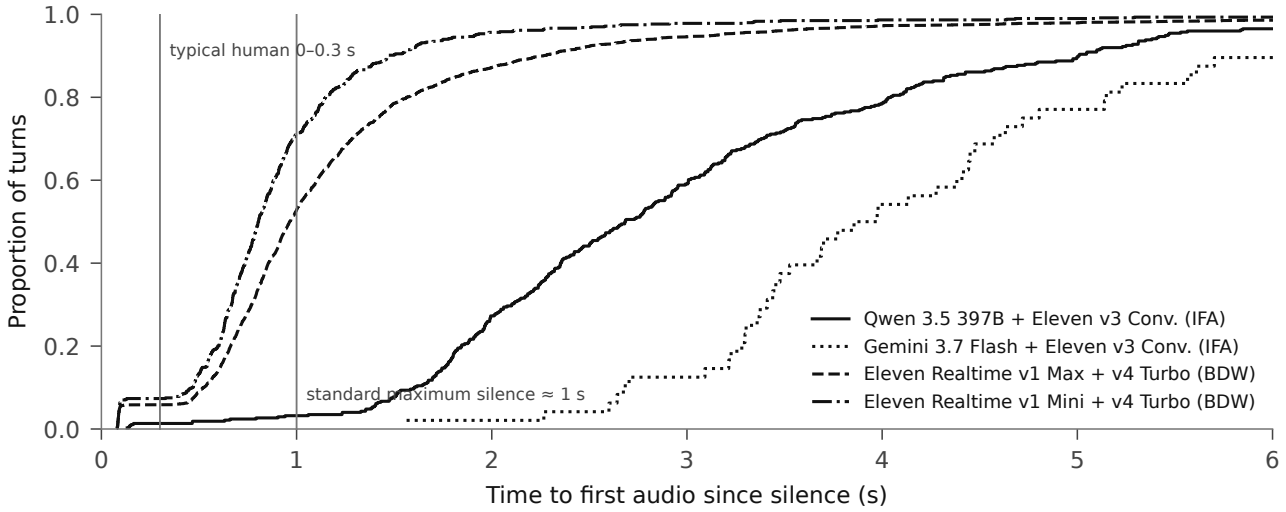


Figure 1. Cumulative distribution of per-turn latency for each stack. Vertical lines mark the typical human answer [2] and Jefferson's standard maximum silence [3], as references, not targets.

Table 3. Pairwise contrasts on each conversation's median. A positive Cliff's delta means the first stack is slower.

Comparison	n conv.	Cliff's δ	Holm p
Qwen vs Gemini	67 / 12	−0.62	<0.001
Qwen vs RT Max	67 / 305	0.93	<0.001
Qwen vs RT Mini	67 / 78	0.98	<0.001
Gemini vs RT Max	12 / 305	0.99	<0.001
Gemini vs RT Mini	12 / 78	1.00	<0.001
RT Max vs RT Mini	305 / 78	0.43	<0.001

Table 4. Median of each recorded stage, in milliseconds. Medians are not summed.

Stage (ms, median)	Qwen	Gemini	RT Max	RT Mini
End-of-turn wait	320	640	160	160
Recognition	56	42	59	47
LLM first token	1,515	2,851	220	181
TTS first byte	187	173	81	80

Table 5. Median share of the call spent waiting for responses.

Stack	Calls	Dead time [95 % CI]
Qwen	51	16 % [15; 18]
Gemini	8	20 % [19; 26]
RT Max	288	8 % [8; 9]
RT Mini	75	7 % [6; 7]

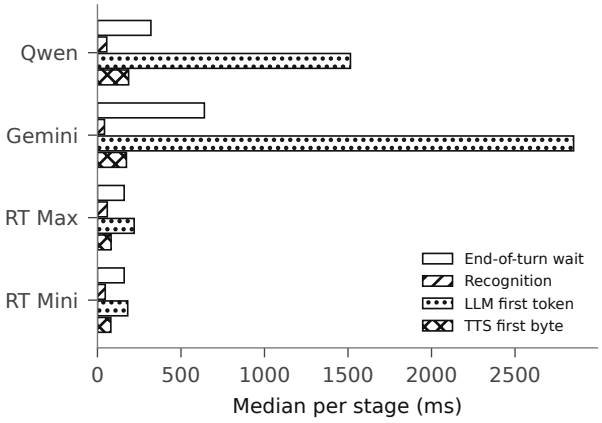


Figure 2. Median per stage and stack. The model's first token dominates in the cascades.

lab TTFA, partly because our metric includes the configured end-of-turn wait [INT]. Second, the evaluator's means and our medians are not the same statistic, and with long tails the mean exceeds the median; the comparison is descriptive and admits no test.

6. Discussion

The improvement between events came from two places: a model that starts speaking sooner and a system that decides sooner that it is its turn [INT]. Both are design decisions, not fixed properties of the hardware. The end-of-turn wait in particular is a chosen parameter; lowering it brings the answer forward and raises the risk of interrupting, a trade-off that DM-2026-01 formalises as a decision under losses [12].

Table 6. Reference and field figures. Each row keeps its metric, its statistic and its source. Tags indicate the type of evidence.

System or population	Figure	What it measures	Comparable with the field
People, ten languages [EMP]	208 ms mean; 100 ms median	End of question to vocal or gestural onset of the answer [1]	No: face to face, rounded to 100 ms
People, maximum silence [EMP]	≈1 s	Threshold beyond which silence carries meaning [3]	As an experience reference, not a system one
Eleven Realtime v1 Max [EVA]	570 ms mean	TTFA on Big Bench Audio [6, 18]	Partly: no end-of-turn wait, no room
Eleven Realtime v1 Mini [EVA]	620 ms mean	TTFA on Big Bench Audio [6, 18]	Partly
GPT-Realtime-2.1 High [EVA]	1.21 s mean	TTFA, High configuration [6]	Partly; not deployed here
GPT-Realtime-2.1 Minimal [EVA]	0.97 s mean	TTFA, Minimal configuration [6]	Partly; not deployed here
Eleven v4 Turbo [FAB]	≈100 ms median	Synthesis inference without network [4, 5]	No: a single stage
Cascade, design target [FAB]	885 ms; 1115 ms	Platform turn gap; mouth to ear [7]	Target, not measurement
GAIA, Realtime v1 Max [MED]	0.96 s median	Detected silence to first audio, server, real room	Field reference
GAIA, Realtime v1 Mini [MED]	0.81 s median	Same	Field reference
GAIA, Qwen cascade [MED]	2.68 s median	Same	Field reference
GAIA, Gemini cascade [MED]	3.92 s median	Same	Field reference

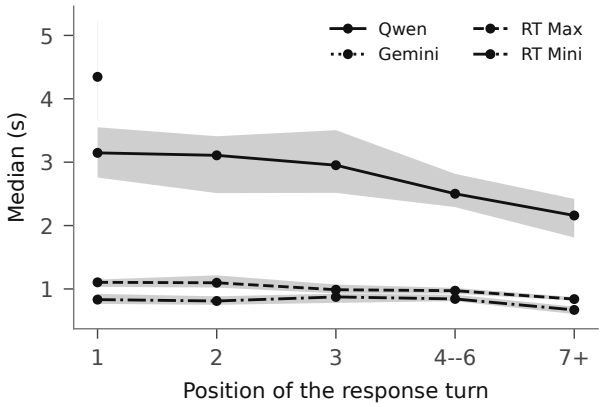


Figure 3. Median latency by position of the response within the call, with 95 % CI by turn bootstrap (grey band). Positions with fewer than 15 turns are omitted.

The first-turn penalty matters more than its size suggests. It is the moment when the visitor decides whether the voice is alive; DM-2026-04 studies that first second. Warming up the connection before pick-up, which we already do, is not enough for the first response turn [HIP].

Dead time translates latency into experience. A median of 2.7 s looks like a technical problem; 16 % of a call in silence is a dramaturgical one.

7. Threats to validity

- *Confounding.* Each stack belongs to one event, with its own network, audience, language and prompt version. Differences between stacks are associations [INT].
- *Measurement.* The metric is server-side. It excludes network and playback in the booth; the audible onset is later.
- *Selection.* Only turns with a recorded metric count; calls answered by Vidiv are not included.
- *Size.* Gemini served 20 conversations; its estimate is the least precise.
- *Content.* A difficult question may take longer; turn length and complexity are not controlled.

8. Recommendations for the community

We propose a minimum standard for reporting field latency of voice agents:

1. Name the start and end events, and state whether the end-of-turn wait is included.
2. Report median and 90th percentile, never the mean alone, with conversation-bootstrap intervals.
3. Break down by stage and state which stages are configured parameters.
4. Separate the first response turn from the rest.
5. Report the dead time per call, which is what the user lives through.
6. Publish the number of conversations, not only the number of turns.

9. Data and code

gaia-2147 dataset with its fingerprint in the manifest [13]. Every figure in this document is generated by `analysis/analyze.py` and `kit/python/dmpaper.py`, seed 2147. Conversation text was never stored.

10. Conclusions

In a real room, with a real audience, a speech-to-speech stack answered in a median of 0.81 s and a cascade in 2.68 s. The difference lay in the model and the end-of-turn wait, not in recognition. Lab figures are useful for choosing, not for promising: what counts is the silence heard by whoever is on the phone.

Signed. Cris Olmedo and Kurt Sommer, Divina Machina Research Group. *How to cite*: Olmedo, C. and Sommer, K. (2026). Field latency of voice agents. Working paper DM-2026-03, v 1.0. Divina Machina Research Group. Divina Machina by Qs Ventures.

References

Sources consulted up to 29 October 2026. Vendor and evaluator figures are not Divina Machina measurements.

[1] T. Stivers et al. Universals and cultural variation in turn-taking in conversation. *PNAS*, 106(26):10587–10592, 2009. doi: 10.1073/pnas.0903616106. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC2705608/.

[2] Language Log. Speaker-change offsets. Language Log, University of Pennsylvania, 2009. URL https://languagelog.ldc.upenn.edu/nll/?p=7997. A reading of the data of Stivers et al. (2009).

[3] G. Jefferson. Preliminary notes on a possible metric which provides for a standard maximum silence of approximately one second in conversation. In D. Roger and P. Bull, editors, *Conversation. An Interdisciplinary Perspective*, pages 166–196. Multilingual Matters, 1989. URL https://emcawiki.net/Jefferson1989a.

[4] ElevenLabs. Models. ElevenLabs documentation, n.d.. URL https://elevenlabs.io/docs/overview/models. Models, identifiers and inference latencies.

[5] M. Staniszewski and P. Dabkowski. Introducing Eleven v4: our most emotive model. ElevenLabs, September 2026. URL https://elevenlabs.io/blog/eleven-v4. Dated 28 September 2026, updated 19 October 2026. Methodology note 3 documents the 150 ms figure.

[6] Artificial Analysis. Speech to speech models and providers analysis. Artificial Analysis, n.d.. URL https://artificialanalysis.ai/speech-to-speech. Live results table; explicit model configurations.

[7] Twilio. A guide to core latency in AI voice agents. Twilio Blog, November 2025. URL https://www.twilio.com/en-us/blog/developers/best-practices/guide-core-latency-ai-voice-agents.

[8] S. C. Levinson and F. Torreira. Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, 6:731, 2015. doi: 10.3389/fpsyg.2015.00731. URL https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2015.00731/full.

[9] E. M. Templeton, L. J. Chang, E. A. Reynolds, M. D. Cone LeBeaumont, and T. Wheatley. Fast response times signal social connection in conversation. *PNAS*, 119(4):e2116915119, 2022. doi: 10.1073/pnas.2116915119. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC8794835/.

[10] ElevenLabs. Realtime models. ElevenLabs documentation, n.d.. URL https://elevenlabs.io/docs/eleven-agents/customization/realtime-models. Custom, Realtime, settings and gradual availability.

[11] Artificial Analysis. Speech to speech benchmarking methodology. Artificial Analysis, n.d.. URL https://artificialanalysis.ai/methodology/speech-to-speech-benchmarking. TTFA definition and Big Bench Audio description.

[12] C. Olmedo and K. Sommer. The art of conversation. from the Before trilogy to the design of conversational agents. Working paper DM-2026-01, v0.2. Divina Machina Research Group, 2026. Conceptual article; no own data.

[13] C. Olmedo and K. Sommer. GAIA-2147. data descriptor of 1527 conversations with a voice agent in public spaces. Working paper DM-2026-09. Divina Machina Research Group, 2026.

[14] B. Efron. Bootstrap methods. another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979. doi: 10.1214/aos/1176344552.

[15] C. A. Field and A. H. Welsh. Bootstrapping clustered data. *Journal of the Royal Statistical Society. Series B*, 69(3):369–390, 2007. doi: 10.1111/j.1467-9868.2007.00593.x.

[16] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[17] N. Cliff. Dominance statistics. ordinal analyses to answer ordinal questions. *Psychological Bulletin*, 114(3):494–509, 1993. doi: 10.1037/0033-2909.114.3.494.

[18] ElevenLabs. Eleven realtime. Product page, 2026. URL https://elevenlabs.io/realtime. Time-to-first-audio figures attributed to Artificial Analysis (Big Bench Audio).