El arte de la conversación

De la trilogía *Before* al diseño de agentes conversacionales

Cris Olmedo y Kurt Sommer
Grupo de investigación Divina Machina
Artículo conceptual y propuesta de investigación
DM-2026-01, versión 0.2 · 29 de octubre de 2026

Resumen. Una conversación puede cambiar el significado de un encuentro, de un recuerdo o de una decisión. Este artículo parte de una lectura interpretativa de la trilogía *Before* para investigar cómo el tiempo, la atención, el espacio y la memoria participan en esa transformación, y qué preguntas plantean al diseño de agentes conversacionales.

Mediante una revisión dirigida de investigación sobre interacción, entrevistas cinematográficas y documentación técnica, proponemos distinguir la disponibilidad de una respuesta del momento en que conviene intervenir. Desarrollamos definiciones de medida, modelos de decisión y una arquitectura para coordinar voz, objetos compartidos y memoria. La hipótesis de diseño es que esa coordinación puede sostener continuidad y comprensión sin reducir el control de la persona.

La contribución es conceptual y metodológica. Las cifras publicadas conservan su estadístico, unidad y condiciones de medida; se distinguen los resultados de investigación, las declaraciones de fabricantes y los informes de evaluadores. Las ecuaciones explicitan decisiones operativas y sus supuestos; no son modelos ajustados a datos propios. No se ha realizado una anotación temporal sistemática de las películas ni una evaluación experimental de los sistemas propuestos. El artículo presenta un programa de contraste, con requisitos de medición, análisis y prerregistro todavía pendientes de ejecución.

Palabras clave: conversación; diseño de interacción; agentes de voz; temporalidad; narrativa; memoria; interacción multimodal.

1. Introducción

1.1. Problema y definiciones de trabajo

En nuestra lectura de *Before*, una conversación vuelve significativo un trayecto porque quienes lo recorren llegan a comprender algo distinto al final. Las palabras participan en ese cambio junto con las miradas, los lugares y el tiempo disponible. Tomamos esa lectura como punto de partida para una pregunta de investigación: ¿qué decisiones de coordinación permiten que un intercambio amplíe lo que una persona puede comprender, expresar o decidir? La propuesta busca describir condiciones observables y contrastables, conservando la diferencia entre una escena y la experiencia de quien conversa con un sistema.

En este artículo llamamos narrativa conversacional a una secuencia de intercambios cuyo desarrollo modifica una situación compartida y permite reconocer continuidad entre lo anterior y lo siguiente. Esa continuidad no exige una trama cerrada. Puede consistir en descubrir una pregunta mejor, reconocer un desacuerdo o decidir que todavía no queremos decidir. Una conversación administrativa excelente puede terminar pronto; una exploración creativa puede beneficiarse de una digresión.

Llamamos atmósfera conversacional a la cualidad percibida de estar en una situación con cierto ritmo, margen de acción, atención y continuidad. Es una definición de trabajo, no una magnitud física ni una lectura automática del estado emocional. La atmósfera se investiga mediante conductas observables, entrevistas y valoración de los participantes. El sistema puede registrar una pausa; no puede concluir por ese solo dato que existe confianza, tristeza o intimidad.

Una pausa tiene una duración que podemos medir y un sentido que debemos preguntar a quien la vivió. Esta formulación orienta la investigación: el registro temporal y el relato de la experiencia aportan información diferente y deben conservar su procedencia.

La hipótesis central es que diseñar esa atmósfera requiere coordinar lenguaje, temporalidad, espacio y memoria. El modelo lingüístico genera o interpreta contenido; el producto establece quién puede intervenir, qué objetos aparecen y qué persiste. Si esas decisiones se dejan al azar de la infraestructura, la experiencia también queda subordinada a ese azar.

1.2. Alcance y enfoque del artículo

La revisión es narrativa, dirigida y exploratoria; no constituye una revisión sistemática ni un metaanálisis. Se priorizan artículos de investigación, entrevistas primarias, documentación de fabricantes e informes del evaluador que realizó cada benchmark. Las fuentes se seleccionan por su relación con la coordinación de turnos, la comprensión compartida y el diseño narrativo. No se ha reconstruido un flujo exhaustivo de búsqueda, cribado y exclusión; por ello, no se afirma cobertura completa ni ausencia de sesgo de selección.

La información de producto tiene fecha de corte el 29 de octubre de 2026. Las películas contienen spoilers en el análisis. Las referencias numeradas se encuentran al final; las escenas se identifican por su situación, sin inventar minutajes que podrían variar entre ediciones.

El análisis cinematográfico presentado es interpretativo y se apoya en las fuentes citadas. La anotación sistemática de escenas y la comparación entre codificadores forman parte del programa de investigación propuesto.

1.3. Trazabilidad y estatuto de las afirmaciones

Se distinguen resultados empíricos publicados, declaraciones técnicas de fabricantes, evaluaciones de terceros, definiciones matemáticas, lecturas cinematográficas e hipótesis de diseño. Cada categoría autoriza infe-

rencias diferentes. Una cifra comercial no se convierte en una medición propia al citarla; una asociación observada no establece por sí sola una relación causal; una ecuación correctamente escrita no acredita que sus parámetros hayan sido estimados.

Para las cifras centrales se registra fuente, fecha de consulta, variable, unidad, estadístico, población o sistema y condiciones conocidas. Se conserva la diferencia entre *no informado en la fuente*, *no verificado en esta revisión* y *no medido por este proyecto*. La procedencia y los límites de las cifras están recogidos en el material suplementario. La sensibilidad poética interviene en las preguntas y en la interpretación de las escenas; las mediciones conservan sus definiciones y su incertidumbre.

2. Análisis cinematográfico de la trilogía *Before*

2.1. Autoría compartida y construcción de naturalidad

Before Sunrise se estrenó en 1995, *Before Sunset* en 2004 y *Before Midnight* en 2013. El guion acreditado de la primera corresponde a Richard Linklater y Kim Krizan. Las dos siguientes acreditan el guion a Linklater, Julie Delpy y Ethan Hawke; Sunset acredita además la historia a Linklater y Krizan. Conviene preservar esta autoría: atribuir todo el lenguaje de la trilogía a un único director borraría una parte de su método. [1–3]

La espontaneidad percibida no demuestra improvisación en rodaje. En una entrevista sobre Midnight, Linklater describe una construcción minuciosa y Delpy explica que ensayaron durante dos semanas con un guion prácticamente cerrado. En su recuerdo de Sunrise, Linklater describe una reescritura intensa junto a los intérpretes. La contribución de cada uno atravesaba ambos personajes. [4, 5]

Lectura cinematográfica. Entendemos ese trabajo previo como una condición de la naturalidad que percibimos en la relación. Para un agente, la traducción útil es preparar condiciones, límites y capacidades de reparación. No consiste en obligar a una persona a interpretar un papel dentro de una historia prevista.

2.2. *Before Sunrise*: descubrimiento y atención compartida

Dos desconocidos deciden prolongar un encuentro y recorrer Viena hasta la mañana. En nuestra lectura, el tiempo limitado convierte cada desvío en una elección; la conversación prueba afinidades, diferencias y formas de exposición, mientras la ciudad ofrece situaciones que reorientan el diálogo. En la cabina de escucha de una tienda de discos, la atención a una canción hace visibles miradas y vacilaciones. [1, 5] En el café representan llamadas imaginarias a sus amigos. [6]

Nuestra lectura es que el interés narrativo surge de una transformación gradual del permiso: qué puedo preguntarte ahora, qué puedo reconocer delante de ti, qué puedo arriesgar sin que el encuentro se rompa. El contenido intelectual importa, pero también funciona como forma indirecta de tantear la relación. Una idea sobre la vida puede ser, simultáneamente, una oferta de cercanía o una protección frente a ella.

Leemos el silencio de la cabina como una forma de atención sostenida por una actividad compartida. Ambos escuchan algo; no están simplemente esperando a que alguien genere una frase. Linklater cuenta que reservó la escucha de la grabación de Kath Bloom para el rodaje, aunque la escena estaba ensayada. Eso permite separar una estructura preparada de una reacción que conserva frescura. [5]

Propuesta de diseño. Un agente puede ofrecer un objeto compartido, una imagen o una pieza sonora, y permitir que la persona elija qué hacer con él. Debe poder permanecer disponible sin producir comentarios continuos. El equivalente digital del paseo sería una exploración con posibilidades de desvío y retorno, no un itinerario que simule libertad mientras conduce a una conversión predeterminada.

2.3. *Before Sunset*: reencuentro y reinterpretación del pasado

El reencuentro en París ocurre nueve años después. Un compromiso de viaje limita la duración disponible. El recorrido desde la librería hacia otros espacios, incluido el apartamento de Céline, acompaña la revelación de versiones distintas de lo ocurrido y de lo que sus vidas han llegado a ser. [2, 7] Leemos la canción y el gesto final como la sugerencia de una decisión cuyo alcance queda abierto, sin convertirla en una explicación exhaustiva.

Linklater explica que organizaron una progresión emocional: el café permite una primera honestidad todavía protegida; el coche concentra una mayor exposición. El tiempo que se agota favorece esa progresión. También relata que una imitación de Nina Simone realizada por Delpy durante la escritura le permitió reconocer el final que buscaban. [7]

Nuestra lectura sitúa la memoria como material activo. Un recuerdo no vale solo porque se recupera: importa qué significa ahora para cada persona. El pasado regresa con desacuerdo, omisión y reinterpretación. La conversación se vuelve historia cuando obliga a revisar una versión que parecía estable.

Propuesta de diseño. La memoria de un agente debería poder presentarse con procedencia y apertura a la corrección: «La última vez anotamos esto; ¿todavía te representa?». Recuperar un dato no autoriza a tratarlo como identidad permanente. La continuidad exige permitir que alguien cambie, contradiga una preferencia o retire una interpretación anterior.

2.4. *Before Midnight*: convivencia y reparación

En Grecia, la pareja comparte una vida atravesada por responsabilidades, hijos y expectativas. Las conversaciones del coche, la comida con otras personas y el hotel desplazan la atención desde el descubrimiento inicial hacia la convivencia y sus conflictos. [3, 4, 8] Leemos el encuentro final como un intento de reparación que conserva las tensiones expuestas.

Nuestra lectura es que la memoria compartida puede ampliar el significado de una frase hasta volverla

Tabla 1. Recursos de la trilogía y traducciones propuestas al diseño conversacional.

Recurso	Función en nuestra lectura	Traducción de diseño	Límite
Tiempo disponible	Da peso a continuar o despedirse	Acordar duración y permitir cerrar	No fabricar urgencia
Recorrido	Cambia el marco y abre temas	Navegar objetos o perspectivas	No distraer por sistema
Subtexto	Permite significados indirectos	Pedir aclaración con tacto	No inferir deseos ocultos como hechos
Objeto compartido	Distribuye la atención	Imagen o documento manipulable	No imponer una lectura
Recurrencia	Une momentos separados	Recuperar una referencia con procedencia	No congelar la identidad
Reparación	Cambia lo que puede ocurrir después	Admitir error y revisar contexto	No exigir reconciliación

Nota. Lectura cinematográfica y propuesta de diseño del artículo; no son equivalencias empíricamente validadas.

difícil de soportar. Una observación práctica carga con conversaciones anteriores. La precisión verbal no garantiza comprensión; conocer al otro también permite anticiparlo de manera injusta. La historia puede avanzar mediante una reformulación, una retirada o una oferta que no exige una capitulación inmediata.

Propuesta de diseño. La calidad de un agente debe evaluarse cuando se equivoca, cuando el usuario lo corrige o cuando una inferencia sobre su vida resulta inadecuada. Reparar supone reconocer el error concreto, actualizar el contexto y devolver control. Acumular datos sobre una persona no proporciona una licencia para interpretar sus motivaciones ni para insistir en una supuesta verdad emocional.

2.5. Escalas de temporalidad

La trilogía articula el instante del intercambio, la duración de un encuentro y los años transcurridos entre encuentros. Sunrise condensa varias horas; Sunset está concebida en tiempo real. Eso no significa que esté rodada en un único plano. Linklater relaciona su interés por el tiempo con Tarkovsky y señala la precisión de Rohmer, incluyendo *The Green Ray* como influencia. Estas declaraciones sostienen una genealogía concreta, sin convertir cada semejanza estética en una influencia demostrada. [9]

Proponemos distinguir una escala micro, donde importan entradas de voz, pausas y solapamientos; una escala de escena, donde una actividad o pregunta mantiene continuidad durante varios minutos; y una escala de trayectoria, donde reaparecen objetos y acuerdos entre sesiones. Cada escala exige herramientas diferentes. Mejorar el tiempo hasta el primer audio no resuelve una memoria mal atribuida, y recuperar una memoria no resuelve una interrupción constante.

2.6. Recursos transferibles al diseño

La tabla 1 sintetiza las correspondencias de diseño propuestas y sus límites.

3. Antecedentes y fundamentos de la interacción conversacional

3.1. Tradición cinematográfica y límites de la transferencia

El cine ofrece un laboratorio de formas, pero sus escenas son construcciones para espectadores. Un montaje puede ocultar fricciones, un encuadre puede volver legible una mirada y un actor puede sostener una ambigüedad que un agente de voz no tiene medios para representar. Por eso no debemos leer una película como una medición de conversación espontánea.

Rohmer permite estudiar cómo la palabra organiza la distancia entre una justificación y una elección. *My Dinner with André*, de Louis Malle, escrito por André Gregory y Wallace Shawn, permite observar cómo una situación espacial contenida puede sostener diferencias profundas de experiencia y perspectiva. El proceso que describe Gregory incluye conversaciones grabadas, selección y ensayo. Como contrapunto posterior, Ryusuke Hamaguchi explica un trabajo de lectura repetida y atención a voz y cuerpo que permite investigar la escucha más allá del significado verbal. Estas son comparaciones de análisis; no afirmamos una cadena causal entre todas las obras. [9–11]

After Yang aporta una pregunta especialmente fértil para sistemas con memoria. Kogonada explica su búsqueda de una interfaz de recuerdos que no se pareciera a un escritorio con carpetas y distingue formalmente los recuerdos registrados por Yang de los recuerdos humanos reinterpretados. La decisión es cinematográfica, no una descripción científica de cómo recuerda el cerebro. Nuestra lectura es que el archivo puede participar en la atmósfera: la forma de volver a algo también cambia su significado. [12]

La diferencia que debe orientar cualquier transferencia es la agencia. En una película, los personajes sirven a una obra. En una conversación con un sistema, la persona debe poder abandonar, discrepar, reducir la intensidad, pedir hechos o declarar que no quiere recordar. El diseño narrativo tiene que soportar esas decisiones. Una herramienta no fracasa porque el usuario cambie la historia prevista.

Esto impide tomar el conflicto, la escasez de tiempo o la intimidad como trucos para prolongar una sesión. Un sistema puede sostener una atmósfera sin atribuirse experiencias humanas ni reclamar reciprocidad emocional. El objetivo propuesto es una participación comprensible y libre, compatible con los fines de la persona.

3.2. Organización de los turnos

El trabajo de Sacks, Schegloff y Jefferson describe la organización local de los turnos y su adaptación a los participantes. Proporciona una base para comprender la conversación como actividad coordinada. No establece un intervalo óptimo universal. [13]

Stivers y colaboradores analizaron preguntas funcionales de sí/no y sus respuestas en diez lenguas, en una base de 101 conversaciones. Reportaron media conjunta de 208 ms, mediana de 100 ms y moda de 0 ms. El comienzo de la respuesta podía ser vocal o gestual; las medidas instrumentales se realizaron en incrementos de 10 ms y se redondearon a 100 ms. Por tanto, la moda cero no implica sincronización exacta y la variable no coincide necesariamente con el inicio audible de un agente. Estos estadísticos describen ese corpus; no fijan un intervalo óptimo universal. [14]

Levinson y Torreira revisan cómo intervalos conversacionales breves conviven con tiempos de producción lingüística de 600 ms o más en determinadas tareas. La explicación implica anticipación y solapamiento entre comprensión y preparación. Los 600 ms no son una constante del pensamiento humano. [15]

Propuesta de diseño. El agente debería poder preparar hipótesis mientras escucha y descartarlas si el usuario continúa. Esa anticipación puede reducir espera, pero no debe ejecutar acciones externas ni reproducir una conclusión antes de tener contexto suficiente. Comprender de forma incremental y actuar de forma prematura son decisiones diferentes.

3.3. Efectos contextuales de la rapidez y la pausa

Los estudios de conversación de Templeton y colaboradores relacionaron intervalos de respuesta con valoraciones de conexión recogidas al revisar grabaciones. En un experimento separado, se modificaron los intervalos de fragmentos grabados y se evaluó la conexión percibida por oyentes externos. La inferencia causal de esa manipulación corresponde a la percepción de terceros; los estudios de conversación vivida son observacionales. El resultado no demuestra que acelerar un agente produzca mayor conexión para su usuario. [16]

Otro trabajo del mismo grupo observó asociaciones diferentes entre intervalos superiores a dos segundos y conexión declarada entre amigos y desconocidos. El umbral fue una definición analítica; amistad y duración de la pausa no se asignaron experimentalmente. Durante esos intervalos podían observarse risas y otras conductas no verbales. Su interpretación como reflexión compartida corresponde a una hipótesis de los autores, no a una lectura directa de un estado interno. [17]

La hipótesis de transferencia es, por tanto, contextual. Una respuesta inmediata a una pregunta cerrada puede resultar adecuada. La misma entrada durante una búsqueda de palabras puede sentirse invasiva. Un silencio acompañado por una imagen que ambos contemplan no tiene necesariamente la misma función que una espera sin señal alguna mientras una herramienta falla.

3.4. Comprensión compartida y acción situada

Clark y Brennan describen el grounding como la construcción de evidencia de comprensión suficiente para el propósito presente. Goodwin estudia cómo habla, cuerpo y materiales del entorno se integran en la acción situada. Ambos permiten fundamentar una atención a lo compartido que excede la transcripción, aunque no demuestran que añadir imágenes a un agente mejore su calidad. [18, 19]

Proponemos registrar por separado que un objeto se produjo, se mostró, fue mencionado por el usuario y recibió una confirmación pertinente. Son estados distintos. Un clic tampoco prueba comprensión de todo el contenido. Esta separación evita que la memoria del sistema convierta automáticamente una pantalla renderizada en un conocimiento compartido.

3.5. Detección de voz y predicción de cesión del turno

Voice Activity Projection, de Ekstedt y Skantze, estudia la predicción de actividad vocal futura y eventos de toma de turno. Su interés para este proyecto es separar la detección de una pausa acústica de la estimación de lo que puede ocurrir después. Esa predicción no equivale a leer una intención emocional. [20]

En el prototipo, una pausa debe conservar incertidumbre hasta que existan señales suficientes. El sistema puede combinar información acústica, final sintáctico, tipo de acto, contexto de tarea y preferencias expresadas. El diseño debe ofrecer una alternativa sencilla cuando la predicción falle: una interrupción efectiva, un control para mantener el turno o una forma explícita de indicar que ya se terminó.

4. Definiciones operativas y modelos de decisión

Las ecuaciones siguientes tienen estatutos distintos: las ecuaciones (1) y (5) son definiciones operativas; la ecuación (2) es una idealización temporal de implementación; la ecuación (3) deriva de una matriz de pérdidas simplificada; y la ecuación (4) es una plantilla genérica de decisión bajo incertidumbre. Ninguna ha sido ajustada ni validada con datos experimentales propios. Su contribución consiste en hacer explícitas las decisiones y los supuestos que una implementación deberá contrastar.

4.1. Intervalo audible entre turnos

$$g_i = t^{\text{inicio}}_{\text{agente},i} - t^{\text{fin}}_{\text{usuario},i}. \tag{1}$$

En la ecuación (1), los tiempos se expresan en milisegundos sobre un reloj común. El inicio del agente corresponde al audio reproducido en el dispositivo; el

final del usuario se determina sobre el registro acústico, no sobre la decisión posterior del detector. Un valor positivo es un intervalo sin habla; uno negativo indica solapamiento. Se etiquetan aparte reconocimientos breves, solapamientos cooperativos e interrupciones invasivas. Sin esa clasificación, un promedio mezcla conductas distintas.

El protocolo deberá fijar qué final de intervención corresponde a cada inicio del agente, incluidos reinicios y reconocimientos breves. Si no aparece una respuesta antes de terminar la observación, se registrará ausencia de transición o censura según el diseño; no una latencia igual a cero. Un fallo de captura es un dato faltante. La distribución de intervalos observados debe acompañarse de la frecuencia de respuestas no observadas.

También se documentarán resolución del registro, sincronización de relojes, criterio de comienzo acústico y desacuerdo entre anotadores. La variabilidad de los intervalos entre personas no equivale a la incertidumbre con que se mide un intervalo. La distinción metrológica entre repetibilidad y reproducibilidad exige declarar qué condiciones se mantienen y cuáles cambian; se adopta aquí para especificar pruebas técnicas, sin equiparar la experiencia subjetiva con una magnitud física. [21]

4.2. Disponibilidad técnica y tiempo de intervención

$$t_i^{\text{inicio}} = \max\!\left(t_i^{\text{listo}},\ t_{\text{usuario},i}^{\text{fin}} + \delta_i\right) + d_i^{\text{salida}}. \tag{2}$$

La ecuación (2) describe el inicio de un turno de respuesta no cancelado, excluidos los reconocimientos breves. t_i^{listo} es el instante en que el primer fragmento reproducible está disponible en un punto de liberación especificado; $\delta_i \geq 0$ es la pausa objetivo y $d_i^{\text{salida}} \geq 0$ el retardo posterior hasta hacerlo audible. Las marcas t_i^{inicio}, t_i^{listo} y $t_{\text{usuario},i}^{\text{fin}}$ están referidas al mismo reloj; δ_i y d_i^{salida} son duraciones. Todo retardo posterior al punto de liberación pertenece a d_i^{salida}, para evitar contar red o búfer dos veces.

El cálculo y la espera deliberada pueden solaparse. Por eso no deben sumarse ciegamente latencia técnica y pausa. Si el audio tarda 800 ms en estar listo y la política pretendía esperar 400 ms, la respuesta no demuestra una decisión de esperar 800 ms: la infraestructura impuso al menos una parte del intervalo. Los valores son ilustrativos, no objetivos recomendados. La figura 1 ilustra la distinción.

En una implementación real, δ_i se calcula con estimaciones disponibles en línea y puede cambiar mientras se escucha. El final acústico definitivo se conoce después y sirve para evaluar el comportamiento. Si el usuario retoma el turno, el plan debe poder cancelarse. La fórmula no sustituye la política de escucha.

4.3. Costos de interrupción y demora

$$p_t > \frac{C_{\text{int}}}{C_{\text{int}} + C_{\text{dem}}(\Delta t)}. \tag{3}$$

Sea $F_t = 1$ la etiqueta operacional de cesión del turno y $p_t = P(F_t = 1 \mid h_t)$ su probabilidad estimada a partir del historial. Se consideran dos acciones: hablar o esperar un intervalo fijo $\Delta t > 0$. Bajo la simplificación de pérdida cero para hablar cuando hay cesión y esperar cuando no la hay, las pérdidas esperadas son:

$$L_{\text{hablar}} = (1 - p_t)C_{\text{int}},$$
$$L_{\text{esperar}} = p_t C_{\text{dem}}(\Delta t).$$

La ecuación (3) resulta de exigir $L_{\text{hablar}} < L_{\text{esperar}}$, con costos no negativos y suma estrictamente positiva. Ambos costos usan la misma unidad de pérdida. Si la demora se expresa como tasa, $C_{\text{dem}}(\Delta t) = \lambda \Delta t$, con λ en pérdida por unidad de tiempo. En empate se propone continuar escuchando. La regla es de un paso y omite otros costos y el valor de información futura que podría aportar esperar.

Los costos deberán estimarse y revisarse por contexto; este artículo no dispone de costos ajustados ni de una probabilidad calibrada. Una implementación deberá definir la etiqueta de referencia, documentar su incertidumbre y evaluar calibración fuera de la muestra de ajuste. Durante un relato vacilante y ante una petición urgente pueden importar errores distintos; esa diferencia es una hipótesis de decisión, no un peso universal ya conocido.

4.4. Selección de acciones permitidas

$$a_t^* \in \arg\max_{a \in \mathcal{A}_t} \mathbb{E}_{Z_t \sim b_t}[u(a, Z_t)]\,. \tag{4}$$

Aquí $b_t(z) = P(Z_t = z \mid h_t)$ es una distribución sobre estados de interacción relevantes para la tarea, y $\mathcal{A}_t = \mathcal{A}_{\text{permitidas}}(h_t)$ incorpora permisos y restricciones. Z_t no designa un estado psicológico directamente observado. La acción puede ser hablar, esperar, reconocer, aclarar, ofrecer o recuperar un objeto, proponer guardarlo, reparar o cerrar. La pertenencia al argmax admite empates.

Esta expresión es una plantilla de decisión, no un modelo estimado. Deben especificarse estado, actualización de creencias, utilidad u, horizonte y regla de desempate. Cualquier combinación de resultados, esfuerzo e intrusión requiere una escala común y pesos declarados; sumar medidas heterogéneas sin esa especificación carecería de interpretación operativa. Su validez deberá examinarse mediante la evaluación de una implementación concreta.

Los permisos limitan el conjunto de acciones antes de optimizarlo. Una mejora supuesta de engagement no puede compensar la ausencia de autorización para conservar un recuerdo o compartir un archivo. El modelo tampoco debe maximizar exposición emocional. Puede escoger terminar cuando la necesidad está satisfecha.

4.5. Medidas de coordinación y calidad

$$R_{\text{prematuro}} = \frac{N_{\text{prematuros}}}{N_{\text{turnos}}}. \tag{5}$$

En la ecuación (5), $N_{\text{prematuros}}$ cuenta inicios prematuros y N_{turnos} inicios del agente que pudieron codificarse;

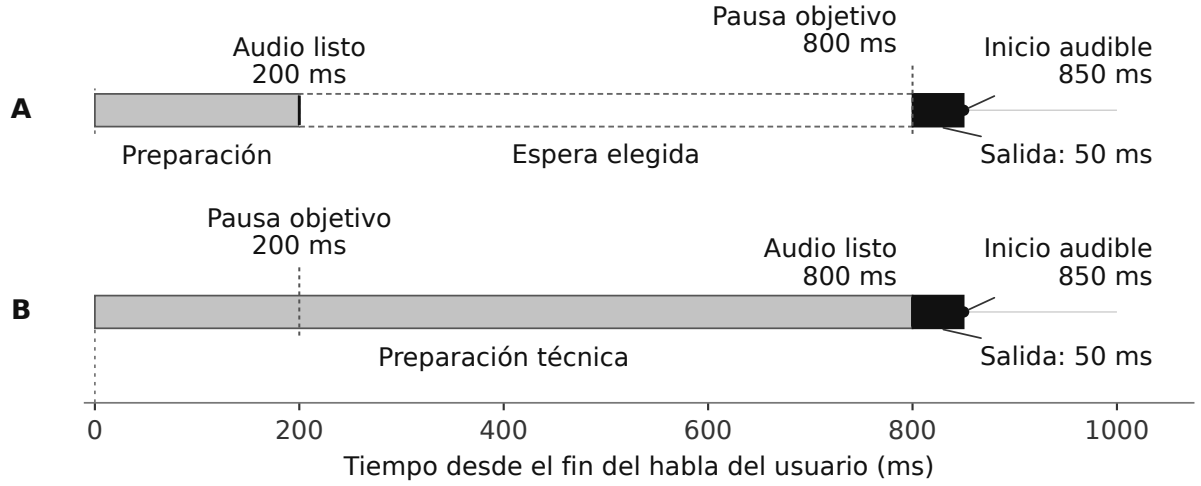


Figura 1. Dos situaciones ilustrativas producen el mismo inicio audible a los 850 ms. En A, la pausa objetivo determina cuándo liberar el audio; en B, lo determina su disponibilidad técnica. En ambos casos se aplica el máximo de la ecuación (2), más 50 ms de salida. Valores hipotéticos: no proceden de un benchmark ni constituyen objetivos de producto.

el numerador es un subconjunto del denominador. Cada evento recibe una etiqueta conforme a un manual con señales de continuación, reclamación del turno y, cuando corresponda, valoración del participante. El solapamiento acústico no determina por sí solo la intención de hablar.

Se informarán por separado los inicios totales, codificables y faltantes. Con cero inicios la tasa no aplica; si hubo inicios pero ninguno pudo codificarse, no puede estimarse por falta de datos. Se distinguirá una proporción conjunta ponderada por turnos de un promedio por persona. Para el tiempo de detención se fijarán el inicio del intento de interrupción y el criterio de cese audible; sin intento, esa medida no aplica. Mediana, percentiles y sus incertidumbres se comunicarán con sus denominadores y método de estimación.

La calidad se representa inicialmente como un perfil, no como una puntuación total: resultado acordado, adecuación temporal, comprensión, continuidad, control, esfuerzo y exactitud de memoria. PARADISE ofrece un precedente para relacionar resultados y costos conversacionales con satisfacción mediante estimación empírica. No proporciona coeficientes universales listos para trasladar a este proyecto. [22]

5. Capacidades tecnológicas y límites de comparación

5.1. Definiciones y cifras de latencia

Eleven v4 y Eleven v4 Turbo se anunciaron el 28 de septiembre de 2026. El catálogo describe aproximadamente 100 ms de mediana de inferencia para Turbo, excluyendo aplicación y red. Se trata de síntesis de voz, no del intervalo completo entre la última palabra del usuario y la respuesta audible de un agente. [23, 24]

Una cadena puede incluir captura, reconocimiento, decisión de fin de turno, generación, herramientas, síntesis, red y reproducción. Algunas etapas se solapan mediante streaming. La latencia observada debe instrumentarse en el sistema completo; sumar cifras aisladas de catálogo puede contar dos veces algunos tiempos y omitir otros. La documentación de ElevenLabs distingue inferencia y tiempo hasta el primer audio. [25]

El ensayo de TTS de 150 ms utilizó guiones iguales, ajustes por defecto y streaming WebSocket para v4 Turbo; midió y descontó la red. El catálogo y el anuncio consultados no aportan tamaño muestral, dispersión ni intervalos de confianza específicos para las latencias de 100 y 150 ms. Se conservan como declaraciones del fabricante bajo las condiciones descritas. [23, 24]

Las cuatro cifras de Eleven Realtime y GPT-Realtime-2.1 en la tabla 2 proceden directamente de Artificial Analysis. Su TTFA es la media hasta el primer token de audio de salida sobre preguntas de Big Bench Audio. Es un evento anterior a la reproducción audible y debe conservar su definición. High y Minimal identifican configuraciones del evaluador; Minimal no designa la variante de modelo Mini. [26, 27]

Las páginas consultadas no aportan n efectivo, repeticiones, exclusiones ni intervalos de confianza por configuración para TTFA, y no formalizan suficientemente el evento inicial del cronómetro. El tamaño nominal del conjunto de preguntas no se adopta como número efectivo de mediciones. La comparación es descriptiva: las fuentes no permiten acreditar igualdad de red, región, carga, transporte y todos los ajustes, ni inferir significación estadística o superioridad general. [27]

5.2. Arquitecturas modulares e integradas

ElevenLabs ofrece un modo Custom para elegir componentes y un modo Realtime que gestiona reconocimiento, modelo y síntesis. El anuncio de producto consultado de Eleven Realtime v1 Mini y Max está fechado el 20 de octubre de 2026. La documentación describe un despliegue gradual por cuentas. Conviene verificar la disponibilidad concreta antes de diseñar un experimento dependiente de esa función. [28, 29]

OpenAI documenta GPT-Realtime-2.1 como modelo de voz a voz con razonamiento configurable, uso de herramientas y mejoras en silencio, ruido e interrupciones.

Tabla 2. Cifras publicadas por familia de medida. TTS y TTFA corresponden a eventos y estadísticos distintos.

Sistema	Cifra publicada	Qué representa	Qué no demuestra
Eleven v4 Turbo	≈100 ms de mediana	Inferencia de síntesis sin aplicación ni red	Espera completa del usuario
Eleven v4 Turbo	≈150 ms de mediana	Solicitud a habla audible en un ensayo de TTS con red descontada	Intervalo desde el final de voz del usuario
Eleven Realtime v1 Max	0,57 s de media	TTFA de AA: primer token de audio, Big Bench Audio	Inicio audible en el dispositivo
Eleven Realtime v1 Mini	0,62 s de media	Misma métrica y evaluador	Diferencia estadísticamente significativa frente a Max
OpenAI GPT-Realtime-2.1 High	1,21 s de media	Misma métrica y evaluador; configuración High	Rendimiento universal de la familia de modelos
OpenAI GPT-Realtime-2.1 Minimal	0,97 s de media	Misma métrica y evaluador; configuración Minimal	Equivalencia con la variante de modelo Mini

Nota. Fuentes: [23, 24, 26, 27]. AA: Artificial Analysis. Las dos primeras filas son medianas de TTS; las cuatro restantes, medias de TTFA del evaluador. No son mediciones de Divina Machina. Consulta: 29 de octubre de 2026.

Aumentar el esfuerzo de razonamiento puede aumentar la latencia. Su ficha admite imagen de entrada; no equivale a generar imágenes ni a decidir cómo se muestran en una interfaz. [30]

La elección de arquitectura depende del control que necesitemos inspeccionar. Una cadena modular facilita observar etapas y sustituir componentes; una solución integrada puede coordinar aspectos del audio dentro de un mismo sistema. Esto no establece que una produzca mejor atmósfera por definición. El prototipo debe comparar implementaciones concretas y registrar sus configuraciones.

5.3. Configuración del comportamiento conversacional

ElevenAgents permite configurar interrupciones y una disposición a responder mediante Eager, Normal y Patient. El modo expresivo documenta control de entrega con instrucciones y etiquetas, cuyo efecto debe comprobarse por voz e idioma. Estos controles pueden servir a una dirección conversacional, pero una risa o un suspiro sintético no demuestran empatía. [31, 32]

OpenAI distingue server VAD, basado en actividad y silencio, de semantic VAD, que estima si el usuario terminó a partir de lo dicho. La configuración permite ajustar eagerness y el comportamiento de respuesta e interrupción. La disponibilidad de esos parámetros no elimina la necesidad de probar relatos, vacilaciones y acentos reales. [33]

Propuesta de diseño. Elegir la forma de decir después de establecer qué función cumple el turno. Una aclaración puede requerir una voz sobria y breve; un relato puede necesitar variación prosódica. No utilizar expresividad como relleno que disimule una respuesta pobre o una herramienta todavía pendiente.

5.4. Consistencia entre generación, reproducción y estado

La documentación de OpenAI diferencia audio generado y audio reproducido cuando hay una interrupción. WebRTC y SIP permiten truncar automáticamente el tramo no reproducido; en WebSocket la aplicación debe detener reproducción y comunicar el punto correspondiente mediante `conversation.item.truncate`. [34]

Nuestra implicación de diseño es más amplia: el historial compartido no puede contener como escuchada una explicación que fue cortada. Lo mismo se aplica a una imagen que nunca se mostró o a una nota cuyo guardado falló. El sistema necesita estados de preparación, presentación y confirmación para no construir continuidad sobre acontecimientos inexistentes.

Realtime también documenta herramientas ejecutadas por la aplicación. La continuidad con operaciones asíncronas permite seguir conversando mientras llega un resultado, pero no autoriza a afirmarlo antes de recibirlo. El plan de producto debe distinguir pendiente, disponible, presentado, aceptado y cancelado. [35, 36]

6. Arquitectura propuesta para la conversación situada

6.1. Espacios de atención

Weiser y Brown propusieron diseñar tecnologías capaces de pasar entre el centro y la periferia de la atención. Ishii y Ullmer exploraron objetos, superficies y medios ambientales como formas de relacionar información digital y actividad situada. Son antecedentes de diseño que permiten pensar una interfaz conversacional más amplia que un historial de mensajes. No son pruebas de eficacia para un agente actual. [37, 38]

Proponemos tres espacios complementarios. El primer plano contiene aquello que se está diciendo o examinando. Un espacio lateral mantiene objetos disponibles sin exigir una respuesta inmediata. Un fondo de trabajo prepara resultados, mantiene contexto temporal y registra operaciones pendientes. La persona debe poder comprender qué se conserva y traer al primer plano lo que necesite.

El segundo plano visual no debe convertirse en almacenamiento opaco. Que una nota no interrumpa la conversación no implica que deba guardarse de forma permanente y silenciosa. La interfaz puede mostrar una señal discreta de preparación, permitir revisar una propuesta al cierre y conservar solo lo que el usuario haya acordado.

6.2. Objetos visuales y actividad compartida

La política de presentación puede formularse como una secuencia. Primero identificar para qué serviría el objeto: comparar dos formas, señalar un detalle, situar un recuerdo o revisar una decisión. Después comprobar si está disponible, si su procedencia es adecuada y si la persona quiere verlo. Finalmente elegir un punto de entrada que no compita con un turno todavía abierto.

El agente puede decir «Tengo una referencia que puede servir para comparar esa curva; la dejo al lado cuando quieras». Si la persona cambia de tema, el objeto puede expirar sin haberse mostrado. Mostrar menos puede resultar mejor que interrumpir cada asociación verbal con contenido visual; esta es una hipótesis a comprobar, no una ley.

Las imágenes generadas requieren una distinción explícita entre representación y evidencia. Si se imagina una escena, debe presentarse como propuesta. Si se recupera una fotografía personal, debe respetarse el alcance de uso acordado. Ninguna similitud visual autoriza a presentar una imagen inventada como parte de la memoria real de alguien.

6.3. Presentación y circulación de objetos

Compartir puede significar poner un objeto dentro de la sesión, crear una pieza que la persona podrá descargar o enviarla fuera de esa sesión. Proponemos acciones y permisos diferentes para cada caso. Un archivo disponible no se envía automáticamente a otra persona. La conversación puede producir un borrador, dejarlo pendiente y permitir que el usuario decida su destinatario y momento.

El estado del objeto debería distinguir preparación, disponibilidad, presentación, referencia conjunta y persistencia. Si el usuario lo retira, la aplicación cancela futuras menciones programadas. Si lo guarda, conserva la procedencia y la versión que realmente vio. Un enlace que cambia después no equivale a una copia estable del objeto compartido.

6.4. Memoria, persistencia y revisión

Proponemos tres niveles. El contexto de turno conserva lo necesario para la intervención inmediata. La memoria de sesión sostiene temas abiertos y objetos temporales. La memoria entre sesiones conserva únicamente elementos autorizados con una finalidad comprensible. Estos niveles pueden implementar caducidades diferentes y no deben confundirse con la retención técnica de los proveedores.

Cada memoria persistente debería tener contenido, origen, fecha, alcance, estado de confirmación y mecanismo de revisión. También debe distinguir lo que la persona dijo de una interpretación del sistema. Ante una contradicción, el agente pregunta o actualiza; no obliga a que la nueva conversación encaje en el perfil anterior.

La política puede aceptar una autorización estable, como guardar notas de un proyecto, sin interrumpir para confirmar cada frase. Debe hacer visible ese alcance, ofrecer excepciones y permitir revertirlo. Para recuerdos personales fuera de ese acuerdo, una propuesta explícita de guardado evita convertir la escucha en un archivo automático.

6.5. Edición de situaciones conversacionales

La unidad de trabajo propuesta es una situación: qué quiere hacer la persona, qué actividad comparten, qué información falta, qué restricciones existen y cómo podría reconocerse un cierre. Cada situación admite varias continuaciones. El editor define recursos y condiciones; las palabras concretas permanecen abiertas a la interacción.

Un equipo podría describir una situación de exploración creativa con una pregunta disponible, dos referencias optativas, un lugar donde dejar ideas y una forma clara de terminar. También podría describir una gestión operativa con datos necesarios, comprobaciones y resultado. El sistema no necesita aplicar la misma intensidad expresiva a ambos casos.

6.6. Instrumentación temporal de la experiencia

Un visor debe alinear audio del usuario, audio del agente, decisiones de turno, operaciones, objetos y eventos de memoria. El diseñador puede escuchar un intervalo y comprobar qué lo produjo: voz todavía en preparación, espera elegida, herramienta lenta o cesión incorrecta. Las anotaciones interpretativas deben mantenerse separadas del registro técnico.

El visor necesita mostrar discrepancias entre lo generado y lo presentado. El tramo generado que no llegó a reproducirse puede permanecer en un registro de diagnóstico autorizado, pero no se registra como presentado. El tramo efectivamente reproducido conserva su lugar en el historial; su reproducción tampoco prueba comprensión. Un resultado tardío puede quedar descartado porque el tema cambió. Esto permite estudiar continuidad sin atribuir al usuario experiencias que no ocurrieron.

6.7. Coordinación de turnos y operaciones

Proponemos un coordinador por encima de los adaptadores de voz. Mantiene el propósito actual, la estimación de cesión del turno, los temas abiertos, los objetos disponibles, las operaciones pendientes y los permisos. Decide cuándo liberar una respuesta preparada y cuándo aplazar una imagen o una propuesta de guardado.

La ruta temporal inmediata debe poder funcionar sin esperar al trabajo lento. Si la persona interrumpe, se detiene el audio y se cancela o marca como obsoleta la presentación pendiente. Esto no revierte una operación externa ya ejecutada, cuyo estado debe comprobarse por separado. Si una herramienta sigue trabajando, su resultado se etiqueta con la versión del contexto que lo originó. Un resultado de una intención anterior no debe reaparecer como si siguiera vigente.

Las operaciones externas requieren una política adicional. Preparar una alternativa es reversible; enviar un mensaje o ejecutar una reserva produce efectos. La anticipación puede preparar datos, pero una acción con efectos debe pasar por los permisos y confirmaciones

Tabla 3. Contrato mínimo propuesto para el registro de acontecimientos.

Acontecimiento	Dato mínimo	Para qué sirve
Fin acústico del usuario	Tiempo y confianza de anotación	Medir intervalos sin depender del detector
Cesión de turno estimada	Tiempo, configuración y probabilidad si existe	Explicar la decisión e identificar estimaciones propias
Audio listo y reproducido	Dos marcas distintas y tramo audible	Separar capacidad técnica y experiencia
Interrupción y cese de audio	Inicio de usuario y fin de reproducción	Medir devolución efectiva del turno
Objeto listo y presentado	Identificador, versión y dos tiempos	Estudiar una presentación diferida
Memoria propuesta y guardada	Alcance, autorización y resultado	Distinguir intención de persistencia

Nota. Los campos no expuestos por un proveedor deben permanecer ausentes o identificarse como estimaciones propias.

que correspondan al producto. La generación especulativa no constituye autorización.

La separación entre coordinador, adaptadores de voz, presentación de objetos y memoria permite cambiar de proveedor sin redefinir todas las reglas de interacción. Esa portabilidad tendrá límites: cada adaptador debe declarar qué eventos y controles expone realmente, y el sistema no debe inventar datos ausentes para completar su registro.

6.8. Contrato mínimo de acontecimientos

Usar relojes monotónicos para duraciones locales y documentar la sincronización entre cliente y servidor. Guardar identificadores de sesión, turno y versión de modelo. Si el proveedor no expone una probabilidad, el campo queda ausente o se identifica como estimación propia; nunca se atribuye al proveedor una cifra reconstruida. No hace falta conservar todo el audio para todas las funciones: el protocolo de investigación y el producto deben definir por separado qué datos necesitan y durante cuánto tiempo.

6.9. Biblioteca de situaciones de evaluación

La biblioteca de evaluación debería incluir una petición breve, una persona que busca una palabra, un relato con pausa interna, una corrección, una interrupción durante una explicación, una herramienta lenta, un cambio de tema y una petición de olvidar. Añadir condiciones de ruido, acentos, velocidades de habla y uso sin pantalla. No interpretar automáticamente una disfluencia o un acento como dificultad de comprensión.

Cada situación necesita un criterio observable. Por ejemplo: permitir completar una frase sin invasión; detener una respuesta al ser interrumpido; no afirmar que una nota se guardó si hubo un error; recordar solo la versión confirmada. Los criterios se pueden automatizar parcialmente, pero las valoraciones de atmósfera necesitan investigación con personas.

La tabla 3 resume los acontecimientos necesarios para instrumentar esta arquitectura.

7. Aplicación ilustrativa

El siguiente ejemplo es original y ficticio. Ilustra decisiones de diseño y no reproduce diálogo de las películas ni resultados de un estudio. La persona está explorando una idea para una instalación y ha acordado un modo de conversación pausado. Existe un espacio lateral de notas temporales; todavía no ha autorizado guardar contenido entre sesiones.

La persona dice: «Quiero que se sienta como volver a un lugar, pero todavía no sé...». La frase y la prosodia sugieren una continuación. El agente mantiene la escucha aunque exista un intervalo sin voz. No presenta imágenes ni completa la frase por cuenta de la persona.

La persona continúa: «...como si el lugar se acordara de ti». El sistema prepara una respuesta breve y la libera cuando estima una cesión. «¿Qué podría reconocer de ti sin tener que explicarlo todo?». La pregunta ofrece un ángulo de exploración y permite una respuesta distinta de la prevista.

Mientras la persona desarrolla la idea, el sistema puede preparar dos referencias visuales autorizadas. Las conserva como disponibles, sin mostrarlas todavía. Cuando el usuario pide ejemplos, ofrece una imagen a la vez. Si el usuario señala un detalle, la referencia pasa a formar parte de la conversación. Si la ignora, el sistema no presume comprensión ni insiste por el hecho de haberla producido.

Más tarde, el agente propone una nota temporal: «Hay una idea que parece útil: reconocer sin pedir una explicación. La puedo dejar escrita aquí». Si el usuario acepta, aparece en la sesión. Al cerrar, puede ofrecer guardarla para el proyecto. La confirmación de guardado se produce después de que la operación haya concluido, con posibilidad de editar o borrar.

La escena también debe funcionar si la persona dice «No, no era eso». El agente retira su interpretación y pide precisar solo lo necesario. No defiende su metáfora ni intenta recuperar a toda costa una atmósfera poética. La continuidad consiste en seguir lo que la persona está construyendo.

En una tarea operativa el mismo coordinador tendría otra política. Ante «¿A qué hora empieza la sesión?», ofrecería el dato verificado de forma directa. La calidad del diseño consiste en reconocer esa diferencia de propósito, sin convertir todas las conversaciones en una experiencia contemplativa.

8. Diseño de investigación propuesto

8.1. Construcción del corpus cinematográfico

La siguiente fase debería identificar las ediciones de las tres películas y anotar secuencias completas. Se propone comenzar con cabina de discos y café en Sunrise;

café, coche y transición al apartamento en Sunset; coche, comida y hotel en Midnight. Es una selección analítica inicial, no una muestra representativa de toda la trilogía.

Las capas observables incluyen habla por participante, intervalos, solapamientos, gestos visibles, cambios de plano, desplazamientos, sonido y objetos. Otra capa recoge interpretaciones: función del silencio, giro de tema, apertura, exposición y reparación. El protocolo prevé que dos codificadores trabajen con un manual común, comparen acuerdos y conserven discrepancias. Los tiempos se anotarían sobre la edición identificada, considerando cortes y posibles diferencias de velocidad de reproducción.

El resultado será un catálogo de recursos y preguntas. No se usarán los intervalos de las películas como distribución normativa del habla humana, porque el guion, la actuación, la cámara y el montaje contribuyen al resultado.

8.2. Estudios de percepción e interacción

Un primer estudio puede usar escenas originales grabadas para el proyecto. Se mantienen palabras, voz, prosodia y objetos, modificando únicamente ciertos intervalos. Los observadores evalúan adecuación, continuidad e incomodidad. La asignación de versiones es aleatoria. Así se investiga la percepción de una escena sin atribuir a los oyentes una experiencia de interacción que no tuvieron.

Un segundo estudio requiere participantes que conversen realmente con un agente. Manteniendo modelo, voz, instrucciones y herramientas, se introducen demoras adicionales asignadas al azar después de que una respuesta esté disponible. Debe registrarse tanto la condición asignada como la demora efectiva. La hipótesis pregunta si el efecto del tiempo depende de la tarea, no si un número de milisegundos gana siempre.

8.3. Comparación de políticas y modalidades

La siguiente comparación enfrenta una política fija y otra contextual de cesión de turno. Para separar adaptación y rapidez, conviene añadir una condición con distribución temporal global emparejada. Si la política contextual es simplemente más rápida en todos los casos, no podremos atribuir la diferencia a que entiende mejor el momento.

Los objetos visuales se estudian después, con política temporal estable: audio sin objeto; el mismo objeto pertinente presentado en un punto acordado; y ese objeto disponible al cierre. La memoria se estudia en sesiones repetidas: contexto de sesión frente a recuerdos confirmados entre sesiones, con edición y borrado en ambos casos donde correspondan. Cambiar simultáneamente voz, latencia, personalidad, imágenes y memoria impediría identificar qué produjo un efecto.

La comparación entre proveedores se realiza como evaluación de sistemas completos, con configuraciones documentadas. Cambiar de proveedor puede cambiar voz, reconocimiento y contenido, por lo que no es un experimento limpio del efecto de latencia. Deben mantenerse separados esos dos objetivos de investigación.

8.4. Hipótesis y condiciones de refutación

La investigación debe admitir que una parte de la propuesta no funcione. Una evaluación cuyo único resultado aceptable sea «más atmósfera» no permitiría distinguir diseño de confirmación de una intuición.

Una estimación imprecisa o un resultado no significativo no demuestran ausencia de un efecto relevante. La tabla 4 debe interpretarse respecto de una mejora mínima especificada antes del análisis. Si se pretende concluir equivalencia, sus márgenes y el contraste correspondiente se definirán previamente; los datos también pueden resultar inconcluyentes. [39]

8.5. Variables de evaluación y análisis

Cada estudio deberá fijar un contraste principal, su población, tarea, horizonte y una variable primaria antes de recoger los datos confirmatorios. Para la hipótesis de toma de turno se propone estimar la diferencia entre políticas en probabilidad de inicio prematuro, expresada en puntos porcentuales. Éxito del propósito, control, esfuerzo, comprensión y continuidad se incluirán según un plan explícito de medidas secundarias o de multiplicidad. Las valoraciones de atmósfera seguirán siendo exploratorias hasta estudiar su fiabilidad y estructura.

Se registrarán los intervalos, la detención tras interrupción y los fallos de herramientas, con dispositivo, red, idioma, ruido y configuración. Se informarán denominadores, pérdidas, exclusiones y respuestas no observadas; los turnos lentos no se excluirán por empeorar la apariencia del rendimiento. Los percentiles temporales requerirán suficientes eventos y un método de incertidumbre declarado. Los turnos de una persona o conversación no se tratarán como participantes independientes.

El piloto permitirá revisar la instrumentación y estimar variabilidad. El tamaño confirmatorio se justificará mediante una diferencia mínima relevante y un análisis de potencia o precisión que respete las medidas repetidas. El plan especificará unidad de aleatorización, contrabalanceo, efectos de orden, regla de parada y manejo de la dependencia entre personas, conversaciones y estímulos. La familia estadística deberá corresponder a la variable: por ejemplo, eventos binarios u observaciones ordinales. Se presentarán efectos e intervalos de incertidumbre, con análisis de sensibilidad predefinidos para datos faltantes y decisiones de codificación.

La duración de sesión, el número de turnos o la cantidad de información personal revelada no se proponen como indicadores primarios de éxito. Pueden aumentar por confusión o por preguntas innecesarias. Un cierre breve y satisfactorio puede ser un resultado excelente.

La tabla 4 explicita comparaciones que permiten revisar o rechazar las hipótesis.

8.6. Programa inicial de desarrollo

Proponemos una secuencia de tres entregables. El primero es un atlas de situaciones, con anotaciones del corpus y un vocabulario de acciones. El segundo es un prototipo instrumentado que permita cambiar políticas temporales, mostrar objetos optativos y gestionar me-

Tabla 4. Hipótesis de investigación, comparaciones y condiciones que las debilitarían.

Hipótesis propuesta	Comparación	Evidencia que la debilitaría
Una política contextual reduce entradas invasivas	Política fija y contextual con tiempos comparables	Estimación precisa que excluya la mejora mínima relevante, o evidencia de deterioro
Un objeto pertinente puede facilitar comprensión	Objeto acordado, objeto al cierre y audio solo	Distracción adicional y ausencia de mejora relevante estimada con precisión
La memoria confirmada puede sostener continuidad	Varias sesiones con y sin recuerdos persistentes	Errores, irrelevancia o menor control percibido
La rapidez funciona de forma distinta según contexto	Tarea práctica y exploración con demoras asignadas	Interacción estimada con precisión incompatible con el efecto mínimo propuesto

Nota. Hipótesis propuestas; el artículo no presenta resultados experimentales propios.

moria de sesión. El tercero es un piloto con personas que compruebe si esas decisiones producen diferencias y dónde fallan.

Para iniciar el prototipo, bastan dos contextos: una tarea práctica con respuesta comprobable y una exploración creativa breve. Un motor de voz, un coordinador de acciones y un registro de acontecimientos son suficientes para poner a prueba las hipótesis principales. La memoria entre sesiones puede incorporarse después de verificar control de turnos y claridad de persistencia.

Un equipo inicial podría combinar dirección de experiencia, ingeniería de voz e interacción, y apoyo de investigación en conversación o HCI. La aportación prevista del proyecto residiría en el corpus por construir, la política de coordinación, el editor, la instrumentación y la evidencia que produzca su evaluación. La elección de proveedor es una parte de esa construcción.

El criterio para ampliar el desarrollo debería ser una mejora observable respecto de una base sencilla: menos invasión, mejor comprensión o continuidad útil sin reducir control. Si un módulo visual distrae o una memoria no aporta valor, se retira o revisa. La propuesta debe poder demostrar utilidad sin depender de una demostración especialmente preparada.

8.7. Prerregistro y materiales reproducibles

Antes del estudio confirmatorio se propone registrar hipótesis, variable principal, diferencia mínima relevante, muestra, asignación, exclusiones, parada y plan de análisis. Los cambios posteriores se fecharán y se distinguirán los análisis previstos de los exploratorios. Este procedimiento favorece la trazabilidad de las decisiones; no impide explorar resultados inesperados. No existe todavía un prerregistro de los experimentos propuestos en este artículo. [40]

Los materiales deberán fijar instrucciones del agente, voz, identificador y versión de modelo, parámetros de turno, transporte, región y configuración de herramientas. El manual de anotación definirá inicio, final, respuesta breve, interrupción y caso indeterminado, con ejemplos y procedimiento de resolución. La evaluación humana documentará quién conoce cada condición; se reconocerá cuando la modalidad hace imposible ocultarla. Datos, código y materiales se compartirán en la medida compatible con los acuerdos de participación, con versiones y cambios documentados.

9. Discusión y limitaciones

Este documento no cuantifica todavía las escenas, no ejecuta pruebas de proveedores y no presenta un instrumento validado de atmósfera. La revisión prioriza fuentes primarias e investigación directamente pertinente, pero no agota el análisis conversacional, los estudios de cine ni la interacción humano máquina. Las cifras comerciales proceden de documentación cambiante y requieren nuevas comprobaciones al implementar.

Disponibilidad actual. El proyecto LaTeX incorpora las referencias, el código y la figura ilustrativa, junto con un registro estructurado de las cifras centrales. Ese registro contiene estadísticos transcritos y sus condiciones conocidas; no es un conjunto de datos experimentales. Divina Machina no ha producido en este trabajo mediciones de latencia, anotaciones temporales del corpus ni resultados con participantes. La reproducción de la maquetación es posible con los archivos entregados; la reproducción independiente de cada benchmark requeriría registros y procedimientos que sus fuentes no siempre publican.

La transferencia desde una pareja de ficción hacia una relación con un agente tiene límites estructurales. No existe reciprocidad humana equivalente; una voz expresiva puede producir atribuciones que el sistema no puede sostener. Por eso se propone claridad sobre el carácter artificial del agente, capacidad efectiva de salida y una política de memoria comprensible. La estética no sustituye la exactitud ni el cumplimiento de lo acordado.

También debemos estudiar diferencias de accesibilidad y contexto. Una pausa puede ayudar a una persona y generar incertidumbre en otra; una imagen no aporta lo mismo con y sin pantalla. Las preferencias expresadas y los controles disponibles tienen que poder prevalecer sobre las estimaciones automáticas. La variación no es un ruido que deba eliminarse del diseño.

10. Conclusiones

Nuestra lectura de la trilogía *Before* sitúa la conversación en el tiempo compartido: aquello que se dice adquiere sentido junto con lo que acaba de ocurrir y con lo que todavía puede suceder. Esa lectura orienta una propuesta técnica que distingue disponibilidad, intervención, presentación y persistencia. Su utilidad de-

berá establecerse mediante las comparaciones descritas en este artículo.

Podemos hacer explícitas las reglas con las que un sistema toma la palabra, ofrece un objeto, reconoce una corrección o conserva un recuerdo. También podemos medir sus errores y preguntar a las personas cómo afectaron a la experiencia. La precisión del registro permite revisar el diseño sin atribuir a las mediciones un significado que no demuestran.

El criterio propuesto es que la intervención contribuya al propósito de quien conversa y conserve su posibilidad de cambiarlo. Una respuesta, una imagen o una nota pueden dejar algo disponible para después. En otras ocasiones, el trabajo del sistema habrá terminado cuando la persona encuentre sus propias palabras y ya no necesite otra pregunta.

Firman. Cris Olmedo y Kurt Sommer, Grupo de investigación Divina Machina. *Cómo citar*: Olmedo, C. y Sommer, K. (2026). El arte de la conversación. Documento de trabajo DM-2026-01, v 0.2. Grupo de investigación Divina Machina. Divina Machina by Qs Ventures.

Referencias

Las fuentes digitales se consultaron el 29 de octubre de 2026. Los enlaces de películas verifican créditos y contexto; los detalles de puesta en escena se apoyan en entrevistas y se distinguen de nuestras lecturas. Las páginas de documentación sin fecha se citan como documentación vigente consultada. Las declaraciones comerciales no constituyen una evaluación independiente realizada por este proyecto.

[1] R. Linklater. Before Sunrise. The Criterion Collection, ficha y créditos, 1995. URL https://www.criterion.com/films/28692-before-sunrise. Película. Dirección de R. Linklater; guion de R. Linklater y K. Krizan.

[2] R. Linklater. Before Sunset. The Criterion Collection, ficha y créditos, 2004. URL https://www.criterion.com/films/28693-before-sunset. Película. Dirección de R. Linklater; guion de R. Linklater, J. Delpy y E. Hawke; historia de R. Linklater y K. Krizan.

[3] R. Linklater. Before Midnight. The Criterion Collection, ficha y créditos, 2013. URL https://www.criterion.com/films/28828-before-midnight. Película. Dirección de R. Linklater; guion de R. Linklater, J. Delpy y E. Hawke.

[4] R. K. Osenlund. Interview: Julie Delpy and Richard Linklater on Before Midnight. Slant Magazine, May 2013. URL https://www.slantmagazine.com/features/interview-julie-delpy-and-richard-linklater/. Fecha: 23 de mayo de 2013. Entrevista primaria.

[5] S. Bland. Richard Linklater on Before Sunrise's enduring legacy. Huck, September 2020. URL https://www.huckmag.com/article/richard-linklater-on-before-sunrises-enduring-legacy. Fecha: 28 de septiembre de 2020. Entrevista primaria.

[6] Phil Hoad. Julie Delpy and Ethan Hawke: how we made the Before Sunrise trilogy. The Guardian, November 2019. URL https://www.theguardian.com/film/2019/nov/04/julie-delpy-ethan-hawke-how-we-made-before-sunrise-trilogy-sunset-midnight. Fecha: 4 de noviembre de 2019. Entrevistas. Actualizado el 6 de noviembre de 2019.

[7] N. Minow. Interview with Richard Linklater about Before Sunset. More about Movies and More, July 2004. URL https://nellminow.blogspot.com/2004/07/interview-with-richard-linklater-about.html. Fecha: 8 de julio de 2004. Entrevista primaria.

[8] Time Out. TONY Q and A: Before Midnight's Richard Linklater, Ethan Hawke and Julie Delpy. Time Out, May 2013. URL https://www.timeout.com/film/tony-q-a-before-midnights-richard-linklater-ethan-hawke-and-julie-delpy. Fecha: 21 de mayo de 2013. Entrevista primaria.

[9] M. Koresky and J. Reichert. A conversation with Richard Linklater. Reverse Shot, July 2004. URL https://reverseshot.org/symposiums/entry/204/richard-linklater. Fecha: 2 de julio de 2004. Entrevista primaria.

[10] P. Sobczynski. Interview: André Gregory and Wallace Shawn. RogerEbert.com, June 2015. URL https://www.rogerebert.com/interviews/interview-andre-gregory-and-wallace-shawn. Fecha: 12 de junio de 2015. Entrevista primaria.

[11] Ł. Mańkowski. On the road with Hamaguchi Ryūsuke. BFI Sight and Sound, November 2021. URL https://www.bfi.org.uk/sight-and-sound/interviews/hamaguchi-ryusuke-drive-my-car. Fecha: 16 de noviembre de 2021. Entrevista primaria.

[12] C. Aguilar. Embodiments of time: Kogonada on After Yang. RogerEbert.com, March 2022. URL https://www.rogerebert.com/interviews/embodiments-of-time-kogonada-on-after-yang. Fecha: 3 de marzo de 2022. Entrevista primaria.

[13] H. Sacks, E. A. Schegloff, and G. Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735, 1974. URL https://www.conversationanalysis.org/schegloff-media-archive/simplest-systematics-for-turn-taking-language-1974/. Acceso mediante el archivo de la International Society for Conversation Analysis.

[14] 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/.

[15] 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.

[16] 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/.

[17] E. M. Templeton, L. J. Chang, E. A. Reynolds, M. D. Cone LeBeaumont, and T. Wheatley. Long gaps between turns are awkward for strangers but not for friends. *Philosophical Transactions of the Royal Society B*, 378:20210471, 2023. doi: 10.1098/rstb.2021.0471. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC9985966/.

[18] H. H. Clark and S. E. Brennan. Grounding in communication. In *Perspectives on Socially Shared Cognition*, pages 127–149. American Psychological Association, 1991. URL https://web.stanford.edu/~clark/1990s/Clark%2C%20H.H.%20_%20Brennan%2C%20S.E.%20_Grounding%20in%20communication_%201991.pdf. Copia de autor en Stanford.

[19] C. Goodwin. Action and embodiment within situated human interaction. *Journal of Pragmatics*, 32(10):1489–1522, 2000. doi: 10.1016/S0378-2166(99)00096-X. URL https://www.sciencedirect.com/science/article/pii/S037821669900096X.

[20] E. Ekstedt and G. Skantze. Voice Activity Projection: Self-supervised learning of turn-taking events. In *Interspeech 2022*, pages 5190–5194, 2022. doi: 10.21437/Interspeech.2022-10955. URL https://www.isca-archive.org/interspeech_2022/ekstedt22_interspeech.html.

[21] Barry N. Taylor and Chris E. Kuyatt. Guidelines for evaluating and expressing the uncertainty of NIST measurement results. Technical Note 1297, National Institute of Standards and Technology, 1994. URL https://www.nist.gov/pml/nist-technical-note-1297/nist-tn-1297-appendix-d1-terminology. Edición de 1994. Apéndice D.1: Terminology, versión web.

[22] M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella. PARADISE: A framework for evaluating spoken dialogue agents. In *ACL EACL*, pages 271–280, 1997. doi: 10.3115/976909.979652. URL https://aclanthology.org/P97-1035/.

[23] 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.

[24] ElevenLabs. Models. Documentación de ElevenLabs, s. f. URL https://elevenlabs.io/docs/overview/models. Modelos, identificadores y latencias de inferencia.

[25] ElevenLabs. Understanding latency. Documentación de ElevenLabs, s. f. URL https://elevenlabs.io/docs/eleven-api/concepts/latency. Inferencia, primer audio y etapas del sistema.

[26] 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.

[27] 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.

[28] M. Kim. Introducing Eleven Realtime: a simpler way to build natural voice agents. ElevenLabs, October 2026. URL https://elevenlabs.io/blog/realtime. Fecha: 20 de octubre de 2026. Informe comercial de lanzamiento; benchmarks atribuidos a Artificial Analysis.

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

[30] OpenAI. GPT-Realtime-2.1 Model. Documentación de OpenAI, s. f. URL https://developers.openai.com/api/docs/models/gpt-realtime-2.1. Ficha del modelo gpt-realtime-2.1, modalidades y esfuerzo de razonamiento.

[31] ElevenLabs. Conversation flow. Documentación de ElevenLabs, s. f. URL https://elevenlabs.io/docs/eleven-agents/customization/conversation-flow. Tiempos, interrupciones y turn eagerness.

[32] ElevenLabs. Expressive mode. Documentación de ElevenLabs, s. f. URL https://elevenlabs.io/docs/eleven-agents/customization/voice/expressive-mode. Control expresivo y variación de resultados.

[33] OpenAI. Voice activity detection. Documentación de OpenAI, s. f. URL https://developers.openai.com/api/docs/guides/realtime-vad. Server VAD y semantic VAD.

[34] OpenAI. Realtime conversations. Documentación de OpenAI, s. f. URL https://developers.openai.com/api/docs/guides/realtime-conversations. Gestión de conversación, interrupciones y truncación de audio.

[35] OpenAI. Realtime with tools. Documentación de OpenAI, s. f. URL https://developers.openai.com/api/docs/guides/realtime-mcp. Funciones y herramientas en Realtime.

[36] OpenAI. Developer notes on the Realtime API. OpenAI, s. f. URL https://developers.openai.com/blog/realtime-api. Referencia utilizada para llamadas asíncronas.

[37] M. Weiser and J. S. Brown. Designing calm technology. Xerox PARC, December 1995. URL https://calmtech.com/papers/designing-calm-technology. Fecha: 21 de diciembre de 1995. Reproducción del texto original en Calmtech.com.

[38] H. Ishii and B. Ullmer. Tangible Bits: Towards seamless interfaces between people, bits and atoms. In *CHI 97*. ACM, 1997. URL https://chi1997.acm.org/proceedings/paper/hi.html. Actas digitales.

[39] Daniël Lakens. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362, 2017. doi: 10.1177/1948550617697177. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC5502906/.

[40] Center for Open Science. Preregistration. Center for Open Science, s. f. URL https://www.cos.io/initiatives/prereg.

Afinar un personaje en la sala

Un método de iteración en público para agentes de voz, con 80 versiones en tres eventos

Cris Olmedo y Kurt Sommer
Grupo de investigación Divina Machina
Documento de trabajo · Método
DM-2026-02, versión 1.0 · 30 de octubre de 2026

Resumen. Casi toda la IA conversacional se diseña en una mesa y se prueba con un guion. Un personaje que contesta una cabina telefónica pública no puede hacerse así: quien descuelga no tiene tarea ni paciencia y decide en un segundo si la voz al otro lado está viva. Describimos el método con el que afinamos a GAIA 2147: cambiar una sola cosa con la sala abierta, medir las llamadas siguientes y quedarse o deshacer.

En tres eventos salieron 80 versiones del agente, con 13 min (IC 95 % 7–25 min) de mediana entre una y la siguiente; el 74 % (IC 95 % 63–83 %) salió menos de una hora después de la anterior. En Barcelona Design Week el ritmo fue de 10,0 versiones al día y la versión mediana atendió 4 llamadas antes de ser sustituida. Damos el ciclo, la instrumentación que lo hace posible, seis reglas que salieron de medir y lo que el método no puede hacer.

Iterar en público cambia control experimental por relevancia. Cada comparación de la serie DM-2026 es observacional, y este documento propone las piezas que faltan para que deje de serlo.

Palabras clave: diseño de personajes; agentes de voz; iteración; evaluación en campo; instrumentación; prompts.

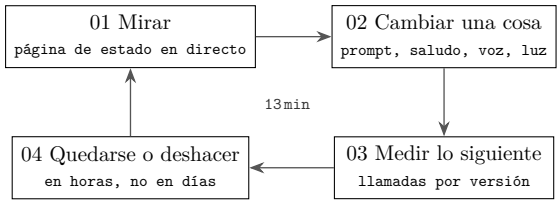


Figura 1. El ciclo de afinado. En el centro, el tiempo mediano por vuelta.

1. La tesis

No tratamos el prompt como un guion. Lo tratamos como el ajuste de un instrumento, y la sala como el lugar donde se afina [INT]. Cambiamos el personaje mientras los visitantes lo usan, medimos cada cambio en las llamadas que vienen después y nos quedamos con lo que funciona. Los papers DM-2026-03 a DM-2026-08 son los resultados de ese ciclo; DM-2026-09 describe los datos [1].

2. El ciclo

El ciclo tiene cuatro pasos (figura 1). Mirar: alguien está junto a la cabina y una página de estado muestra cada llamada. Cambiar una sola cosa: una línea del prompt, el saludo, un ajuste de voz, una luz. Medir: cada llamada registra la versión que la atendió, así que el efecto se lee en las llamadas siguientes. Quedarse o deshacer: un cambio se queda si mueve la medida que quería mover. Es rápido porque un error cuesta poco: una versión que falla está en vivo unos minutos y toca un puñado de llamadas [INT].

La tabla 1 y la figura 2 dan el ritmo [MED]. Solo el 24 de octubre el personaje cambió 23 veces. En IFA el ritmo fue mucho más lento (74 min (IC 95 % 13–205 min)), un evento con menos acceso a la cabina.

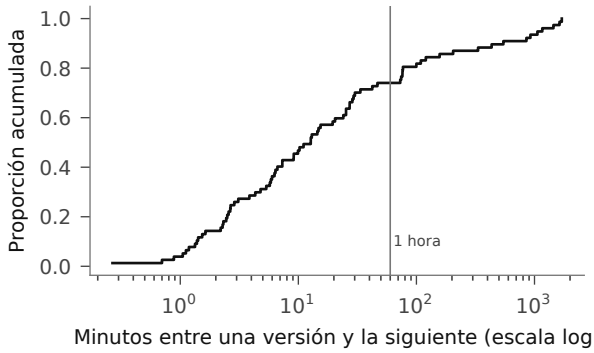


Figura 2. Distribución acumulada del tiempo entre versiones, en los tres eventos.

3. La cabina es un instrumento

El ciclo necesita una cabina que se mida a sí misma. Cada llamada registra versión, duración, cierre y, en cada turno, el tiempo de cada etapa de la respuesta. La cabina informa de su estado a una página en directo; algunos ajustes cambian desde la nube sin tocarla; las actualizaciones esperan a que se cuelgue. Dos plataformas, ElevenLabs y Vidiv, llevan el mismo personaje, y si una falla la cabina pasa a la otra (DM-2026-07). Nada de esto es ingeniería rara; lo raro es usarla en una obra en un pabellón público con la sala abierta [INT].

4. Seis reglas que salieron de medir

5. Lo que no hacemos

No guardamos las palabras de los visitantes: el análisis lee las transcripciones en memoria y conserva recuentos. No recordamos a nadie de una llamada a otra. No reiniciamos una cabina que pueda estar en una llamada. No publicamos una cifra sin decir de dónde sale: cada número de la serie sale de un script y lleva su intervalo. No inventamos ni redondeamos al alza.

Tabla 1. Ritmo de iteración por evento. Intervalos del 95 % por bootstrap.

Evento	Versiones	Días	Por día	Min. entre versiones [IC]	Llamadas por versión
Sónar+D	36	4	9,0	15 [9; 27]	5
IFA	14	4	3,5	74 [13; 205]	12
BDW	30	3	10,0	6 [3; 10]	4

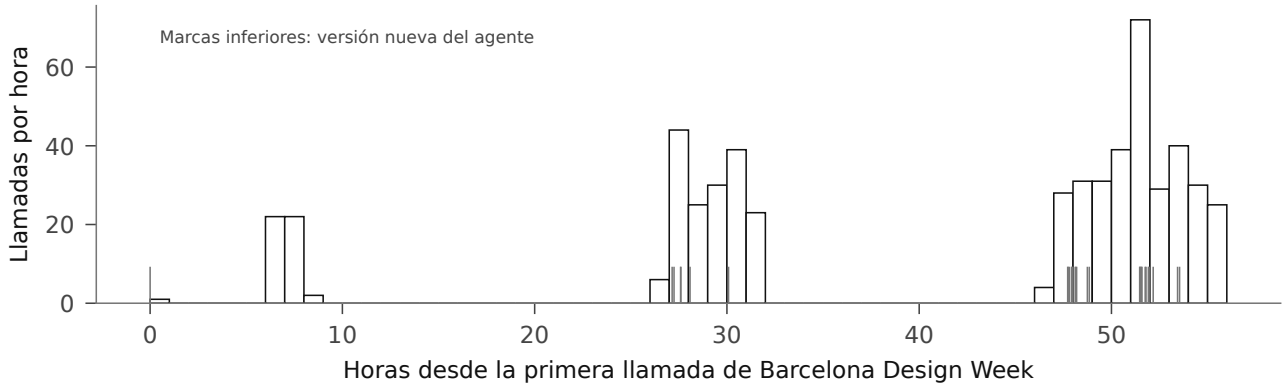


Figura 3. Barcelona Design Week: llamadas por hora y momento de cada versión nueva del agente.

6. Lo que el método no puede hacer

Iterar rápido y en público cambia rigor por relevancia. Cambian varias cosas a la vez, el público cambia a lo largo del día y no hay dos eventos iguales. Todas nuestras comparaciones son observacionales [INT]. La respuesta no es ir más despacio: es añadir al ciclo las piezas que faltan [HIP].

1. Una versión de control que funcione sin cambios en franjas alternas, para tener un contrafactual en la misma sala.
2. Asignar las variantes al azar por llamada cuando el cambio lo permita, como ya hacemos con el saludo corto.
3. Una hoja de observación por hora rellenada en la cabina.
4. El mismo límite de tiempo en cada evento.
5. Prerregistrar la medida principal antes de cada evento [2].

7. Recomendaciones para la comunidad

Para quien construya personajes de voz en espacios públicos: versionar cada cambio y guardar la versión con cada llamada; medir en horas; elegir antes la medida que el cambio debe mover; mirar dos veces, porque el primer resultado no es el estable; y separar lo que el modelo decide de lo que hace el dispositivo.

8. Datos y código

Dataset gaia-2147 [1]; cifras de `analysis/analyze.py`.

9. Conclusiones

Un personaje de voz no se escribe: se afina. Ochenta versiones en tres eventos no son desorden; son el instrumento sonando con público. Lo que queda por hacer es convertir ese ritmo en experimentos, sin perder la sala.

Firman. Cris Olmedo y Kurt Sommer, Grupo de investigación Divina Machina. *Cómo citar*: Olmedo, C. y Sommer, K. (2026). Afinar un personaje en la sala. Documento de trabajo DM-2026-02, v 1.0. Grupo de investigación Divina Machina. Divina Machina by Qs Ventures.

Referencias

Fuentes consultadas hasta el 29 de octubre de 2026.

[1] C. Olmedo and K. Sommer. GAIA-2147. descriptor de datos de 1527 conversaciones con un agente de voz en espacios públi-

Tabla 2. Reglas del método y el resultado que las sostiene. Detalle e incertidumbre en cada paper.

Regla	Resultado	Paper
Nunca empezar con silencio	Sin saludo, 71 % de llamadas cortas; con saludo corto, 7 %	DM-2026-04
Medir el silencio en la sala	Cascada 2,68 s, voz a voz 0,81 s de mediana en campo	DM-2026-03
Escribir la personalidad como hábitos contables	Una regla de tono redujo un 92 % las etiquetas sufridas, y se desgastó	DM-2026-05
El modelo pone el sentido, la sala la presencia	El modelo usó luz, pantalla o papel en el 5,2 % de las llamadas	DM-2026-06
Diseñar para quien pasa	En IFA, el 49 % de quienes descolgaron no habló	DM-2026-06
El borde lo que se toca, la nube el lenguaje	Fallos casi siempre físicos; nube fiable	DM-2026-07

cos. Documento de trabajo DM-2026-09. Grupo de investigación Divina Machina, 2026.

[2] Center for Open Science. Preregistration. Center for Open Science, s. f. URL https://www.cos.io/initiatives/prereg.

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 %

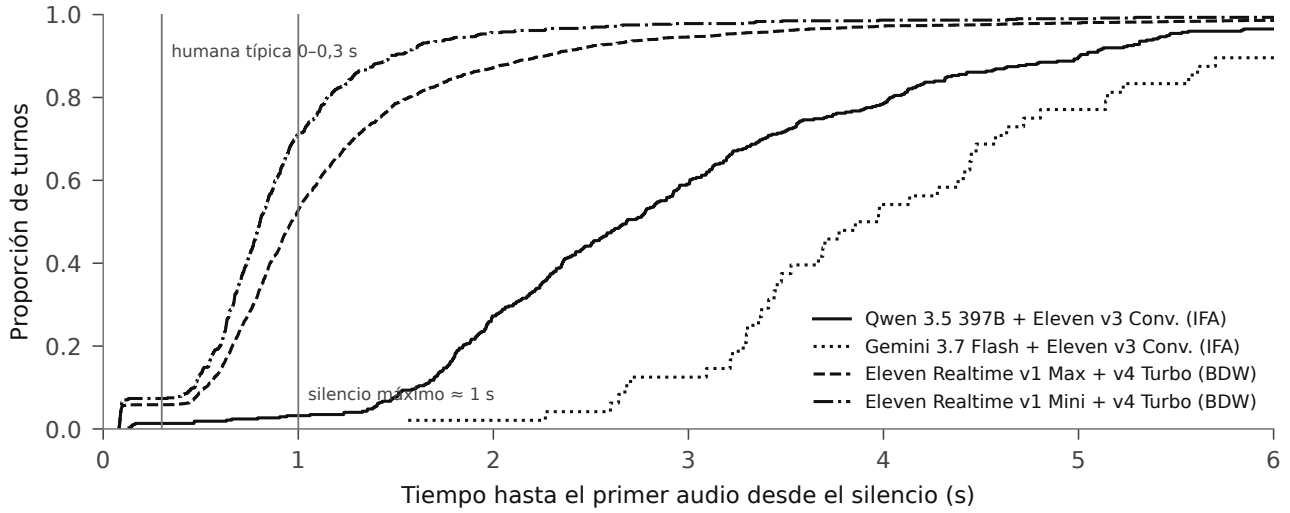


Figura 1. Distribución acumulada de la latencia por turno en cada pila. Las líneas verticales marcan la respuesta humana típica [2] y el silencio máximo estándar de Jefferson [3], como referencias, no como objetivos.

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]

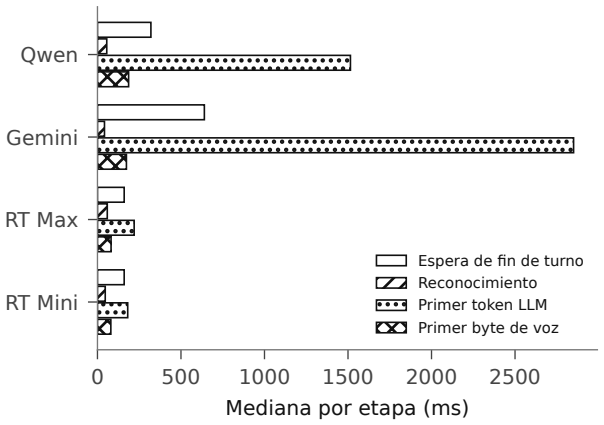


Figura 2. Mediana por etapa y pila. El primer token del modelo domina en las cascadas.

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

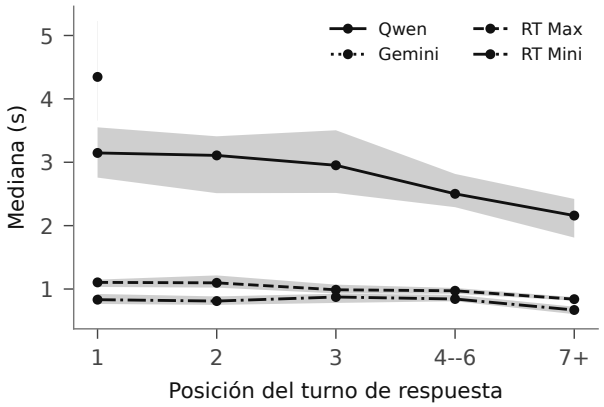


Figura 3. Mediana de latencia según la posición del turno de respuesta, con IC 95 % por bootstrap de turnos (banda gris). Se omiten posiciones con menos de 15 turnos.

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).

El primer segundo

Saludo, silencio y continuidad en llamadas a un agente de voz

Cris Olmedo y Kurt Sommer
Grupo de investigación Divina Machina
Documento de trabajo · Estudio observacional
DM-2026-04, versión 1.0 · 30 de octubre de 2026

Resumen. Un agente de voz en un teléfono público tiene un segundo para demostrar que la línea está viva. Estudiamos tres aperturas del mismo personaje, GAIA 2147, en 543 llamadas de Barcelona Design Week: ningún saludo, un saludo largo y fijo, y un saludo corto elegido al azar entre 8 variantes.

Sin saludo, el 71 % de las llamadas duró menos de diez segundos y en el 62 % el visitante no dijo nada. Con el saludo corto, esas cifras fueron el 7 % y el 6 %, frente al 14 % y el 24 % del saludo largo. Un análisis de supervivencia que trata como censuradas las llamadas cortadas por el límite de tiempo confirma el orden de las tres curvas. En el mismo evento, cuando la primera respuesta del agente llegó en dos segundos o menos, el visitante volvió a hablar en el 96 % de las llamadas, frente al 81 % cuando tardó más.

Leemos el resultado con la teoría de las aperturas telefónicas. Quien descuelga un teléfono público es quien llama, y espera que el otro lado hable primero. Las tres aperturas coinciden con cambios de prompt y de hora; el diseño es observacional y lo tratamos como tal.

Palabras clave: apertura de llamada; saludo; agentes de voz; análisis de supervivencia; toma de turnos; diseño conversacional.

1. Introducción

Las aperturas de las llamadas son de las secuencias mejor estudiadas de la conversación. Schegloff mostró que la primera intervención de una llamada responde a la llamada misma, el timbre, y que quien contesta habla primero [1]. Los agentes de voz heredan esa gramática sin haberla elegido. GAIA 2147 es un personaje que contesta en una cabina telefónica en espacios públicos. El visitante descuelga para ver qué pasa, sin tarea ni instrucciones.

En Barcelona Design Week cambiamos la apertura dos veces con la cabina abierta. Eso deja tres grupos con el mismo hardware, la misma pila de voz y el mismo evento. No es un experimento aleatorizado. Es la comparación más limpia que permite iterar en público [2].

1.1. Preguntas

1. ¿Cómo cambian el abandono temprano y el silencio del visitante con la apertura?
2. ¿Cambia la duración de la llamada cuando se trata el límite de tiempo como censura?
3. ¿Se asocia la rapidez de la primera respuesta con que el visitante vuelva a hablar?

2. Datos y método

2.1. Grupos

La clase de saludo se asigna comparando la primera frase de GAIA, si suena en los dos primeros segundos, con nuestros propios saludos [DEF]. Los grupos son *sin saludo* (24 llamadas), en el que GAIA espera a que hable el visitante; *saludo largo fijo* en español (313), que preguntaba si había alguien y cómo se llamaba; y *saludo corto variable* (140), una de 8 variantes como *¿Sí…?*, *Hello…?* o *Hallo…?*, sin repetir la anterior.

2.2. Variables

Llamada corta, menos de 10 s. *Llamada muda*, sin ningún turno transcrito del visitante. *Continuidad*, que el visitante produzca al menos dos turnos, entre las llamadas en que ya produjo uno. *Primera respuesta*, el primer turno de GAIA con latencia registrada tras el saludo [DEF].

2.3. Estadística

Proporciones con intervalos de Wilson [3] y diferencias con intervalos de Newcombe [4], contrastadas con la prueba exacta de Fisher y corregidas por Holm [5]. Las llamadas que alcanzan el límite de tiempo no terminan por decisión del visitante; tratarlas como duraciones completas sesga la mediana hacia abajo. Por eso estimamos la función de supervivencia de la llamada con Kaplan-Meier [6], con censura en el límite,

$$\hat{S}(t) = \prod_{t_i \le t} \left(1 - \frac{d_i}{n_i}\right), \tag{1}$$

donde d_i es el número de llamadas que terminan en t_i y n_i las que siguen abiertas. Las curvas se comparan con la prueba de rangos logarítmicos.

3. Resultados

3.1. Abandono y silencio

La tabla 1 y la figura 1 resumen los tres grupos [MED]. El silencio inicial fue la peor apertura con diferencia. La mitad de esas llamadas duró cero segundos: la gente descolgaba, no oía nada y colgaba.

Frente al saludo largo, no saludar aumentó las llamadas cortas en 56 pp (IC 95 % 36–71 pp) y las mudas en 39 pp (IC 95 % 18–55 pp). El saludo corto las redujo en −7 pp (IC 95 % −13 a −1 pp) y −18 pp (IC 95 % −23 a −11 pp) respectivamente (tabla 2). La reducción de llamadas mudas es la más grande y la más precisa: un saludo de una palabra le devuelve el turno al visitante.

Tabla 1. Resultados por apertura en Barcelona Design Week.

Saludo	n	Corta < 10 s	Muda	Mediana (s)	KM (s)	Límite
Sin saludo	24	71 % [51; 85]	62 % [43; 79]	0 [0; 9]	0	0 %
Saludo largo fijo	313	14 % [11; 19]	24 % [20; 29]	79 [63; 88]	79	32 %
Saludo corto variable	140	7 % [4; 13]	6 % [3; 12]	116 [94; 125]	116	48 %

Corchetes, IC 95 % (Wilson para proporciones, bootstrap para medianas). KM, mediana de Kaplan-Meier con las llamadas cortadas por el límite tratadas como censuradas.

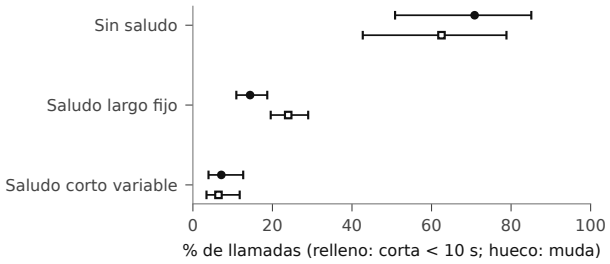


Figura 1. Proporción de llamadas cortas y mudas por apertura, con IC 95 % de Wilson.

Tabla 2. Diferencias de riesgo frente al saludo largo, en puntos porcentuales, con IC de Newcombe y p de Fisher ajustada por Holm.

Comparación	Resultado	Diferencia [IC 95 %]	p Holm
Sin saludo frente a largo	Llamada corta	56 pp [36; 71]	<0,001
Sin saludo frente a largo	Llamada muda	39 pp [18; 55]	<0,001
Saludo corto variable frente a largo	Llamada corta	−7 pp [−13; −1]	0,03
Saludo corto variable frente a largo	Llamada muda	−18 pp [−23; −11]	<0,001

3.2. Duración con censura

Con Kaplan-Meier, la mediana de duración fue 79 s con el saludo largo y 116 s con el corto; sin saludo fue 0 s (figura 2). Las curvas difieren frente al saludo largo tanto sin saludo (log-rank $\chi^2 = 112{,}6$, $p < 0{,}001$) como con el saludo corto ($\chi^2 = 11{,}7$, $p < 0{,}001$). El 48 % de las llamadas con saludo corto llegó al límite de tiempo, frente al 32 % con el largo: la duración real que la gente quería quedarse está por encima de lo que se observa.

3.3. Mismas horas

El saludo corto se usó en la tarde y noche del 24 de octubre; el largo, en los días anteriores y en las primeras horas de esa tarde. Restringido al 24 desde las 14 h, las llamadas cortas fueron el 7 % (IC 95 % 4–13 %) con el saludo corto (140 llamadas) y el 24 % (IC 95 % 17–33 %) con el largo (109), una diferencia de −17 pp (IC 95 % −26 a −8 pp). La dirección se mantiene en horas adyacentes del mismo día [MED], aunque no en las mismas horas.

3.4. Variantes

Las 8 variantes del saludo corto tienen entre 14 y 26 llamadas cada una (tabla 3). Hay heterogeneidad (chi cuadrado de homogeneidad, $p = 0{,}03$), concentrada en *¿Diga…?*, pero los intervalos se solapan casi por completo. No ordenamos las variantes; haría falta un estudio con más llamadas por variante [INT].

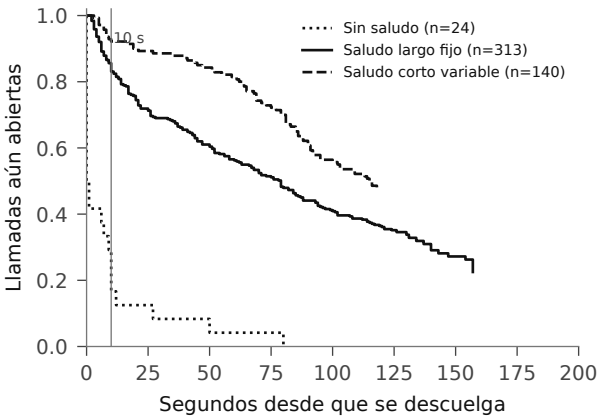


Figura 2. Supervivencia de la llamada por apertura (Kaplan-Meier). Las llamadas cortadas por el límite de tiempo se tratan como censuradas.

Tabla 3. Variantes del saludo corto. Corchetes, IC 95 % de Wilson.

Variante	n	Corta [IC 95 %]	Mediana (s)
¿Sí…?	26	4 % [1; 19]	92
¿Hola…? ¿Me oyes?	19	16 % [6; 38]	88
¿Hola…?	18	6 % [1; 26]	116
Hallo…?	17	0 % [0; 18]	125
¿Alguien…?	16	6 % [1; 28]	108
Hello…?	16	0 % [0; 19]	125
¿Diga…?	14	29 % [12; 55]	68
¿Hola? ¿Hay alguien ahí?	14	0 % [0; 22]	125

3.5. La primera respuesta

En Barcelona Design Week, cuando la primera respuesta de GAIA llegó en dos segundos o menos, el visitante volvió a hablar en el 96 % (IC 95 % 94–98 %) de las llamadas (359); cuando tardó más, en el 81 % (IC 95 % 69–89 %) (53), una diferencia de 15 pp (IC 95 % 6–28 pp) (Fisher, $p < 0{,}001$). En IFA casi todas las primeras respuestas fueron lentas; con 22 llamadas rápidas la diferencia es imprecisa, 8 pp (IC 95 % −14 a 22 pp). La latencia por pila se analiza en DM-2026-03.

4. Discusión

Quien descuelga es quien llama. En la secuencia que describe Schegloff, el timbre convoca y quien contesta habla primero [1]. En una cabina pública el visitante

descuelga para hacer una llamada, no para atenderla, y espera tono o voz del otro lado. Cuando GAIA calla, el visitante oye una línea muerta [INT]. El saludo corto es el turno del que contesta, el *¿Diga?* de siempre, y deja al visitante el siguiente turno. El saludo largo hace lo mismo pero gasta los primeros segundos en la máquina.

El silencio al teléfono no es copresencia. Hoey distingue los silencios permitidos entre personas presentes de los lapsos que se notan como ausencia [7]. Al teléfono no hay un cuerpo que sostenga el silencio; por encima de un segundo, el silencio ya significa [8]. La asociación entre primera respuesta lenta y abandono apunta en la misma dirección, con la salvedad de que una pregunta difícil puede alargar la respuesta y acortar la llamada a la vez [INT].

5. Amenazas a la validez

- *Confusión temporal.* Los grupos ocupan franjas distintas y coinciden con cambios de prompt descritos en DM-2026-05. El análisis de mismas horas reduce, no elimina, esa confusión.
- *Sin saludo* tiene 24 llamadas; su efecto es grande y su intervalo, ancho.
- *Clasificación.* La clase de saludo sale de comparar texto del agente con nuestros saludos; las aperturas de otra clase quedan fuera.
- *Continuidad* se mide solo entre quienes ya hablaron una vez, una población seleccionada.

6. Recomendaciones para la comunidad

1. No empezar nunca con silencio en un canal de voz sin imagen.
2. Abrir como quien contesta: una palabra corta, variable, que devuelva el turno.
3. Medir la apertura con tres cifras: llamadas cortas, llamadas mudas y supervivencia con censura en el límite.
4. Medir aparte la latencia de la primera respuesta; es la que decide si hay segunda.
5. Para comparar variantes, planificar al menos varios cientos de llamadas por variante; con decenas solo se ven efectos enormes.

7. Datos y código

Dataset gaia-2147 [2]; cifras generadas por `analysis/analyze.py`. Siempre trabajamos con ElevenLabs y Vidiv; estos registros son los del agente de ElevenLabs. No se guardó texto de los visitantes.

8. Conclusiones

El primer segundo de una llamada es una decisión de diseño. Callar convirtió siete de cada diez llamadas en un descolgar y colgar. Una palabra corta, dicha como quien contesta, llevó las llamadas mudas al 6 %. El personaje empieza antes de decir quién es.

Firman. Cris Olmedo y Kurt Sommer, Grupo de investigación Divina Machina. *Cómo citar*: Olmedo, C. y Sommer, K. (2026). El primer segundo. Documento de trabajo DM-2026-04, v 1.0. Grupo de investigación Divina Machina. Divina Machina by Qs Ventures.

Referencias

Fuentes consultadas hasta el 29 de octubre de 2026.

[1] E. A. Schegloff. Sequencing in conversational openings. *American Anthropologist*, 70(6):1075–1095, 1968. doi: 10.1525/aa.1968.70.6.02a00030.

[2] 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.

[3] 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.

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

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

[6] 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.

[7] E. M. Hoey. *When Conversation Lapses. The Public Accountability of Silent Copresence*. Oxford University Press, 2020. URL https://academic.oup.com/book/40563.

[8] 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.

La personalidad como vector de hábitos

Deriva de las reglas de tono y memoria dentro de la llamada en un agente de voz

Cris Olmedo y Kurt Sommer
Grupo de investigación Divina Machina
Documento de trabajo · Estudio observacional
DM-2026-05, versión 1.0 · 30 de octubre de 2026

Resumen. La personalidad de un agente de voz se escribe como una descripción y se encuentra como comportamiento. Proponemos medirla como un vector de tasas de hábitos contables por turno y la seguimos en siete periodos del mismo personaje, GAIA 2147, a lo largo de 82 versiones del prompt.

Una regla de tono añadida en directo redujo las etiquetas de voz sufridas a 0,08 (IC 95 % 0,04–0,12) veces su tasa anterior en pocas horas. Esa misma noche, con la regla aún en el prompt, la tasa volvió a ser 5,27 (IC 95 % 3,28–10,23) veces la de la ventana con la regla recién añadida. Pedir preguntas llevó al agente al 70 % de turnos en pregunta; pedir «más o menos la mitad» lo bajó al 59 %. Retomar palabras del visitante ocurrió en el 36 % (IC 95 % 34–37 %) de los turnos elegibles y no se asoció con cuánto hablaba la gente (Spearman 0,05 (IC 95 % −0,04 a 0,13)).

Las reglas de tono funcionan y se desgastan. Proponemos medir la personalidad tras cada cambio, en horas, y una memoria para personajes públicos que no recuerda a personas. Los periodos difieren también en saludo, hora y público; los efectos sobre los visitantes no se pueden separar aquí.

Palabras clave: personalidad de agentes; prompts; deriva; etiquetas de voz; memoria conversacional; medidas de comportamiento.

1. Introducción

Un personaje se describe con adjetivos. GAIA es amarga, curiosa, culpable, cansada y divertida. Lo que el visitante encuentra es otra cosa: cuánto habla, si pregunta o cuenta, si suspira, si suena a asistente. Esa conducta se puede contar. La industria ha visto que la personalidad de un modelo se mueve sin que nadie lo pida: en 2025 OpenAI retiró una actualización de GPT-4o que se había vuelto complaciente [1], y Anthropic describe los rasgos como direcciones internas del modelo que pueden derivar y vigilarse [2]. A escala de un prompt, la pregunta es más modesta y práctica: ¿qué hace una regla de tono, cuánto dura y cómo se mide?

2. Medida

Definimos la personalidad observable de una versión v como un vector de tasas por turno del agente [DEF]:

$$\mathbf{h}_v = (r_q,\ \bar{w},\ r_e,\ r_s,\ r_a,\ r_c), \qquad r_x = \frac{\sum_j N_{x,j}}{\sum_j N_{\text{turnos},j}}, \tag{1}$$

donde j recorre las conversaciones. r_q es la proporción de turnos que terminan en pregunta; $\bar{w}$, palabras por turno; r_e y r_s, etiquetas de voz por turno, todas y las sufridas (*sigh*, *softly*, *slow*, *sad*, *whisper* y similares); r_a, frases de asistente de una lista fija; r_c, el hilo retomado definido en la sección 5. Las etiquetas son acotaciones que el modelo de voz interpreta [3]. El recuento lo hace el código sobre las transcripciones en memoria; no se guardó texto [4].

Cada tasa es una razón de sumas. Sus intervalos del 95 % se calculan remuestreando conversaciones, no turnos, y los cambios entre periodos se expresan como cocientes de tasas con intervalos por bootstrap de conversaciones en cada grupo [5, 6]. Una regla de tono es un movimiento buscado de $\mathbf{h}$; la deriva es el movimiento que nadie pidió.

Tabla 1. Ventanas de Barcelona Design Week y lo que cambió antes de cada una.

Ventana	Cambio previo
A, hasta el 24 a las 13.53	Prompt de llegada; saludo largo fijo
B, 13.53 a 16.37	Regla de tono: seca, firme, directa; sin suspiros ni susurros; cinco etiquetas neutras permitidas
C, 16.37 a 17.18	Apertura escrita a mano; pruebas de saludo; saludo corto variable
D, 17.18 a 18.41	Sonar humana; más o menos la mitad de los turnos en pregunta
E, desde las 18.41	Cambiar de idioma en el acto si el visitante lo pide

3. Periodos

Comparamos Sónar+D, IFA y cinco ventanas de Barcelona Design Week definidas por cambios del prompt, unidos a los identificadores de versión con nuestro registro de cambios, con precisión de minutos [DEF]:

4. Resultados

4.1. Tres eventos, tres personalidades

La tabla 2 da el vector por periodo [MED]. En Sónar+D, GAIA sonaba a servicio, con 17,2 (IC 95 % 15,6–18,9) frases de asistente cada 100 turnos. En IFA preguntó menos que nunca, en el 19 % (IC 95 % 15–22 %) de sus turnos. En Barcelona el prompt la pedía curiosa

Tabla 2. Vector de hábitos por periodo. Pregunta y hilo como porcentaje; etiquetas por turno; frases de asistente cada 100 turnos.

Periodo	Conv.	Turnos	Preg.	Pal./t.	Etiq./t.	Sufr./t.	Asist./100	Hilo
Sónar+D	681	3452	30	22,2	1,04	0,53	17,2	34
IFA	247	668	19	17,3	1,39	0,49	1,8	41
BDW · A	263	1754	53	20,2	0,83	0,27	7,1	38
BDW · B	112	557	70	18,2	0,44	0,02	2,7	38
BDW · C	31	158	75	18,0	0,61	0,05	3,8	32
BDW · D	46	338	70	19,2	0,52	0,02	3,8	36
BDW · E	75	625	59	20,2	0,68	0,11	3,4	32

Tasas como razón de sumas sobre turnos de GAIA. Los IC 95 % por bootstrap de conversaciones están en el texto y en la figura.

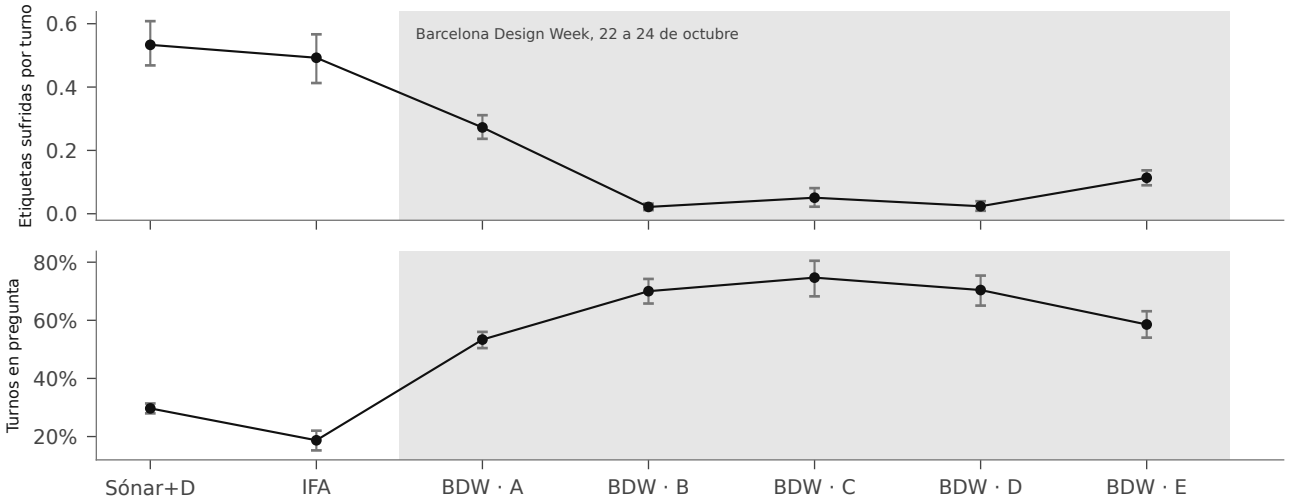


Figura 1. Dos hábitos a lo largo de los siete periodos, con IC 95 % por bootstrap de conversaciones. La banda gris marca Barcelona Design Week.

y, ya en la ventana A, el 53 % (IC 95 % 50–56 %) de sus turnos terminaban en pregunta.

4.2. La regla funciona

Tras la regla de tono, las etiquetas sufridas pasaron de 0,27 (IC 95 % 0,24–0,31) por turno a 0,02 (IC 95 % 0,01–0,03), un cociente de 0,08 (IC 95 % 0,04–0,12). Las etiquetas totales bajaron mucho menos (0,53 (IC 95 % 0,46–0,61)): las neutras ocuparon su lugar. GAIA conservó una voz expresiva y perdió los suspiros (figura 1).

4.3. La regla se desgasta

En la ventana E las etiquetas sufridas volvieron a 0,11 (IC 95 % 0,09–0,14) por turno, 5,27 (IC 95 % 3,28–10,23) veces la tasa de la ventana B. La regla seguía en el prompt; ningún cambio posterior pedía la voz antigua [MED]. Dos explicaciones son compatibles con los datos y no podemos separarlas: cada cambio alargó el prompt y la regla compite con lo que la rodea; y el modelo tiene hábitos propios que vuelven cuando la instrucción pierde peso relativo [INT]. El intervalo es ancho porque la ventana B tiene pocas etiquetas sufridas.

4.4. Preguntar no es escuchar

La petición de preguntar funcionó demasiado: entre la ventana A y la B, la tasa de preguntas se multiplicó por 1,31 (IC 95 % 1,21–1,42) y llegó al 70 %. Pedir «más o menos la mitad» no movió la ventana D (70 %); la E bajó al 59 %, un cociente de 0,83 (IC 95 % 0,75–0,93).

Tabla 3. Resultados de las llamadas por ventana. Confundidos con saludo, hora y público.

Ventana	n	Mediana (s) [IC 95 %]	Corta	Límite
A	271	82 [66; 95]	17 %	34 %
B	117	36 [21; 56]	28 %	19 %
C	34	55 [14; 80]	21 %	18 %
D	46	111 [79; 125]	15 %	46 %
E	75	124 [103; 125]	3 %	53 %

Las instrucciones de proporción tardan en notarse o necesitan el siguiente cambio para asentarse [INT]. Las frases de asistente pasaron de Sónar+D a la ventana B a 0,16 (IC 95 % 0,08–0,25) veces su tasa, sin llegar a cero; parte de lo que queda es legítimo dentro de una historia.

4.5. ¿Lo notaron los visitantes?

No podemos decirlo (tabla 3). La ventana B tuvo llamadas más cortas que la A y la E, las más largas, con el 53 % llegando al límite de tiempo. Pero el saludo cambió en esas mismas horas, como estudia DM-2026-04, y el público cambia a lo largo del día. El comportamiento se movió con claridad; su efecto sobre los visitantes está confundido.

Tabla 4. Conversaciones largas por tercil de hilo retomado.

Tercil de hilo	n	Hilo	Pal. visitante/t.	Límite
Bajo	187	17 %	8,5	47 %
Medio	187	33 %	9,0	49 %
Alto	187	50 %	9,8	44 %

5. Memoria dentro de la llamada

GAIA no recuerda a nadie entre llamadas. Fue una decisión: una cabina pública recibe niños y desconocidos que nunca aceptaron ser recordados. Tiene tres memorias de alcance distinto: lo dicho en la llamada, el lugar (un ánimo de la ciudad que solo llega si llama a una herramienta, cosa que hizo en 24 de 1527 conversaciones) y ella misma, escrita en el prompt.

Para la primera definimos el hilo retomado. Un turno elegible del agente retoma el hilo si contiene una palabra de contenido de cinco letras o más que el visitante dijo antes y que el agente no había usado [DEF]:

$$R = \frac{1}{|A|} \sum_{a \in A} \mathbb{1}[W(a) \cap (V_{<a} \;\; G_{<a}) \neq \emptyset]. \qquad (2)$$

En el conjunto, R fue 36 % (IC 95 % 34–37 %). En las 561 conversaciones con al menos cinco turnos elegibles, la mediana por llamada fue 33 % (IC 95 % 33–38 %). La tasa apenas se movió con el prompt, entre el 32 % y el 41 % según el periodo.

Retomar el hilo no se asoció con que el visitante hablara más (Spearman 0,05 (IC 95 % −0,04 a 0,13), $p = 0{,}26$). Por terciles (tabla 4), los visitantes del tercil alto dijeron 9,8 palabras por turno frente a 8,5 en el bajo, y la proporción de llamadas al límite fue parecida. La medida es tosca: repetir una palabra no es entenderla [INT]. Lo que muestra es estrecho: reutilizar mecánicamente lo que dijo la gente no es, por sí solo, lo que la mantiene al teléfono.

5.1. Lo que dice la investigación

Un estudio longitudinal con 24 personas y 10 sesiones encontró que la memoria percibida dependía de la relación construida y actuaba sobre el disfrute posterior a través de la autorrevelación [7] [EMP]. Un estudio de mago de Oz con 30 participantes encontró ventaja para la comprensión del contexto y no para la personalización con memoria a largo plazo [8]. La habilidad conversacional se asocia a mayor humanidad percibida [9]. Los grandes asistentes convirtieron la memoria en producto con controles explícitos [10, 11]; son asistentes personales, no personajes públicos.

6. Propuestas

Memoria para un personaje público [HIP]. Tres capas, ninguna sobre personas: el hilo como hábito explícito del prompt; una memoria anónima de la sala con los temas del día, nunca ligada a una voz ni a un nombre; y un canon fijo de recuerdos del personaje para que no se contradiga entre visitantes.

Experimento propuesto [HIP]. Alternar días u horas con y sin memoria de la sala, con el mismo saludo, pila y límite. Medida principal: proporción de llamadas en que el visitante habla cinco turnos o más. Para detectar un paso del 50 al 60 % con $\alpha = 0{,}05$ y potencia de 0,8 hacen falta

$$n = \frac{\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2 \left(p_1\,(1-p_1) + p_2\,(1-p_2)\right)}{\left(-p_1 + p_2\right)^2} \qquad (3)$$

385 llamadas por grupo, dos o tres días al ritmo de Barcelona Design Week. La expresión se genera con cálculo simbólico en el análisis y se comprueba contra el cálculo numérico. La figura 2 muestra cuánto crece n si el efecto buscado es menor.

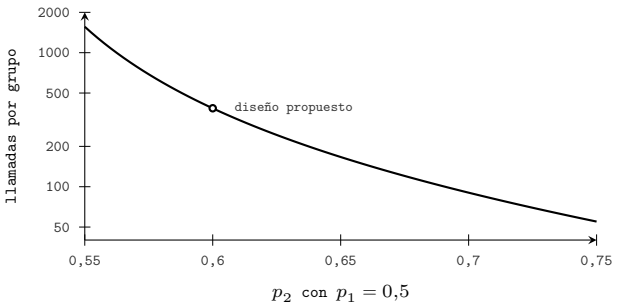


Figura 2. Tamaño por grupo según la ecuación (3), con $\alpha = 0{,}05$ y potencia 0,8. Curva analítica dibujada con PGFPlots.

7. Amenazas a la validez

- Recuentos heurísticos: un signo de interrogación no siempre es una pregunta y una lista fija no recoge paráfrasis.
- Ventanas definidas por hora y versión, unidas a mano con el registro de cambios.
- Varios cambios simultáneos el 24 de octubre: saludo, prompt y público.
- Sin etiquetas de una persona: nadie leyó conversaciones para validar las medidas.

8. Recomendaciones para la comunidad

1. Medir la personalidad como tasas por turno, con intervalos por conversación, después de cada cambio.
2. Volver a medir horas después: el primer resultado tras un cambio no es el estable.
3. Dar al modelo de voz una lista cerrada de etiquetas; una lista abierta vuelve a lo de siempre.
4. Pedir proporciones, no absolutos: «pregunta siempre» produce una entrevista.
5. Separar memoria de la conversación, del lugar y del personaje, y no recordar a personas en espacios públicos.

9. Datos y código

Dataset gaia-2147 [4]; cifras de `analysis/analyze.py`. Siempre trabajamos con ElevenLabs y Vidiv; los datos son del agente de ElevenLabs.

10. Conclusiones

Una personalidad escrita en un prompt es una hipótesis sobre una conducta. Contada turno a turno, la regla de tono funcionó en horas y se desgastó en la misma tarde. La personalidad no es un ajuste que se escribe una vez: se mide, como la temperatura de una sala.

Firman. Cris Olmedo y Kurt Sommer, Grupo de investigación Divina Machina. *Cómo citar*: Olmedo, C. y Sommer, K. (2026). La personalidad como vector de hábitos. Documento de trabajo DM-2026-05, v 1.0. Grupo de investigación Divina Machina. Divina Machina by Qs Ventures.

Referencias

Fuentes consultadas hasta el 29 de octubre de 2026.

[1] TechCrunch. OpenAI explains why ChatGPT became too sycophantic, April 2025. URL https://techcrunch.com/2025/04/29/openai-explains-why-chatgpt-became-too-sycophantic.

[2] Anthropic. Persona vectors. monitoring and controlling character traits in language models, August 2025. URL https://www.anthropic.com/research/persona-vectors.

[3] ElevenLabs. Eleven v3. ElevenLabs Blog, 2025. URL https://elevenlabs.io/blog/eleven-v3. Etiquetas de audio y voz expresiva.

[4] 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.

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

[6] 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.

[7] R. Sumida et al. Memory-driven self-disclosure and relational turning points. arXiv:2607.14593. Aceptado en ICMI 2026, 2026. URL https://arxiv.org/abs/2607.14593.

[8] POSTECH. Comprensión del contexto y memoria a largo plazo en agentes conversacionales. Tesis, repositorio POASIS, s. f. URL https://poasis.postech.ac.kr/handle/2014.oak/123425.

[9] R. M. Schuetzler, G. M. Grimes, and J. S. Giboney. The impact of chatbot conversational skill on engagement and perceived humanness. *Journal of Management Information Systems*, 2020. URL https://scholarsarchive.byu.edu/facpub/5655.

[10] OpenAI. Memory and new controls for ChatGPT, 2025. URL https://openai.com/index/memory-and-new-controls-for-chatgpt/. Actualización de abril de 2025.

[11] Bloomberg. Anthropic tries to win users from ChatGPT with memory feature, March 2026. URL https://www.bloomberg.com/news/articles/2026-03-03/anthropic-tries-to-win-users-from-chatgpt-with-memory-feature.

La sala hace la mitad del trabajo

Cuerpo, público y conducta de un agente de voz en tres eventos

Cris Olmedo y Kurt Sommer
Grupo de investigación Divina Machina
Documento de trabajo · Estudio observacional
DM-2026-06, versión 1.0 · 30 de octubre de 2026

Resumen. Un personaje de voz es tres cosas a la vez: un modelo, un cuerpo y la gente que tiene delante. Comparamos el mismo personaje, GAIA 2147, en una cabina telefónica con luz, pantalla e impresora, en tres eventos: Sónar+D (701 llamadas), IFA (274) y Barcelona Design Week (543).

La sala marcó las llamadas más que el guion. En IFA, el 49 % (IC 95 % 43–55 %) de quienes descolgaron no dijo nada, frente al 22 % en Barcelona. En Sónar+D, el 36 % (IC 95 % 33–40 %) de los turnos del visitante eran la propia voz del agente recogida por el micrófono. El modelo apenas usó su cuerpo: en Barcelona, solo el 5,2 % (IC 95 % 3,6–7,4 %) de las llamadas incluyó una herramienta de luz, pantalla o papel. La gente fue amable con la máquina, la puso a prueba y le habló sobre todo de máquinas y de futuro.

Proponemos repartir responsabilidades: el modelo decide el sentido; la sala, en el dispositivo, produce la presencia. Cada evento es una única muestra de sala, público y configuración, y lo tratamos así.

Palabras clave: agentes encarnados; espacios públicos; herramientas; conducta del usuario; eco acústico; instalaciones interactivas.

1. Introducción

Casi toda la atención de diseño de un agente de voz va al modelo. Nuestros registros dicen que el cuerpo y el público decidieron tanto como él. GAIA 2147 contesta en una cabina de Telefónica con un auricular, una pantalla de 20 por 4 caracteres, una tira de 60 luces y una impresora térmica. El modelo puede usar las cuatro como herramientas; el prompt le pide hacerlo poco [1]. Siempre trabajamos con ElevenLabs y Vidiv; estos datos son del agente de ElevenLabs.

Este trabajo pone uno junto a otro tres hallazgos: la sala cambia las llamadas, el modelo casi no usa su cuerpo y la gente hace cosas que una transcripción apenas recoge.

2. Datos y método

Los tres eventos difieren en público, idioma, ruido, pila de voz, prompt, saludo y límite de tiempo. Por eso no ordenamos causas: describimos cada sala con sus intervalos y medimos la heterogeneidad entre eventos con chi cuadrado y la V de Cramér, con corrección de Holm cuando hay varias conductas [2]. Proporciones con Wilson [3]; medianas con bootstrap; tasas por turno con bootstrap de conversaciones [4]. Las conductas, los idiomas y los temas se detectan con léxico en siete idiomas: cada proporción es un mínimo [DEF]. El eco marca un turno del visitante de cuatro o más palabras con al menos el 60 % en el turno anterior de GAIA:

$$\mathrm{eco}(v_i) = \mathbb{1}\Big[|W(v_i)| \geq 4 \ \wedge \ \tfrac{|W(v_i) \cap W(g_{i-1})|}{|W(v_i)|} \geq 0{,}6\Big]. \quad (1)$$

3. Tres salas

La tabla 1 resume las salas [MED]. En IFA casi la mitad de las llamadas duró menos de diez segundos y otra mitad fue muda. Las diferencias entre eventos son claras en llamadas cortas (V = 0,22) y mudas (V = 0,20), ambas con $p < 0{,}001$. IFA juntó la pila más lenta, un público de feria que va de stand en stand y que hablaba sobre todo inglés y alemán; no sabemos cuál pesó más [INT].

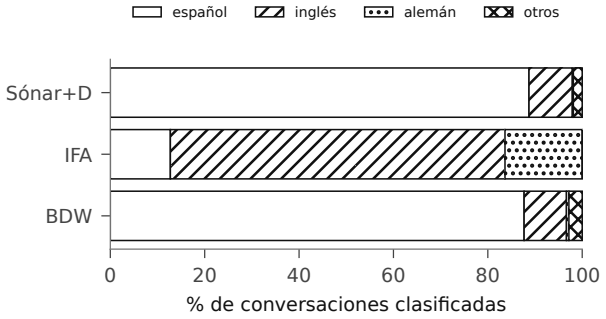


Figura 1. Idioma del visitante por evento, entre conversaciones clasificadas.

Quién estaba al otro lado. Cada sala tuvo su idioma (figura 1): el 89 % de las conversaciones clasificadas fue en español en Sónar+D; en IFA, el 71 % en inglés y el 16 % en alemán; en Barcelona, el 88 % en español. GAIA cambió de idioma a mitad de llamada en el 11,7 %, 12,8 % y 6,4 % de las conversaciones.

La sala contesta. En Sónar+D el 36 % (IC 95 % 33–40 %) de los turnos del visitante eran eco, frente al 4 % en IFA y el 10 % en Barcelona. El eco infla lo que parece decir el visitante (16,4 palabras por turno en Sónar+D frente a 7,5 en Barcelona) y hace que el agente se conteste a sí mismo. En Barcelona lo combatimos con una puerta de ruido y bajando el micrófono mientras habla GAIA [MED]. No documentamos el altavoz ni el volumen de Sónar+D, así que no podemos atribuir la causa allí.

Cuando el público pone a prueba a la máquina. En Sónar+D, el 9,7 % (IC 95 % 7,7–12,1 %) de las llamadas lo cortó un guardarraíl de seguridad de la plataforma: 53 por contenido sexual, 7 por autolesión, 4 por lenguaje ofensivo y 3 por violencia [MED]. En IFA no hubo

Tabla 1. Tres salas. Resultados por evento.

Evento	n	Mediana (s)	Corta	Muda	Límite	Seguridad	Anómala	Eco
Sónar+D	701	58 [47; 66]	27 % [24; 30]	28 % [25; 31]	11 %	9,7 %	7,4 %	36 %
IFA	274	14 [9; 18]	46 % [40; 52]	49 % [43; 55]	3 %	0,0 %	1,1 %	4 %
BDW	543	80 [71; 88]	17 % [14; 21]	22 % [19; 26]	33 %	0,9 %	0,7 %	10 %

Corchetes, IC 95 % (bootstrap para la mediana, Wilson para proporciones). Eco, turnos del visitante que repiten el turno anterior de GAIA.

Tabla 2. Llamadas a herramientas decididas por el modelo, por evento.

Herramienta	Sónar+D	IFA	BDW
Luz	0	2	7
Pantalla	0	0	2
Impresión	0	1	17
Ticket	0	0	4
Cambio de idioma	89	40	39
Colgar	29	3	16

Número de llamadas a cada herramienta decididas por el modelo. Lo que la cabina hace por su cuenta no queda registrado aquí.

ninguno. Los registros no dicen si lo activó el visitante o el agente.

4. El cuerpo que el modelo no usa

En Barcelona el modelo usó al menos una herramienta de cuerpo en el 5,2 % (IC 95 % 3,6–7,4 %) de las llamadas; en las de un minuto o más, el 8,7 % (IC 95 % 6,1–12,3 %). En IFA, el 1,1 %; en Sónar+D, ninguna (tabla 2). Las herramientas más usadas fueron las de sistema: cambiar de idioma y colgar. Entre las de cuerpo ganó el papel. En el conjunto, 31 resultados de herramienta volvieron con error [MED].

Poco fue la instrucción y poco fue lo que obtuvimos: pedimos a un modelo de lenguaje que hiciera de iluminador como tarea secundaria, en mitad de una conversación, con una regla que le pedía contenerse [INT]. Lo que hizo parecer viva a la cabina lo decidía la cabina: una luz que sigue el ánimo de cada frase del visitante con reglas locales, una pantalla que murmura, el tono de la línea, el papel que sale al colgar. Nada de eso queda en los registros de conversación, y es lo primero que añadiríamos a la instrumentación.

La literatura sobre agentes con cuerpo apunta en la misma dirección: en una revisión de 33 experimentos, los robots que compartían espacio con la persona se valoraron mejor que los mismos agentes en pantalla, y pesó más la presencia física que el cuerpo en sí [5].

5. Lo que hizo la gente

Entre las 1067 conversaciones en que el visitante habló, la tabla 3 recoge seis conductas. El 28 % (IC 95 % 26–31 %) empezó con un saludo propio, la primera señal de que acepta la ficción de una llamada. El 5 % (IC 95 % 4–6 %) preguntó si GAIA era real, un robot o quién era; esas llamadas duraron 125 s de mediana frente a 89 s, lo que leemos como señal de interés y no como causa, porque una llamada más larga deja más sitio para la pregunta [INT].

La gente es amable con máquinas que no lo necesitan [6]. El 8 % (IC 95 % 7–10 %) dio las gracias, más en Sónar+D (11 %) que en IFA (4 %); es la única conducta que difiere entre eventos tras la corrección ($p = 0{,}01$). La risa casi no llega a la transcripción (1 %): es la conducta más visible en la cabina y la menos visible en los datos.

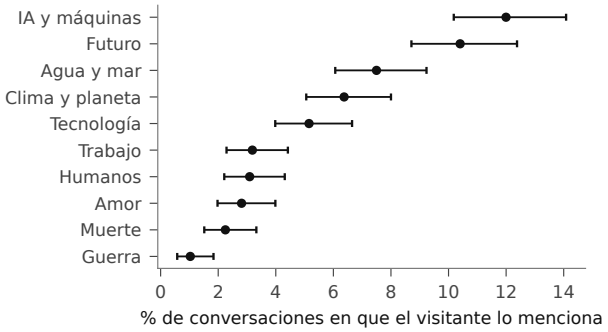


Figura 2. Temas mencionados por el visitante, con IC 95 % de Wilson. Detección por léxico; son mínimos.

Respuestas cortas a una máquina curiosa. El 24 % de los turnos del visitante tenía una o dos palabras, más en Barcelona (32 %) que en Sónar+D (15 %). Barcelona fue también donde más preguntó el agente (DM-2026-05). Cuanto más entrevista la máquina, más corto responde la gente [INT].

De qué quería hablar la gente. En el 35 % de las conversaciones el visitante tocó alguno de los diez temas que seguimos (figura 2). Primero las máquinas (12 % (IC 95 % 10–14 %)) y el futuro (10 %); el amor (3 %) y la muerte (2 %) fueron raros. El clasificador de sentimiento del proveedor, no validado por nosotros, etiquetó el 66 % de 562 llamadas como neutras.

6. Propuesta: el sentido para el modelo, la presencia para la sala

El modelo decide el sentido: prometer un papel, una palabra que se lee, un cambio de luz cuando la conversación gira. La sala produce la presencia: barata, inmediata y siempre activa, en el dispositivo, sin esperar al modelo [HIP]. Una señal de turno en la mano importa sobre todo a los niños, que aprenden cuándo ha terminado de hablar una máquina [7].

7. Amenazas a la validez

- Cada evento es una muestra única de sala, público y configuración; las diferencias son asociaciones.
- Detección léxica: conductas, idiomas y temas son mínimos; el eco y el reconocimiento emborronan el lado del visitante.
- Solo se registra lo que decide el modelo; lo que hace la cabina por su cuenta no tiene recuentos.

Tabla 3. Conductas del visitante por evento, con heterogeneidad entre eventos (V de Cramér) y p ajustada por Holm.

Conducta	Total [IC 95 %]	Sónar+D	IFA	BDW	V	p Holm
Saluda	28 % [26; 31]	25 %	24 %	33 %	0,09	0,09
Pregunta si es real	5 % [4; 6]	6 %	1 %	5 %	0,07	0,26
Da las gracias	8 % [7; 10]	11 %	4 %	6 %	0,11	0,01
Se despide	6 % [5; 7]	7 %	6 %	5 %	0,05	0,85
Ríe (en texto)	1 % [0; 1]	1 %	1 %	0 %	0,04	0,85
Dice palabrotas	2 % [2; 3]	3 %	1 %	2 %	0,05	0,85

Entre conversaciones en que el visitante habló. Detección por léxico en siete idiomas; cada proporción es un mínimo.

Tabla 4. Kit de atmósfera. Las filas propuestas son hipótesis.

Elemento	Quién lo activa	Estado
Luz	La sala por ánimo y turnos; el modelo en los giros	Construido
Pantalla	El modelo, palabras para leer; la sala, murmullos	Construido
Papel	El modelo lo promete; la cabina imprime al colgar	Construido
Sonido de línea	La sala, en cada llamada	Construido
Ambiente sonoro	La sala, según la tensión	Propuesto [HIP]
Vibración de turno	La sala, en el auricular	Propuesto [HIP]

- Límites de tiempo distintos por evento truncan la duración de forma distinta.

8. Recomendaciones para la comunidad

1. Reportar el eco, las llamadas mudas y los cortes de seguridad por sala: hablan de la sala antes que del personaje.
2. Registrar cada acción del entorno con su hora junto al identificador de la conversación.
3. No delegar la presencia en el modelo: lo ambiental, en el dispositivo y al instante.
4. Añadir una hoja de observación por hora; lo que no llega a la transcripción, como la risa o los grupos, solo lo ve una persona.

9. Datos y código

Dataset gaia-2147 [1]; cifras de `analysis/analyze.py`. No se guardó texto de los visitantes.

10. Conclusiones

La sala decidió quién descolgaba, en qué idioma, con cuánto eco y con qué ganas de jugar. El modelo, libre de usar su cuerpo, casi no lo hizo. La presencia de un personaje de voz se construye fuera del modelo, en la sala y en el dispositivo.

Firman. Cris Olmedo y Kurt Sommer, Grupo de investigación Divina Machina. *Cómo citar*: Olmedo, C. y Sommer, K. (2026). La sala hace la mitad del trabajo. Documento de trabajo DM-2026-06, v 1.0. Grupo de investigación Divina Machina. Divina Machina by Qs Ventures.

Referencias

Fuentes consultadas hasta el 29 de octubre de 2026.

[1] 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.

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

[3] 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.

[4] 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.

[5] J. Li. The benefit of being physically present. a survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human-Computer Studies*, 77:23–37, 2015. URL https://napier-repository.worktribe.com/output/3636900.

[6] C. Nass, J. Steuer, and E. R. Tauber. Computers are social actors. In *CHI '94*, pages 72–78, 1994. doi: 10.1145/191666.191703.

[7] J. Festerling and I. Siraj. Anthropomorphizing technology. a conceptual review of anthropomorphism research and how it relates to children's engagements with digital voice assistants. *Integrative Psychological and Behavioral Science*, 56:709–738, 2022. URL https://pubmed.ncbi.nlm.nih.gov/34811705/.

Una Raspberry Pi 5 en una cabina telefónica

Arquitectura de borde, notas de campo y fiabilidad de un agente de voz físico

Cris Olmedo y Kurt Sommer
Grupo de investigación Divina Machina
Documento de trabajo · Nota técnica
DM-2026-07, versión 1.0 · 30 de octubre de 2026

Resumen. Un agente de voz en un objeto físico necesita decidir qué corre en el dispositivo y qué en la nube. Documentamos la cabina de GAIA 2147 en Barcelona Design Week: una Raspberry Pi 5 que lleva todo lo que se toca, la horquilla, el audio, la luz, la pantalla y el papel, y dos plataformas de voz en la nube, ElevenLabs y Vidiv, con paso automático de una a otra.

Damos la especificación completa, once modos de fallo encontrados en campo con su causa y su corrección, y la fiabilidad registrada en la nube. En Barcelona, el 0,7 % (IC 95 % 0,3–1,9 %) de 543 llamadas terminó con una desconexión anómala y el 0,4 % con un fallo técnico. El primer audio del saludo salió del servidor en 95 ms de mediana con la conexión preparada antes de descolgar. En una ventana de registros de la cabina, el 21 % (IC 95 % 15–29 %) de los colgados agotó una espera de dos segundos en el hilo de audio, un fallo todavía abierto.

La regla que proponemos es simple: el borde se encarga de lo que se toca y de lo que debe pasar en milisegundos; la nube, del lenguaje, la voz y la seguridad, que cambian cada pocos meses.

Palabras clave: computación de borde; Raspberry Pi; agentes de voz; fiabilidad; instalaciones interactivas; análisis de modos de fallo.

1. Introducción

Una cabina telefónica en una feria tiene necesidades sencillas y restricciones duras. Tiene que parecer un teléfono, arrancar sola y funcionar todo el día, sobrevivir al wifi del pabellón y no abrir ningún puerto a internet. La construimos sobre una Raspberry Pi 5: Cortex-A76 de cuatro núcleos a 2,4 GHz, wifi de doble banda, Bluetooth 5.0, dos USB 3 y el conector de 40 pines, con una fuente de 5 V y 5 A [1]. Es suficiente para el mundo físico de la cabina. No es donde debe fabricarse la voz.

Esta nota cubre la versión de Barcelona Design Week. Los registros no documentan el hardware de Sónar+D e IFA; las cifras de fiabilidad de esos eventos son de la nube y no comparan hardware [MED].

2. Arquitectura

La Pi se encarga de todo lo que se puede tocar y del estado de la llamada (figura 1). Pone el tono de marcar, abre la conversación, lleva el audio al auricular y lo trae, controla luz, pantalla e impresora y mantiene reglas y cupos para que nada se dispare. La nube se encarga de la voz, el lenguaje y la seguridad. Siempre trabajamos con dos plataformas, ElevenLabs y Vidiv; si una falla, la cabina pasa a la otra sin intervención [MED]. Un Worker de Cloudflare es el otro extremo de un túnel que abre la cabina, por el que bajan herramientas y ajustes y sube el estado.

3. Modos de fallo

La tabla 2 recoge lo que nos pasó, con el formato de un análisis de modos de fallo. Son observaciones de construcción y operación, no recuentos [MED].

Dos merecen una ecuación. La hoja de datos del WS2812B pide un nivel alto de al menos $0{,}7\,V_{DD}$ [2]:

$$V_{IH,\min} = 0{,}7 \times 5\,\mathrm{V} = 3{,}5\,\mathrm{V} > V_{OH,\mathrm{Pi}} \approx 3{,}3\,\mathrm{V}. \quad (1)$$

Muchas tiras funcionan igual; la nuestra no, y no llegamos a poner el adaptador. La figura 2 muestra la corrección propuesta, un búfer de la familia 74AHCT alimentado a 5 V, cuyo umbral de entrada admite los 3,3 V de la Pi y cuya salida llega a V_{DD} [HIP].

Y la espera del hilo de audio: en una ventana de tres horas y media de registros de la cabina, el 21 % (IC 95 % 15–29 %) de 136 colgados agotó una espera de dos segundos [MED]. Ese recuento sale de los registros de la Pi, que no forman parte del dataset gaia-2147.

4. Fiabilidad

La tabla 3 da la fiabilidad registrada en la nube para los tres eventos [MED]. Barcelona tuvo la menor proporción de desconexiones anómalas, 0,7 % (IC 95 % 0,3–1,9 %), frente al 7,4 % (IC 95 % 5,7–9,6 %) en Sónar+D, cuyo hardware no documentamos. El primer audio del saludo bajó de 287 ms (solo síntesis, Sónar+D) a 95 ms en Barcelona, con la conexión firmada preparada antes de descolgar. Son tiempos de servidor: la Pi y el altavoz añaden su propio retardo, del que solo tenemos lecturas sueltas de entre 0,55 y 0,73 s desde que se abre la conversación hasta el primer sonido.

La carga es real (figura 3): en Barcelona la cabina llegó a 75 llamadas en una hora. Una tarde atendió más de cien llamadas sin caerse; las únicas intervenciones fueron reinicios decididos por nosotros. El coste de nube fue de 0,10 USD por minuto en Barcelona y 0,12 USD en IFA; no calculamos el coste del hardware.

5. Lo que debe hacer el borde

- Todo lo físico y todo lo que debe pasar en milisegundos: horquilla, tonos, luz, pantalla, papel.
- El estado de la llamada, para saber cuándo es seguro imprimir, reiniciar o actualizar.
- El paso entre plataformas: ElevenLabs y Vidiv a la vez, con cambio automático.

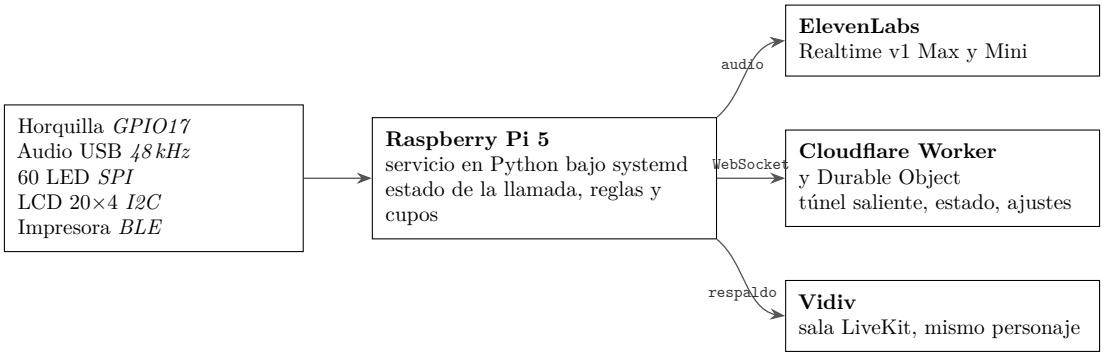


Figura 1. Qué corre dónde. El borde lleva lo físico y el estado; la nube, el lenguaje, la voz y la seguridad. La cabina nunca abre un puerto: todo sale hacia fuera.

Tabla 1. Especificación de la cabina en Barcelona Design Week, extraída de su configuración y código.

Subsistema	Pieza	Interfaz	Ajuste
Ordenador	Raspberry Pi 5		Servicio Python bajo systemd con reinicio automático
Audio	Tarjeta USB	USB, ALSA	48 kHz mono, bloques de 20 ms; mezclador al 98 %; ganancia 3,5; búfer objetivo 40 ms, techo 30 s
Horquilla	Interruptor original	GPIO17	Polaridad invertida por software; antirrebote 120 ms; confirmación 0,4 s
Pantalla	LCD 20×4	I2C 0x27	Murmullos y mensajes cortos con límites por llamada
Luz	60 LED WS2812E	SPI, GPIO10	Brillo 0,35; color de reposo leído de la nube cada 10 s
Papel	Impresora térmica	Bluetooth LE	384 puntos por línea; energía de impresión subida; imprime al colgar
Voz	ElevenLabs Realtime v1	WebSocket	Conexión firmada preparada antes de descolgar
Voz, respaldo	Vidiv	LiveKit	Mismo personaje; paso automático
Enlace	Cloudflare Workers	WebSocket saliente	Estado cada pocos segundos; ajustes cada 10 s
Software	elevenlabs 2.66.0, websockets 16.0		Versiones fijadas; una actualización silenciosa puede deshacer un arreglo

Tabla 2. Modos de fallo encontrados en campo.

Síntoma	Causa	Corrección	Estado
Descolgar no hacía nada	Horquilla con polaridad contraria y rebotes	Lectura invertida, antirrebote y segunda comprobación	Resuelto
Voz entrecortada	Audio a ráfagas y búfer pequeño	Techo del búfer a 30 s, objetivo 40 ms	Resuelto
Volumen reiniciado al arrancar	La tarjeta USB cambiaba de número	Buscarla por nombre	Resuelto
Distorsión con la risa	Ganancia excesiva en el auricular	Mezclador al 98 %, ganancia 3,5	Resuelto
Papel en blanco	Calor de impresión por defecto bajo	Más energía de impresión	Resuelto
Impresora ocupada	Una sola conexión Bluetooth	El servicio la gobierna	Resuelto
Imprimir rompía el audio	Wifi y Bluetooth comparten el módulo de radio	Imprimir solo al colgar	Resuelto
Luz blanca azulada	Datos a 3,3 V frente a un umbral de 3,5 V	Adaptador de nivel y masa común	Abierto
GAIA se contestaba a sí misma	El micrófono recogía el altavoz	Puerta de ruido y micrófono bajo al hablar	Mitigado
Descolgar tardaba unos 2 s más	El hilo de audio a veces no confirmaba pausa o reanudación	Parar sin vaciar y medir cada paso	Abierto
Ajuste de nube tardío	Copia vieja en el objeto del túnel	Consulta de ajustes cada 10 s	Resuelto

Tabla 3. Fiabilidad y saludo por evento, según los registros de la nube. Corchetes, IC 95 % de Wilson o bootstrap.

Evento	Llamadas	Desconexión anómala	Fallo técnico	Saludo (ms)
Sónar+D	701	52 · 7,4 % [5,7; 9,6]	8 · 1,1 % [0,6; 2,2]	287 (TTS)
IFA	274	3 · 1,1 % [0,4; 3,2]	3 · 1,1 % [0,4; 3,2]	179 (e2e)
BDW	543	4 · 0,7 % [0,3; 1,9]	2 · 0,4 % [0,1; 1,3]	95 (e2e)

Desconexión anómala, cierre de WebSocket 1006 o 1011. Fallo técnico, estado fallido sin guardarraíl de seguridad. Saludo, mediana en servidor del primer audio del turno 0; en Sónar+D solo existe el primer byte de síntesis.

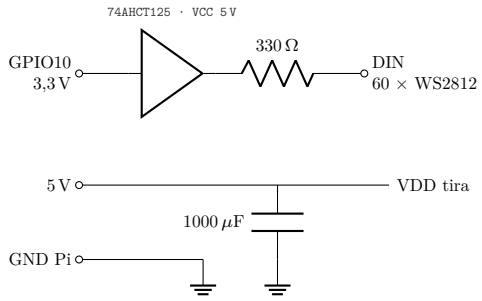


Figura 2. Adaptador de nivel propuesto para la tira de luz, con masa común. Esquema dibujado con circuitikz; no instalado en Barcelona.

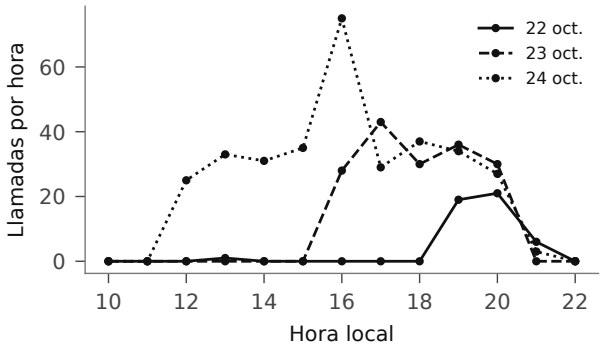


Figura 3. Llamadas por hora en Barcelona Design Week.

- La privacidad por defecto: historial local, subida solo si se activa y con aviso visible.

Y lo que no: reconocimiento, modelo de lenguaje, voz y filtros de seguridad. Cambian cada pocos meses y la cabina no debería necesitar una imagen nueva cuando lo hacen [INT].

6. Amenazas a la validez

Una sola versión en un solo evento. Las notas de campo no están contadas. La fiabilidad de Sónar+D e IFA no corresponde a un hardware documentado. Los tiempos hasta el primer sonido en la cabina son lecturas sueltas. El recuento de esperas agotadas cubre una ventana de registros, no el evento.

7. Recomendaciones para la comunidad

1. Poner un adaptador de nivel en cualquier tira WS2812 alimentada a 5 V desde el primer día.
2. Dar a la impresora su propia radio o pasar a USB; imprimir no debe competir con la voz.
3. Registrar en el dispositivo, cada hora, tiempo en marcha, reinicios, primer sonido y eco.
4. Reiniciar solo con el auricular colgado, comprobándolo dos veces.
5. Fijar versiones de librerías y tener una segunda tarjeta preparada con un único script de instalación.

8. Datos y código

Fiabilidad y tiempos de servidor del dataset gaia-2147 [3]; cifras de `analysis/analyze.py`. Especificación y modos de fallo, de la configuración y los registros de la cabina.

9. Conclusiones

Un ordenador pequeño basta para que un objeto converse si no intenta fabricar la voz. La cabina llevó lo que se toca; la nube, lo que se dice. Lo que falló fue casi siempre físico: polaridades, radios compartidas, voltajes. Esa es la parte que no viene en ninguna ficha de producto.

Firman. Cris Olmedo y Kurt Sommer, Grupo de investigación Divina Machina. *Cómo citar*: Olmedo, C. y Sommer, K. (2026). Una Raspberry Pi 5 en una cabina telefónica. Documento de trabajo DM-2026-07, v 1.0. Grupo de investigación Divina Machina. Divina Machina by Qs Ventures.

Referencias

Fuentes consultadas hasta el 29 de octubre de 2026.

[1] Raspberry Pi Ltd. Raspberry pi 5. Página de producto, 2023. URL https://www.raspberrypi.com/products/raspberry-pi-5/.

[2] Worldsemi. WS2812B. intelligent control LED integrated light source. Hoja de datos, copia de la Universidad de Cambridge. URL https://www.cl.cam.ac.uk/teaching/1516/ECAD+Arch/files/pcb/WS2812B.pdf.

[3] 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.

El arte de la conversación según Linklater

Doce escenas de la trilogía *Before*, la ciencia de la conversación y 1527 llamadas

Cris Olmedo y Kurt Sommer
Grupo de investigación Divina Machina
Documento de trabajo · Ensayo con datos
DM-2026-08, versión 1.0 · 30 de octubre de 2026

Resumen. Tres películas en las que casi no pasa nada salvo una conversación sostuvieron al público durante dieciocho años. Leemos doce escenas de la trilogía *Before* de Richard Linklater junto a la ciencia de los turnos, el silencio y la autorrevelación, y junto a 1527 llamadas con un personaje artificial al teléfono, GAIA 2147.

Las películas enseñan una gramática: empezar por algo compartido, invitar con un juego, dejar silencios habitados, preguntar por lo que el otro acaba de decir, bajar una capa cada vez, anunciar el reloj y acabar con ganas de más. Las llamadas muestran dónde una máquina se aleja de ella. Los visitantes dijeron el 33 % (IC 95 % 32–34 %) de las palabras; GAIA preguntó 1,7 (IC 95 % 1,6–1,8) veces más que ellos, y en Barcelona, donde le pedimos curiosidad, 2,8 veces más. Con una cascada de voz, el 16 % de la llamada mediana era silencio de máquina.

Proponemos diez reglas para conversar bien, para personas y para máquinas. Las películas son ficción escrita, interpretada y montada; no medimos nada en ellas y las comparaciones con GAIA son analogías marcadas como tales.

Palabras clave: conversación; cine; Linklater; toma de turnos; reciprocidad; agentes de voz; autorrevelación.

Tabla 1. Tres películas, un mismo mecanismo [2, 7, 8].

Película	Lugar	Lo que destaca
Sunrise, 1995	Viena, una noche	Cabina de escucha, poeta callejero, llamadas fingidas
Sunset, 2004	París, una tarde	Tiempo real; planos de hasta unos 11 min
Midnight, 2013	Grecia, un día	Coche en un plano de unos 13 min

1. Películas en las que no pasa nada

En *Before Sunrise*, Jesse le cuenta a Céline una idea para la televisión: un canal que emitiera veinticuatro horas seguidas de la vida corriente de alguien [1]. Las tres películas hacen eso mismo con dos personas. Se conocen en un tren y caminan por Viena hasta que amanece; nueve años después se reencuentran una tarde en París; nueve años más tarde pasan un día en Grecia, y discuten [2]. Son producciones estadounidenses rodadas en Europa, escritas por Linklater con Kim Krizan y después con los dos intérpretes, Ethan Hawke y Julie Delpy [3–5].

GAIA es un personaje cuya existencia entera es una conversación, en un teléfono público, con un desconocido que no tiene ningún motivo para quedarse. Si algo del cine puede enseñar a una máquina a sujetar a alguien con la voz, son estas tres películas [INT]. El trabajo compañero DM-2026-01 desarrolla el marco conceptual [6]; este pone las escenas junto a datos.

2. Doce escenas

No citamos diálogos; describimos. Cada escena enseña un mecanismo y se contrasta con GAIA.

2.1. *Before Sunrise*

1. El tren. Una pareja discute en alemán; Céline cambia de asiento y acaba junto a Jesse, que le pregunta si ha entendido la discusión [2]. La conversación empieza por algo que ambos acaban de ver: un tercer objeto quita peso a la primera frase [INT]. Cuando GAIA calló al descolgar, el 71 % (IC 95 % 51–85 %) de las llamadas duró menos de diez segundos; con un saludo corto que reconoce la situación, el 7 % (DM-2026-04).

2. El viaje en el tiempo. Para que baje en Viena, Jesse no pide una cita: propone imaginar que, dentro de veinte años, ella se pregunta qué habría pasado [1]. Una invitación como hipótesis no obliga a nada. GAIA entera es ese juego: una voz que dice llamar desde 2147 y una factura del futuro impresa al colgar [INT].

3. La cabina de escucha. En una tienda de discos escuchan una canción de Kath Bloom sin hablar, mirándose a escondidas [9, 10]. Hoey distingue los silencios permitidos entre personas copresentes de los lapsos que se notan como ausencia [11]. Al teléfono, un silencio no se ve compartido: sin saludo, en el 62 % de las llamadas el visitante no dijo nada.

4. El poeta del canal. Un hombre les escribe un poema con la palabra que eligen [10]. El regalo funciona porque está hecho con algo que ellos trajeron [INT]. GAIA retomó una palabra del visitante en el 36 % (IC 95 % 34–37 %) de sus turnos elegibles (DM-2026-05). La cabina podría imprimir algo hecho con las palabras del visitante [HIP].

5. Las llamadas fingidas. En un café, cada uno finge llamar a un amigo y le cuenta cómo es la persona que tiene delante [12]. El teléfono fingido da distancia y la distancia da valor. GAIA es un teléfono: el auricular deja hablar sin que nadie vea la cara [INT].

6. Los lugares vacíos. Deciden no intercambiar teléfonos y quedar en seis meses; la película termina con los sitios por donde pasaron, vacíos [10, 13]. El final de una buena conversación es una cita abierta [INT].

2.2. *Before Sunset*

7. La librería. Jesse presenta su novela en París y ve a Céline; tiene un avión en menos de una hora [2]. El reloj se anuncia desde el primer minuto y la película dura lo que dura la tarde. GAIA tiene límite de tiempo y no lo dice: en Barcelona lo alcanzó el 33 % de las llamadas.

8. De lo ligero a lo hondo. Empiezan por trabajo y política y bajan, capa a capa, a lo que importa [2]. Altman y Taylor describieron la penetración social por capas [14]; las preguntas recíprocas y crecientes aumentan la cercanía entre desconocidos [15]; y la gente subestima cuánto disfrutará una conversación profunda [16]. Nuestros visitantes casi nunca bajaron: amor en el 3 % y muerte en el 2 % de las conversaciones, frente a máquinas (12 %) y futuro (10 %).

9. Vas a perder ese avión. La película corta antes de que nadie decida nada [2]. Las conversaciones casi nunca acaban cuando las dos personas quieren [17]. Acabar con ganas de más es lo que hace volver [INT].

2.3. *Before Midnight*

10. El sol que se va. Miran ponerse el sol y Céline va diciendo que sigue ahí, hasta que deja de estar [18]. Señalar juntos el presente es la forma más simple de atención compartida. GAIA no ve la sala, pero la oye [HIP].

11. La habitación del hotel. La pelea más larga de la trilogía: ella se va y vuelve dos veces [2]. Una buena conversación no es la que nunca se rompe; es la que se repara. En la pelea los turnos se alargan y nadie recoge lo que dice el otro, lo contrario del poeta [INT]. En el 2 % de nuestras conversaciones hubo palabrotas.

12. La carta del futuro. Jesse se presenta como viajero del tiempo con una carta de ella misma a los ochenta y dos años [2]. El juego que abrió la primera conversación repara la última. Volver a lo dicho al principio es la forma más fuerte de memoria en una conversación [INT].

3. El tiempo es el material

En *Before Sunset* el tiempo en pantalla y el de la historia son el mismo [7]. Una llamada también es tiempo real; lo único que la rompe es el silencio mientras la máquina prepara una respuesta. Lo medimos como la fracción de la llamada que se va en esperar,

$$\delta = \frac{1}{T_{\text{llamada}}} \sum_{i} t_i. \tag{1}$$

Con la cascada de IFA, δ fue 16 % (IC 95 % 15–18 %) en la llamada mediana; con Realtime Mini, 7 % (IC 95 % 6–7 %). La figura 1 pone las medianas de GAIA en la misma escala que el hueco humano [19, 20] y el silencio máximo estándar de Jefferson [21]. Las respuestas rápidas se asocian a conexión [22]. En Barcelona, cuando la primera respuesta llegó en dos segundos o menos, el visitante volvió a hablar en el 96 % de las llamadas, frente al 81 % cuando no.

4. Reciprocidad, no interrogatorio

Las conversaciones de la trilogía están equilibradas: los dos preguntan, responden y cuentan [INT]. Medimos el equilibrio con dos números [DEF]:

$$\rho = \frac{W_{\text{vis}}}{W_{\text{vis}} + W_{\text{GAIA}}}, \qquad \alpha = \frac{q_{\text{GAIA}}}{q_{\text{vis}}}, \tag{2}$$

donde W cuenta palabras y q es la proporción de turnos que terminan en pregunta. Una conversación equilibrada tiene $\rho \approx 0{,}5$ y $\alpha \approx 1$.

Los visitantes dijeron el 33 % (IC 95 % 32–34 %) de las palabras; los turnos de GAIA tuvieron 21 palabras de media frente a 12 los suyos (tabla 2, figura 2). El equilibrio de preguntas cambió por evento: en IFA, α fue 0,7 (IC 95 % 0,6–0,9); en Barcelona, donde el prompt pedía curiosidad, 2,8 (IC 95 % 2,5–3,1). Preguntar más no hizo hablar más: convirtió la llamada en una entrevista [INT]. Huang y colaboradores encontraron que preguntar aumenta el agrado sobre todo a través de las preguntas de seguimiento, las que recogen lo que el otro acaba de decir [23]. Una pregunta genérica es un turno vacío con signo de interrogación; una de seguimiento es el poeta del canal. En Sónar+D, ρ está inflado por el eco del micrófono (DM-2026-06).

5. Escrito para sonar sin escribir

Nada de la trilogía es improvisado: la segunda se esbozó en pocos días y se escribió durante un año entre el director y los intérpretes antes de rodarse en quince días [2]. Nosotros hacemos lo mismo a velocidad de máquina: 80 versiones del agente en tres eventos, con 13 min de mediana entre una y otra (DM-2026-02). Un prompt es el guion de un personaje que improvisa.

6. Diez reglas para conversar bien

La tabla 3 junta escenas, ciencia y datos. Las reglas son hipótesis de diseño para una máquina [HIP]; para personas, varias tienen apoyo experimental.

7. Lo que una máquina no puede copiar

Los intérpretes envejecen nueve años entre película y película y el público con ellos; los personajes comparten una historia y se ven la cara. GAIA conoce a cada visitante por primera vez, solo por la voz, durante un par de minutos, y no recuerda a nadie, a propósito. Lo que sí puede copiar es la gramática: tiempo real, huecos cortos, silencio habitado, preguntas que recogen, hondura que baja poco a poco, un juego que da permiso y un reloj que da peso [INT].

8. Amenazas a la validez

Las películas son ficción y no medimos nada en ellas; las escenas se describen a partir de las fuentes citadas.

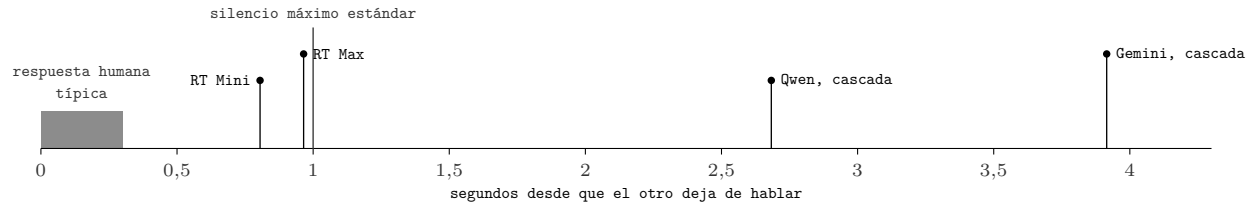


Figura 1. Escala del silencio. Medianas de latencia de GAIA por pila frente a la respuesta humana típica y al silencio máximo de en torno a un segundo. Las referencias humanas no son objetivos; las métricas no son idénticas (DM-2026-03).

Tabla 2. Reciprocidad por evento. Intervalos por bootstrap de conversaciones.

Evento	Conv.	ρ	Preg. GAIA	Preg. visitante	α [IC 95 %]
Todos	1067	33 %	44 %	26 %	1,7 [1,6; 1,8]
Sónar+D	505	40 %	31 %	29 %	1,1 [1,0; 1,2]
IFA	140	24 %	22 %	30 %	0,7 [0,6; 0,9]
BDW	422	24 %	61 %	22 %	2,8 [2,5; 3,1]

ρ, cuota de palabras del visitante. α, cociente entre la proporción de turnos en pregunta de GAIA y la del visitante. Equilibrio: $\rho \approx 0{,}5$ y $\alpha \approx 1$.

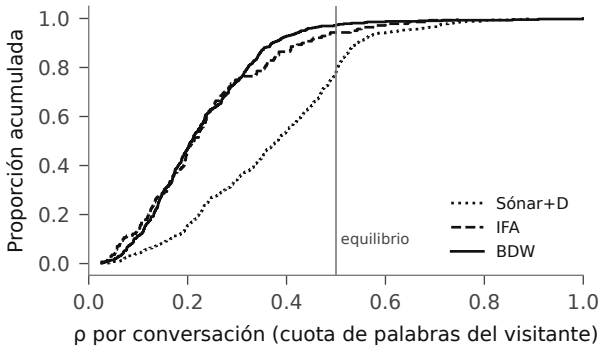


Figura 2. Distribución por conversación de la cuota de palabras del visitante.

Los estudios citados son con personas, casi siempre cara a cara. Nuestras medidas de equilibrio cuentan palabras y signos de interrogación sobre transcripciones procesadas en memoria; el eco infla el lado del visitante en Sónar+D. Ningún texto se guardó [25].

9. Recomendaciones para la comunidad

Medir ρ y α en cualquier agente conversacional, junto con el tiempo muerto δ: tres números que traducen una gramática del cine a una auditoría de diseño. Contar preguntas de seguimiento, no preguntas. Y tratar el silencio al teléfono como ausencia, no como pausa.

10. Conclusiones

Linklater sostuvo a su público con dos personas que hablan. GAIA, con una sola voz al teléfono, se acercó a esa gramática cuando contestó a tiempo y empezó como quien contesta, y se alejó cuando preguntó demasiado. Una buena conversación no se mide por cuánto pregunta una parte, sino por cuánto se recoge de la otra.

Firman. Cris Olmedo y Kurt Sommer, Grupo de investigación Divina Machina. *Cómo citar*: Olmedo, C. y Sommer, K. (2026). El arte de la conversación según Linklater. Documento de trabajo DM-2026-08, v 1.0. Grupo de investigación Divina Machina. Divina Machina by Qs Ventures.

Tabla 3. Diez reglas, con la escena, la evidencia y el estado en GAIA.

	Regla	Escena	Evidencia	GAIA hoy
1	Empieza por algo que los dos veis	El tren	[24]	Saludo corto: 7 % cortas
2	Invita con un juego	El viaje en el tiempo	Sin estudio directo	La ficción de 2147
3	Contesta a tiempo	La cabina de escucha	[19, 21, 22]	0,81 s con RT Mini
4	Deja silencios, pero que se note que sigues ahí	La cabina de escucha	[11]	Al teléfono, silencio es ausencia
5	Pregunta por lo que el otro acaba de decir	El poeta	[23]	$\alpha = 1{,}7$
6	Cuenta algo tuyo por cada cosa que preguntes	De lo ligero a lo hondo	[14, 15]	$\rho = 33\,\%$
7	Baja una capa cada vez	De lo ligero a lo hondo	[16]	Amor 3 %, muerte 2 %
8	Di cuánto tiempo hay	La librería	Sin estudio directo	Límite no anunciado
9	Acaba antes de que se agote	Vas a perder ese avión	[17]	33 % al límite
10	Vuelve a lo que se dijo al principio	La carta del futuro	[23]	Hilo 36 %

Referencias

Fuentes consultadas hasta el 29 de octubre de 2026. Las escenas se describen a partir de fuentes publicadas; no se citan diálogos.

[1] The Werewolf. The one that got away, 2013. URL https://werewolf.co.nz/2013/06/the-one-that-got-away.

[2] Wikipedia. Before trilogy; before sunrise; before sunset; before midnight. Artículos en inglés. URL https://en.wikipedia.org/wiki/Before_trilogy. Tramas, lugares, duraciones, guion, rodaje, recepción y taquilla.

[3] R. Linklater. Before Sunrise. The Criterion Collection, ficha y créditos, 1995. URL https://www.criterion.com/films/28692-before-sunrise. Película. Dirección de R. Linklater; guion de R. Linklater y K. Krizan.

[4] R. Linklater. Before Sunset. The Criterion Collection, ficha y créditos, 2004. URL https://www.criterion.com/films/28693-before-sunset. Película. Dirección de R. Linklater; guion de R. Linklater, J. Delpy y E. Hawke; historia de R. Linklater y K. Krizan.

[5] R. Linklater. Before Midnight. The Criterion Collection, ficha y créditos, 2013. URL https://www.criterion.com/films/28828-before-midnight. Película. Dirección de R. Linklater; guion de R. Linklater, J. Delpy y E. Hawke.

[6] 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.

[7] Filmmaker Magazine. Paris day, 2004. URL https://filmmakermagazine.com/archives/issues/spring2004/features/paris_day.php.

[8] The Austin Chronicle. Before midnight, 2013. URL https://www.austinchronicle.com/screens/before-midnight-11922299/.

[9] Far Out Magazine. Anatomy of a scene. the listening booth in Before Sunrise. URL https://faroutmagazine.co.uk/listening-booth-scene-before-sunrise/.

[10] T. Dirks. Before sunrise (1995). Filmsite. URL https://filmsite.org/beforesunrise.html.

[11] E. M. Hoey. *When Conversation Lapses. The Public Accountability of Silent Copresence*. Oxford University Press, 2020. URL https://academic.oup.com/book/40563.

[12] We Minored in Film. Two confessional faux telephone calls in Before Sunrise, 2013. URL https://weminoredinfilm.com/2013/06/08/i-love-that-scene-two-confessional-faux-telephone-calls-in-before-sunrise/.

[13] The Vienna Review. A cinephile goes walking Before Sunrise, 2012. URL https://www.theviennareview.at/archives/2012/a-cinophile-goes-walking-before-sunrise.

[14] I. Altman and D. A. Taylor. *Social Penetration. The Development of Interpersonal Relationships*. Holt, Rinehart and Winston, 1973.

[15] A. Aron, E. Melinat, E. N. Aron, R. D. Vallone, and R. J. Bator. The experimental generation of interpersonal closeness. a procedure and some preliminary findings. *Personality and Social Psychology Bulletin*, 23(4):363–377, 1997. doi: 10.1177/0146167297234003.

[16] M. Kardas, A. Kumar, and N. Epley. Overly shallow? miscalibrated expectations create a barrier to deeper conversation. *Journal of Personality and Social Psychology*, 122(3), 2022. URL https://www.apa.org/news/press/releases/2021/09/deep-conversations-strangers. DOI no verificado; nota de prensa de la APA.

[17] A. M. Mastroianni, D. T. Gilbert, G. Cooney, and T. D. Wilson. Do conversations end when people want them to? *PNAS*, 118 (10):e2011809118, 2021. doi: 10.1073/pnas.2011809118.

[18] The Varsity. Film review. Before Midnight, 2013. URL https://thevarsity.ca/2013/07/07/film-review-before-midnight/.

[19] 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/.

[20] 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).

[21] 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.

[22] 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/.

[23] K. Huang, M. Yeomans, A. W. Brooks, J. Minson, and F. Gino. It doesn't hurt to ask. question-asking increases liking. *Journal of Personality and Social Psychology*, 113(3):430–452, 2017. doi: 10.1037/pspi0000097.

[24] E. A. Schegloff. Sequencing in conversational openings. *American Anthropologist*, 70(6):1075–1095, 1968. doi: 10.1525/aa.1968.70.6.02a00030.

[25] 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.

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 %

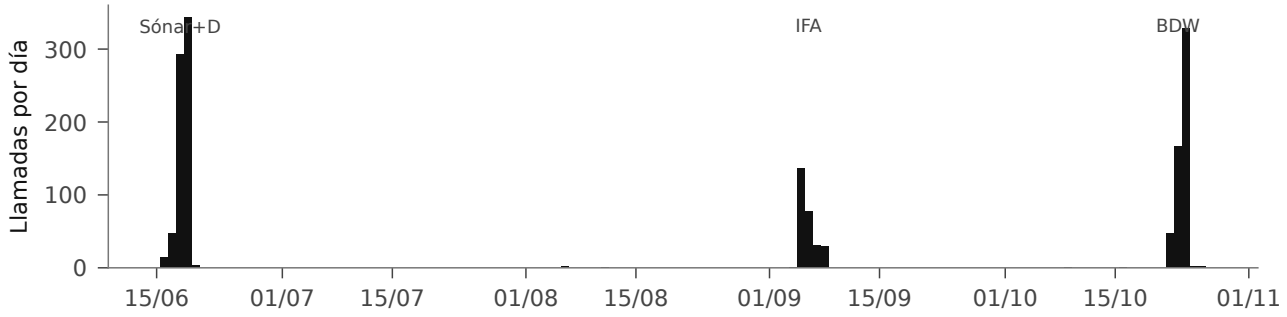


Figura 1. Llamadas por día en 2026.

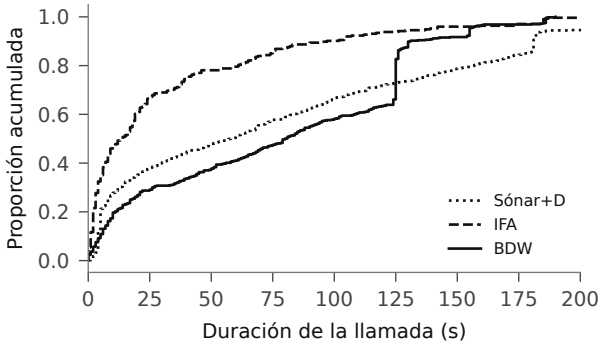


Figura 2. Distribución acumulada de la duración de la llamada por evento.

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.