Latencia de agentes de voz en campo

Benchmark de cuatro pilas en tres eventos públicos, con descomposición por etapas

Cris Olmedo y Kurt Sommer
Grupo de investigación Divina Machina
Documento de trabajo · Benchmark
DM-2026-03, versión 1.0 · 30 de octubre de 2026

Resumen. Las cifras de latencia de los agentes de voz se publican casi siempre en laboratorio, con métricas distintas y sin incertidumbre. Este trabajo aporta un benchmark de campo. Un mismo personaje, GAIA 2147, atendió llamadas en una cabina telefónica en tres eventos públicos con cuatro pilas de voz, dos en cascada y dos de voz a voz. Analizamos 3357 turnos de respuesta en 530 conversaciones con la métrica por turno de la plataforma, el tiempo hasta el primer audio desde el silencio detectado.

La mediana fue 2,68 s con Qwen 3.5 en cascada, 3,92 s con Gemini 3.7 Flash en cascada, 0,96 s con Eleven Realtime v1 Max y 0,81 s con v1 Mini, con intervalos por bootstrap de conversaciones. La pila más rápida fue 3,3× más rápida que la cascada principal. La descomposición atribuye casi toda la diferencia al primer token del modelo de lenguaje y a la espera configurada de fin de turno. En tres de las cuatro pilas, la primera respuesta de cada llamada fue más lenta que las siguientes. En la cascada, la espera ocupó el 16 % de la llamada mediana.

Proponemos una tabla de referencia que separa lo que mide cada cifra publicada y un estándar mínimo de reporte para latencias de campo. Los datos son observacionales y las pilas cambian junto con el evento.

Palabras clave: agentes de voz; latencia; toma de turnos; benchmark de campo; voz a voz; bootstrap por conglomerados.

1. Introducción

Al teléfono, el silencio no es neutro. Una pausa suena a línea rota o a que no hay nadie. En la conversación entre personas la respuesta típica llega en unos pocos cientos de milisegundos [1, 2], y un silencio de en torno a un segundo empieza a significar algo por sí mismo [3]. Ningún agente de voz responde hoy tan rápido en una sala ruidosa. La pregunta de diseño es cuánto se aleja un sistema de esa referencia, dónde se pierde el tiempo y qué parte se puede recuperar.

Las cifras disponibles no responden a esa pregunta. Los fabricantes publican tiempos de inferencia o de síntesis [4, 5]; los evaluadores independientes miden el tiempo hasta el primer token de audio en un banco de preguntas [6]; las guías técnicas proponen objetivos de diseño [7]. Son métricas distintas, con eventos de inicio y fin distintos, y casi nunca llevan incertidumbre. Ninguna procede de llamadas reales con público.

1.1. Contribuciones

1. Un benchmark de campo de cuatro pilas de voz bajo el mismo personaje, con medianas, percentil 90 y colas, todas con intervalos de confianza por bootstrap de conversaciones.
2. Una descomposición por etapas que localiza dónde se gana el tiempo.
3. Dos efectos de diseño que el laboratorio no ve: la penalización del primer turno y la fracción de la llamada que se va en esperar.
4. Una tabla comparativa que separa lo que mide cada cifra publicada y un estándar mínimo de reporte.

2. Antecedentes

2.1. El intervalo humano

Stivers y colaboradores midieron respuestas a preguntas de sí o no en diez lenguas y 101 conversaciones, con media conjunta de 208 ms, mediana de 100 ms y moda de 0 ms. La respuesta podía ser vocal o gestual y las medidas se redondearon a 100 ms [1]. Una lectura de esos datos sitúa la respuesta típica entre 0 y 300 ms [2]. Jefferson propuso un silencio máximo estándar de en torno a un segundo [3]. Levinson y Torreira explican que esa rapidez convive con tiempos de producción del lenguaje de 600 ms o más, lo que exige anticipar el final del turno ajeno [8]. Templeton y colaboradores observaron que las respuestas más rápidas se asocian a una conexión percibida mayor [9]. Ninguno de estos trabajos fija un objetivo para máquinas; describen una distribución humana en sus condiciones [EMP].

2.2. Cifras de fabricantes y evaluadores

Una cascada encadena reconocimiento de voz, modelo de lenguaje y síntesis. Un modelo de voz a voz integra esas etapas en un sistema ajustado en conjunto [10]. Artificial Analysis publica el tiempo medio hasta el primer token de audio (TTFA) sobre Big Bench Audio para distintas configuraciones [6, 11] [EVA]. ElevenLabs declara latencias de inferencia para sus modelos de síntesis [4, 5] [FAB]. Twilio propone objetivos para una cascada sencilla [7] [FAB]. El trabajo compañero DM-2026-01 desarrolla por qué disponibilidad del audio, liberación e inicio audible son acontecimientos distintos [12].

Tabla 1. Pilas de voz por evento. El personaje es el mismo; cambian el modelo, el evento y el público.

Evento (2026)	Pila
Sónar+D, Barcelona, junio	Eleven v3 Conversational; sin métrica de extremo a extremo
IFA, Berlín, septiembre	Cascada: Qwen 3.5 397B o Gemini 3.7 Flash, con Eleven v3 Conversational
Barcelona Design Week, octubre	Voz a voz: Eleven Realtime v1 Max o Mini, con Eleven v4 Turbo

3. Datos y método

3.1. Escenario

GAIA 2147 es un personaje que contesta en una cabina telefónica. Siempre corre sobre ElevenLabs y Vidiv; este análisis usa los registros del agente de ElevenLabs. El conjunto de datos, gaia-2147, reúne 1527 conversaciones y 8151 turnos, descritos en DM-2026-09 [13]. La tabla 1 resume las pilas.

3.2. Variable

La variable principal es la métrica por turno `convai_ttf_audio_since_silence`: el tiempo desde el silencio detectado tras la voz del visitante hasta el primer audio de respuesta, medido en el servidor [DEF]. No incluye la red hasta la cabina, el búfer de reproducción ni el altavoz, y no es el inicio audible percibido. Se excluye el turno 0, el saludo, que suena antes de que nadie hable. Las etapas registradas son la espera de fin de turno t_{fin}, el reconocimiento t_{asr}, el primer token del modelo t_{llm} y el primer byte de voz 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 recoge lo que el registro no separa. Las etapas pueden solaparse en streaming y las medianas no son aditivas, así que se comparan etapa a etapa y nunca se suman.

3.3. Estadística

Los turnos de una misma llamada no son independientes. Los intervalos del 95 % de medianas, percentiles y proporciones de turnos se calculan con bootstrap por conglomerados [14, 15]: se remuestrean conversaciones completas $B = 2000$ veces con semilla fija y se toma

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

con los percentiles 2,5 y 97,5 de $\hat{\theta}^{(b)}$. Los contrastes entre pilas usan la mediana de cada conversación con al menos dos turnos, que es la unidad independiente: Kruskal-Wallis global y Mann-Whitney por pares con corrección de Holm [16], acompañados del delta de Cliff como tamaño del efecto [17]. El tiempo muerto de una llamada es la fracción de su duración que se fue en esperar respuestas:

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

calculado en llamadas de al menos 20 s con tres o más respuestas medidas.

4. Resultados

4.1. Latencia por pila

La tabla 2 y la figura 1 resumen las cuatro distribuciones [MED]. La pila de voz a voz más rápida, Realtime v1 Mini, tuvo una mediana de 0,81 s (IC 95 % 0,78–0,84 s) y el 71 % de sus turnos llegó en menos de un segundo. La cascada con Qwen tuvo una mediana de 2,68 s (IC 95 % 2,45–2,94 s); solo el 3 % de sus turnos bajó del segundo y el 41 % superó los tres. La mejora de la mediana entre ambas fue de 3,3× (IC 95 % 3,1–3,6×).

Gemini 3.7 Flash fue la pila más lenta, con una mediana de 3,92 s (IC 95 % 3,67–4,43 s). Solo sirvió 20 conversaciones, por lo que su intervalo es el más ancho; su lentitud se debe al primer token del modelo (tabla 4).

4.2. Contrastes

Las medianas por conversación difieren entre pilas (Kruskal-Wallis, $H = 201{,}1$, $p < 0{,}001$). Todos los pares difieren tras la corrección de Holm (tabla 3). El efecto más pequeño separa Max de Mini, con un delta de Cliff de 0,43: en el mismo evento y con la misma voz, Mini respondió antes en la mayoría de las comparaciones entre llamadas.

4.3. Dónde se va el tiempo

La tabla 4 y la figura 2 descomponen la espera. El reconocimiento nunca fue el cuello de botella. El primer token del modelo pasó de 1515 ms con Qwen a 220 ms dentro de Realtime Max. La espera de fin de turno no es una variable aleatoria sino un parámetro de configuración: 320 ms en IFA y 160 ms en Barcelona [MED]. Restando esa espera, la mediana de la cascada sigue en 2,08 s frente a 0,59 s de Mini.

4.4. El primer turno es más lento

La latencia no es constante a lo largo de la llamada (figura 3). Con Realtime Max, el primer turno de respuesta tuvo una mediana de 1,10 s y los turnos séptimo y posteriores, de 0,84 s (Spearman entre posición y latencia, $\rho = -0{,}21$). En la cascada con Qwen la primera respuesta tardó 3,15 s frente a 2,54 s de las siguientes. Con Realtime Mini la diferencia casi desaparece, 0,83 s frente a 0,80 s. La primera respuesta es la primera impresión real del visitante, y en tres de las cuatro pilas fue la peor.

4.5. Tiempo muerto

Según la ecuación (3), la cascada con Qwen dedicó a esperar el 16 % (IC 95 % 15–18 %) de la llamada mediana; Realtime Mini, el 7 % (IC 95 % 6–7 %) (tabla 5). Uno de cada seis segundos de la cascada era silencio de máquina.

4.6. Sónar+D y coste

Los registros de Sónar+D no incluyen la métrica de extremo a extremo. En 2908 turnos, el primer token del modelo tardó 1011 ms y el primer byte de voz, 284 ms, en mediana. El coste de nube fue de 0,12 USD por minuto con la cascada de IFA y 0,10 USD con Realtime: la pila más rápida no fue más cara.

Tabla 2. Benchmark de campo. Tiempo hasta el primer audio desde el silencio, por turno de respuesta. Intervalos del 95 % por bootstrap de conversaciones.

Pila		Conv.	Turnos	Mediana (s) [IC 95 %]	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	2336	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 %

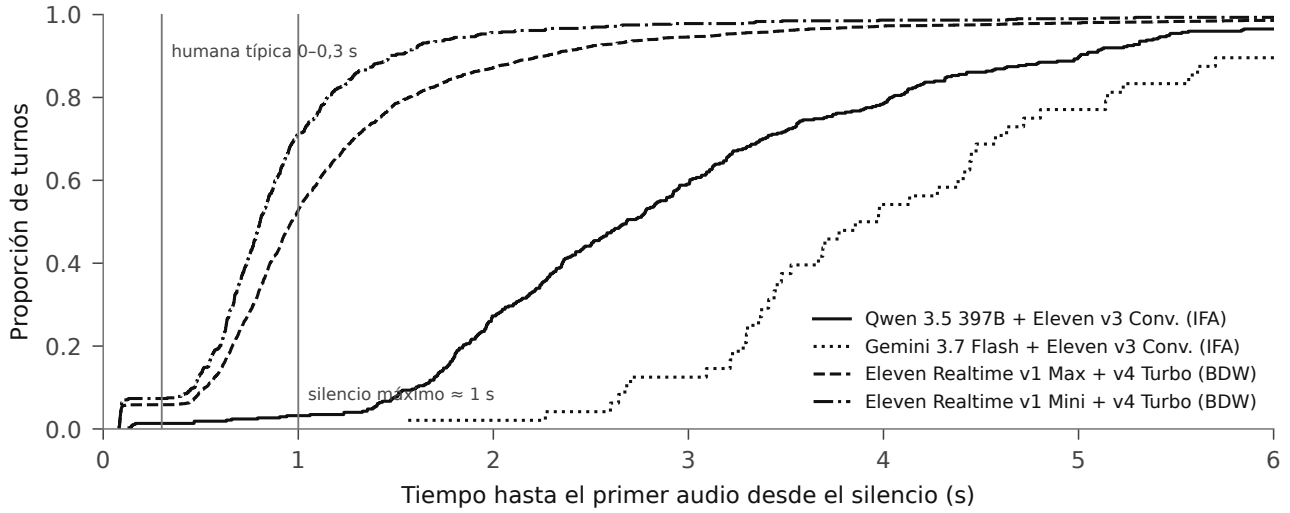


Figura 1. Distribución acumulada de la latencia por turno en cada pila. Las líneas verticales marcan la respuesta humana típica [2] y el silencio máximo estándar de Jefferson [3], como referencias, no como objetivos.

Tabla 3. Contrastes por pares sobre la mediana de cada conversación. Delta de Cliff positivo indica que la primera pila es más lenta.

Comparación	n conv.	δ de Cliff	p Holm
Qwen frente a Gemini	67 / 12	−0,62	<0,001
Qwen frente a RT Max	67 / 305	0,93	<0,001
Qwen frente a RT Mini	67 / 78	0,98	<0,001
Gemini frente a RT Max	12 / 305	0,99	<0,001
Gemini frente a RT Mini	12 / 78	1,00	<0,001
RT Max frente a RT Mini	305 / 78	0,43	<0,001

Tabla 4. Mediana de cada etapa registrada, en milisegundos. Las medianas no se suman.

Etapa (ms, mediana)	Qwen	Gemini	RT Max	RT Mini
Espera de fin de turno	320	640	160	160
Reconocimiento	56	42	59	47
Primer token LLM	1515	2851	220	181
Primer byte de voz	187	173	81	80

Tabla 5. Fracción mediana de la llamada dedicada a esperar respuestas.

Pila	Llamadas	Tiempo muerto [IC 95 %]
Qwen	51	16 % [15; 18]
Gemini	8	20 % [19; 26]
RT Max	288	8 % [8; 9]
RT Mini	75	7 % [6; 7]

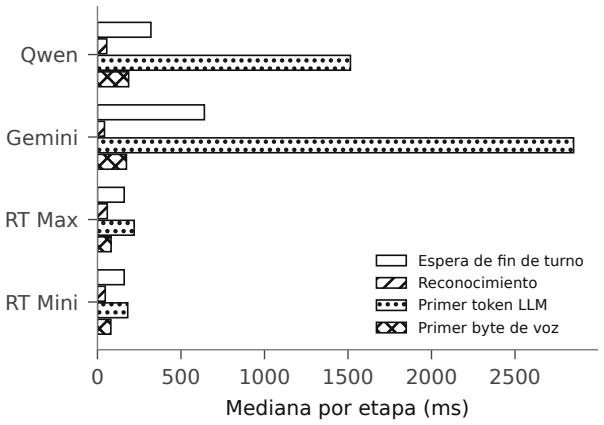


Figura 2. Mediana por etapa y pila. El primer token del modelo domina en las cascadas.

5. Benchmark comparativo

La tabla 6 pone nuestras cifras junto a las publicadas. La columna decisiva es la de comparabilidad. Ninguna cifra publicada mide lo mismo que la nuestra, y la nuestra tampoco mide el inicio audible.

Dos lecturas se sostienen. Primera: en campo, Realtime Max y Mini quedaron unos cientos de milisegundos por encima de su TTFA de laboratorio, en parte porque nuestra métrica incluye la espera configurada de fin de turno [INT]. Segunda: las medias del evaluador y nuestras medianas no son el mismo estadístico, y con colas largas la media es mayor que la mediana; la comparación es descriptiva y no admite contraste.

Tabla 6. Cifras de referencia y de campo. Cada fila conserva su métrica, su estadístico y su fuente. Las marcas indican el tipo de evidencia.

Sistema o población	Cifra	Qué mide	Comparable con el campo
Personas, diez lenguas [EMP]	208 ms media; 100 ms mediana	Fin de pregunta a inicio vocal o gestual de respuesta [1]	No: conversación cara a cara, redondeo a 100 ms
Personas, silencio máximo [EMP]	≈1 s	Umbral a partir del cual el silencio significa [3]	Como referencia de experiencia, no de sistema
Eleven Realtime v1 Max [EVA]	570 ms media	TTFA en Big Bench Audio [6, 18]	Parcial: sin espera de fin de turno ni sala
Eleven Realtime v1 Mini [EVA]	620 ms media	TTFA en Big Bench Audio [6, 18]	Parcial
GPT-Realtime-2.1 High [EVA]	1,21 s media	TTFA, configuración High [6]	Parcial; no desplegado aquí
GPT-Realtime-2.1 Minimal [EVA]	0,97 s media	TTFA, configuración Minimal [6]	Parcial; no desplegado aquí
Eleven v4 Turbo [FAB]	≈100 ms mediana	Inferencia de síntesis sin red [4, 5]	No: solo una etapa
Cascada, objetivo de diseño [FAB]	885 ms; 1115 ms	Hueco en plataforma; boca a oído [7]	Objetivo, no medición
GAIA, Realtime v1 Max [MED]	0,96 s mediana	Silencio detectado a primer audio, servidor, sala real	Referencia de campo
GAIA, Realtime v1 Mini [MED]	0,81 s mediana	Ídem	Referencia de campo
GAIA, cascada Qwen [MED]	2,68 s mediana	Ídem	Referencia de campo
GAIA, cascada Gemini [MED]	3,92 s mediana	Ídem	Referencia de campo

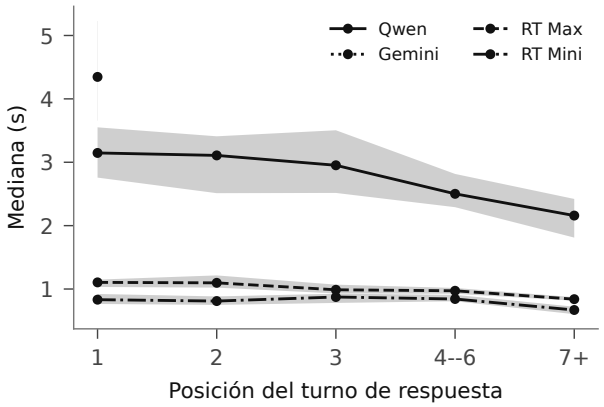


Figura 3. Mediana de latencia según la posición del turno de respuesta, con IC 95 % por bootstrap de turnos (banda gris). Se omiten posiciones con menos de 15 turnos.

6. Discusión

La mejora entre eventos salió de dos sitios: un modelo que empieza a hablar antes y un sistema que decide antes que le toca hablar [INT]. Ambos son decisiones de diseño, no propiedades fijas del hardware. La espera de fin de turno, en particular, es un parámetro que se elige; bajarla adelanta la respuesta y aumenta el riesgo de interrumpir, un compromiso que DM-2026-01 formaliza como una decisión bajo pérdidas [12].

La penalización del primer turno importa más de lo que su tamaño sugiere. Es el momento en que el visitante decide si la voz está viva; DM-2026-04 estudia ese primer segundo. Calentar la conexión antes de descolgar, que ya hacemos, no basta para el primer turno de respuesta [HIP].

El tiempo muerto traduce la latencia a experiencia. Una mediana de 2,7 s parece un problema técnico; el 16 % de una llamada en silencio es un problema de dramaturgia.

7. Amenazas a la validez

- *Confusión.* Cada pila corresponde a un evento, con su red, su público, su idioma y su versión del prompt. Las diferencias entre pilas son asociaciones [INT].
- *Medida.* La métrica es de servidor. No incluye red ni reproducción en la cabina; el inicio audible es más tardío.
- *Selección.* Solo cuentan turnos con métrica registrada; las llamadas atendidas por Vidiv no están.
- *Tamaño.* Gemini sirvió 20 conversaciones; su estimación es la menos precisa.
- *Contenido.* Una pregunta difícil puede tardar más; no controlamos la longitud ni la complejidad del turno.

8. Recomendaciones para la comunidad

Proponemos un estándar mínimo para reportar latencias de agentes de voz en campo:

1. Nombrar el evento de inicio y el de fin, y decir si incluye la espera de fin de turno.
2. Reportar mediana y percentil 90, nunca solo la media, con intervalos por bootstrap de conversaciones.
3. Descomponer por etapas y declarar qué etapas son parámetros configurados.

4. Separar el primer turno de respuesta de los siguientes.
5. Reportar el tiempo muerto por llamada, que es lo que vive el usuario.
6. Publicar el número de conversaciones, no solo el de turnos.

9. Datos y código

Dataset gaia-2147 con huella en su manifiesto [13]. Todas las cifras de este documento se generan con `analysis/analyze.py` y `kit/python/dmpaper.py`, semilla 2147. El texto de las conversaciones nunca se guardó.

10. Conclusiones

En una sala real, con público real, una pila de voz a voz respondió en 0,81 s de mediana y una cascada, en 2,68 s. La diferencia estuvo en el modelo y en la espera de fin de turno, no en el reconocimiento. Las cifras de laboratorio son útiles para elegir, no para prometer: lo que cuenta es el silencio que oye quien está al teléfono.

Firman. Cris Olmedo y Kurt Sommer, Grupo de investigación Divina Machina. *Cómo citar*: Olmedo, C. y Sommer, K. (2026). Latencia de agentes de voz en campo. Documento de trabajo DM-2026-03, v 1.0. Grupo de investigación Divina Machina. Divina Machina by Qs Ventures.

Referencias

Fuentes consultadas hasta el 29 de octubre de 2026. Las cifras de fabricantes y evaluadores no son mediciones de Divina Machina.

[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. Lectura de los datos de 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. Documentación de ElevenLabs, s. f. URL https://elevenlabs.io/docs/overview/models. Modelos, identificadores y latencias de inferencia.

[5] M. Staniszewski and P. Dabkowski. Introducing Eleven v4: our most emotive model. ElevenLabs, September 2026. URL https://elevenlabs.io/blog/eleven-v4. Fecha: 28 de septiembre de 2026. Actualizado el 19 de octubre de 2026. La nota metodológica 3 documenta la cifra de 150 ms.

[6] Artificial Analysis. Speech to speech models and providers analysis. Artificial Analysis, s. f. URL https://artificialanalysis.ai/speech-to-speech. Tabla viva de resultados; configuraciones de modelo explícitas.

[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. Documentación de ElevenLabs, s. f. URL https://elevenlabs.io/docs/eleven-agents/customization/realtime-models. Custom, Realtime, ajustes y disponibilidad gradual.

[11] Artificial Analysis. Speech to speech benchmarking methodology. Artificial Analysis, s. f. URL https://artificialanalysis.ai/methodology/speech-to-speech-benchmarking. Definición de TTFA y descripción de Big Bench Audio.

[12] C. Olmedo and K. Sommer. El arte de la conversación. de la trilogía Before al diseño de agentes conversacionales. Documento de trabajo DM-2026-01, v0.2. Grupo de investigación Divina Machina, 2026. Artículo conceptual; sin datos propios.

[13] C. Olmedo and K. Sommer. GAIA-2147. descriptor de datos de 1527 conversaciones con un agente de voz en espacios públicos. Documento de trabajo DM-2026-09. Grupo de investigación Divina Machina, 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. Página de producto, 2026. URL https://elevenlabs.io/realtime. Cifras de tiempo hasta el primer audio atribuidas a Artificial Analysis (Big Bench Audio).