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;

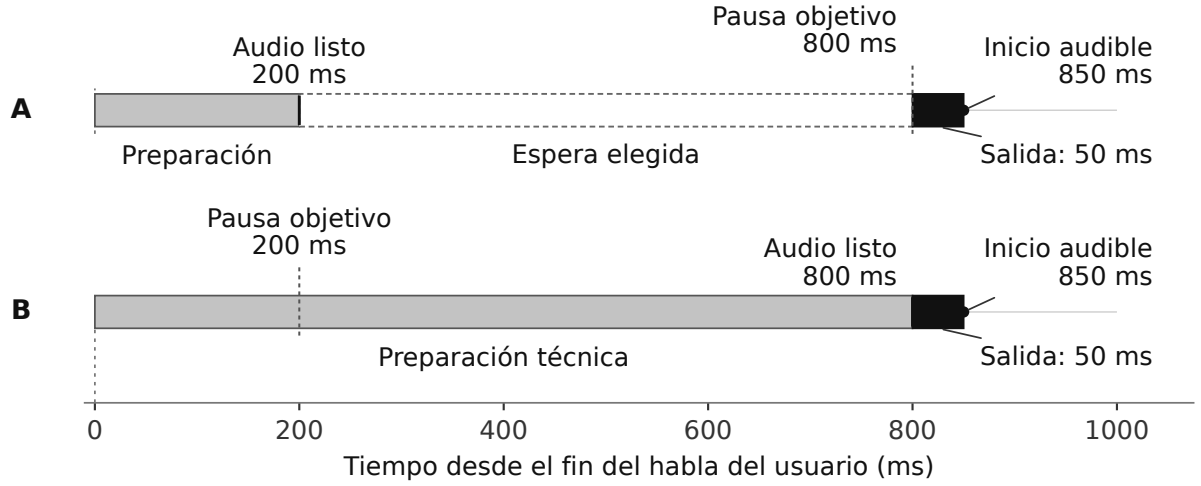


Figura 1. Dos situaciones ilustrativas producen el mismo inicio audible a los 850 ms. En A, la pausa objetivo determina cuándo liberar el audio; en B, lo determina su disponibilidad técnica. En ambos casos se aplica el máximo de la ecuación (2), más 50 ms de salida. Valores hipotéticos: no proceden de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