GAIA-2147

Descriptor de datos de 1527 conversaciones con un agente de voz en espacios públicos

Cris Olmedo y Kurt Sommer
Grupo de investigación Divina Machina
Documento de trabajo · Descriptor de datos
DM-2026-09, versión 1.0 · 30 de octubre de 2026

Resumen. Los datos de agentes de voz en uso real, con público y en salas ruidosas, casi no existen fuera de las empresas que los operan. Describimos gaia-2147: 1527 conversaciones y 8151 turnos con métricas de latencia por etapa, recogidos entre el 16/06/2026 y el 26/10/2026 en una cabina telefónica que contesta un personaje artificial, GAIA 2147, en tres eventos públicos. Suman 29,2 h de llamada y 82 versiones del agente.

El conjunto contiene solo números y categorías por conversación y por turno: duraciones, cierres, versiones, saludos, recuentos de turnos, palabras, preguntas, etiquetas de voz, herramientas, reacciones léxicas y la descomposición de cada respuesta en espera de fin de turno, reconocimiento, primer token del modelo y primer byte de voz. No contiene texto de las transcripciones ni audio.

Seguimos la estructura de las fichas de datos: motivación, composición, recogida, preprocesado, validación técnica, usos, distribución y mantenimiento. Las 8 comprobaciones técnicas se cumplen. Este conjunto sostiene los documentos DM-2026-02 a DM-2026-08.

Palabras clave: conjunto de datos; agentes de voz; latencia; interacción en público; ficha de datos; privacidad.

1. Motivación

gaia-2147 se construyó para estudiar cómo conversa un agente de voz con desconocidos en espacios públicos: latencia, apertura, personalidad, uso del cuerpo y conducta del visitante. Lo recogió el Grupo de investigación Divina Machina sobre su propia obra, GAIA 2147, sin financiación externa específica. Seguimos la propuesta de fichas de datos de Gebru y colaboradores [1].

2. Composición

Cada instancia es una conversación telefónica entre un visitante y el agente, o un turno de esa conversación (tabla 1, figura 1). Las conversaciones de *Pruebas* se hicieron fuera de evento y se excluyen de las comparaciones. La figura 2 muestra la duración por evento; los límites de tiempo distintos truncan cada distribución de forma distinta.

2.1. Esquema

`raw/features.json` tiene 46 campos por conversación y `raw/turns.json`, 14 por turno (tablas 2 y 3).

Datos faltantes. La métrica de extremo a extremo no existe en Sónar+D y falta en una parte de los turnos de la cascada (tabla 4). Los análisis de la serie usan los turnos con la métrica presente y lo declaran.

3. Recogida

GAIA corre siempre sobre dos plataformas de voz, ElevenLabs y Vidiv. El conjunto procede de la API de conversaciones del agente de ElevenLabs y se descargó el 27 de octubre de 2026; las llamadas atendidas por Vidiv no están incluidas. Las llamadas se asignan a cada evento por fecha local. Los visitantes descolgaron voluntariamente un teléfono en un espacio público; no hubo selección ni reclutamiento.

4. Preprocesado y privacidad

La extracción lee cada transcripción en memoria, calcula números y categorías y descarta el texto. Lo que llega al disco son duraciones, recuentos, tiempos, nombres de herramientas, versiones y la clase del saludo, que es texto nuestro. Nadie leyó las conversaciones para construir el conjunto y no se publica ninguna frase de un visitante. Los identificadores de conversación son seudónimos de la plataforma; no hay voz, nombre ni imagen de nadie.

Definiciones léxicas [DEF]. Una pregunta es un turno que termina en signo de interrogación. El eco marca un turno del visitante de cuatro o más palabras con al menos el 60 % en el turno anterior del agente. El hilo retomado marca un turno del agente que usa una palabra de contenido de cinco o más letras dicha antes por el visitante y no usada antes por el agente. Las reacciones, temas e idiomas usan listas fijas en siete idiomas; son mínimos.

5. Validación técnica

8 de 8 comprobaciones de consistencia se cumplen (tabla 5). La integridad de los archivos se fija con SHA-256 en `MANIFEST.json`: `features.json` empieza por `9159aabef8531ef8` (1,4 MB) y `turns.json` por `ac9c51a92762b719` (3,0 MB). Cada análisis de la serie cita esa huella.

5.1. Métodos estadísticos de la serie

Proporciones con el intervalo de Wilson [2],

$$\hat{p}_{\pm} = \frac{\hat{p} + \frac{z^2}{2n} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}}, \qquad (1)$$

diferencias de proporciones con Newcombe [3], medianas con bootstrap [4] y, cuando hay varios turnos por

Tabla 1. Composición por evento.

Evento	Conv.	Turnos	Turnos visit.	Días	Horas	Versiones	Fallidas
Sónar+D	701	3732	3014	5	15,1	36	10,8 %
IFA	274	780	471	5	2,5	14	1,1 %
BDW	543	3622	2805	3	11,6	30	1,3 %
Pruebas	9	17	10	6	0,1	4	11,1 %

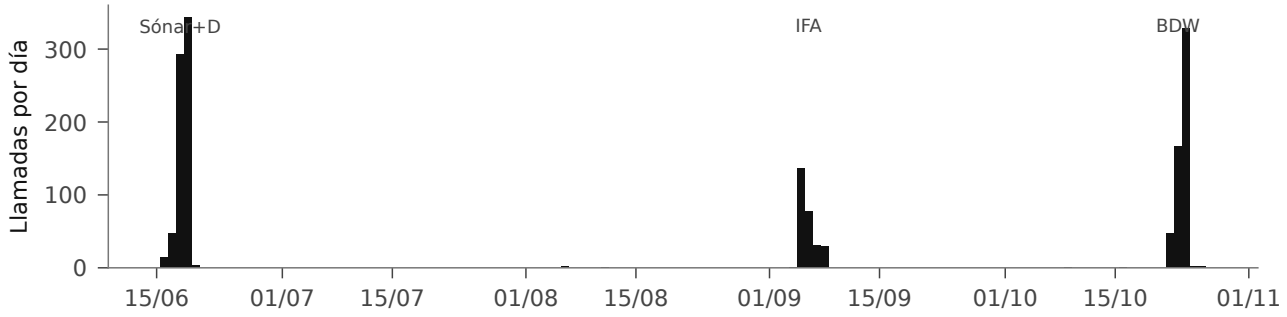


Figura 1. Llamadas por día en 2026.

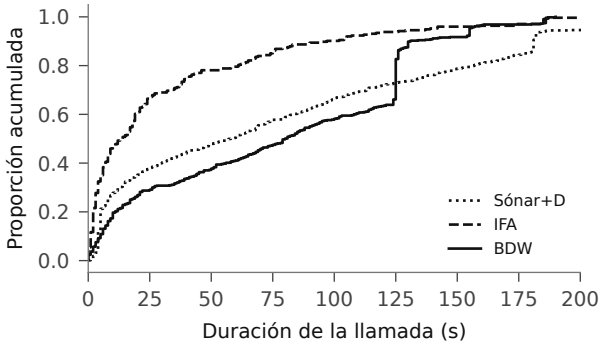


Figura 2. Distribución acumulada de la duración de la llamada por evento.

llamada, bootstrap por conglomerados [5]. Contrastes con Mann-Whitney, Fisher, chi cuadrado o Kruskal-Wallis, con corrección de Holm [6] y tamaño de efecto (delta de Cliff [7] o V de Cramér). Duraciones con censura en el límite de tiempo, con Kaplan-Meier [8]. Semilla 2147 y $B = 2000$ remuestreos.

6. Usos

Adecuado para: latencia de agentes de voz en campo; diseño de aperturas; medidas de comportamiento de personajes; uso de herramientas por modelos; conducta agregada de visitantes. No adecuado para: inferencias causales entre eventos, porque evento, pila, prompt, público y sala cambian juntos; estudios sobre personas concretas; evaluación de comprensión, porque las medidas léxicas son toscas; latencia percibida en el dispositivo, porque las métricas son de servidor.

7. Distribución y mantenimiento

El conjunto está en el repositorio interno del grupo (`datasets/gaia-2147`) con su ficha y su manifiesto. No se ha decidido licencia ni publicación; cualquier distribución requiere el visto bueno de Cris Olmedo. Las versiones nuevas conservarán la anterior y su huella. Contacto: los autores.

Tabla 2. Campos principales por conversación.

Campo	Tipo	Definición
`id`, `start`, `evento`, `hora`	texto, entero, categoría	Identificador seudónimo, inicio (epoch UTC), evento y hora local
`dur`, `status`, `fin`	entero, categoría	Duración (s), estado y motivo de cierre informados por la plataforma
`version`, `saludo`	texto, categoría	Versión del agente; clase del saludo inicial
`n_ag`, `n_vis`, `pal_ag`, `pal_vis`	entero	Turnos y palabras de cada parte
`preguntas`, `etiquetas`, `asistente`, `metafora`	entero, recuento	Hábitos del agente por léxico
`eco`, `callback_ok`, `callback_eleg`	entero	Eco acústico; hilo retomado y turnos elegibles
`herr`, `imagenes`, `colores`, `err_herr`	recuento	Herramientas decididas por el modelo y errores
`idioma_vis`, `idiomas_det`	categoría	Idioma del visitante (heurística) y cambios detectados
`r_*`	booleano, entero	Reacciones del visitante por léxico: saludo, si es real, risa, gracias, despedida, palabrotas, preguntas, turnos cortos, temas
`sent_label`, `exito`, `coste_usd`	categoría, número	Evaluación automática y coste de la plataforma

Tabla 3. Campos por turno. Métricas de servidor de la plataforma.

Campo	Definición
`id`, `i`, `t`	Conversación, índice en la transcripción, segundo
`llm`, `tts_model`	Modelo de lenguaje y de voz
`e2e`	Primer audio desde el silencio detectado (s)
`silencio`	Espera configurada de fin de turno (s)
`asr`	Latencia de reconocimiento (s)
`llm_ttfb`	Primer token del modelo (s)
`tts_ttfb`	Primer byte de síntesis (s)

8. Datos y código

Extracción con `docs/whitepapers/data/extract.py`; cifras de este documento con `analysis/analyze.py`.

9. Conclusiones

gaia-2147 es un registro pequeño y preciso de algo poco documentado: una máquina conversando con desconocidos en público, turno a turno. Lo que no guarda es tan deliberado como lo que guarda.

Firman. Cris Olmedo y Kurt Sommer, Grupo de investigación Divina Machina. *Cómo citar*: Olmedo, C. y Sommer, K. (2026). GAIA-2147. Documento de trabajo DM-2026-09, v 1.0. Grupo de investigación Divina Machina. Divina Machina by Qs Ventures.

Tabla 4. Presencia de cada métrica por grupo, en turnos de respuesta.

Grupo (turnos de respuesta)	Turnos	e2e	Silencio	ASR	LLM	TTS
Sónar+D	3051	0 %	0 %	0 %	95 %	92 %
IFA · Qwen 3.5 397B	427	88 %	88 %	77 %	98 %	88 %
IFA · Gemini 3.7 Flash	48	100 %	100 %	100 %	100 %	100 %
BDW · Eleven Realtime v1 Max	2404	97 %	97 %	90 %	94 %	98 %
BDW · Eleven Realtime v1 Mini	607	99 %	99 %	91 %	93 %	99 %

Porcentaje de turnos de respuesta con la métrica presente. Sónar+D no registró la métrica de extremo a extremo.

Tabla 5. Comprobaciones técnicas automáticas.

Comprobación	Resultado
Identificadores de conversación únicos	cumple
Todo turno pertenece a una conversación del conjunto	cumple
Duraciones no negativas	cumple
Latencias no negativas	cumple
Hilo retomado no supera los turnos elegibles	cumple
Preguntas no superan los turnos de GAIA	cumple
Eco no supera los turnos del visitante	cumple
Sin campos de texto libre de visitantes	cumple

Referencias

[1] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021. doi: 10.1145/3458723.

[2] E. B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. doi: 10.1080/01621459.1927.10502953.

[3] R. G. Newcombe. Interval estimation for the difference between independent proportions. comparison of eleven methods. *Statistics in Medicine*, 17(8):873–890, 1998.

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

[5] 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.

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

[7] 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.

[8] E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958. doi: 10.1080/01621459.1958.10501452.