Inteligencia Artificial Centrada en el Usuario: Agentes Inteligentes para el Soporte en Deterioro Cognitivo

Álvaro Luque Buzón[1], Bienvenue Bukasa Muepu[1,2], Pablo Reina-Jiménez[1][0009−0008−8770−3767], Belén Vega-Márquez[1][0000−0002−2466−6486], and Cristina Rubio-Escudero[1][0000−0001−5732−9139]

[1] Escuela Técnica Superior de Ingeniería Informática, Universidad de Sevilla, España
[2] Université de Mbuji-Mayi, República Democrática del Congo {alvluqbuz, bukasa}@alum.us.es {prjimenez, bvega, crubioescudero}@.us.es

Resumen Las enfermedades neurodegenerativas como el Alzheimer provocan un deterioro progresivo de la memoria y la orientación, reduciendo la autonomía del paciente y aumentando la carga física y emocional de sus cuidadores. Los sistemas tradicionales de teleasistencia presentan limitaciones al depender de una activación consciente por parte del usuario, lo que los hace poco adecuados en fases moderadas de la enfermedad.
Este trabajo presenta un agente de inteligencia artificial para la monitorización pasiva y local de personas con deterioro cognitivo, diseñado para ofrecer asistencia contextualizada en tiempo real preservando la privacidad. La solución emplea una arquitectura cliente-servidor en la que una aplicación Android captura y segmenta el audio mediante detección de actividad de voz (VAD), mientras que el procesamiento intensivo se realiza en un servidor para reducir el consumo energético.
El backend integra un flujo de procesamiento compuesto por transcripción automática de voz (STT), identificación del hablante y un modelo de lenguaje de gran escala (LLM) que analiza el contexto mediante memoria a corto y largo plazo. Gracias a este enfoque, el sistema identifica episodios de confusión personalizados, distingue falsas alarmas y genera alertas con distintos niveles de gravedad para los cuidadores. Además, proporciona orientación verbal interactiva al paciente mediante síntesis de voz (TTS).
El prototipo, validado en un entorno local y privado, demuestra la viabilidad técnica de los agentes conversacionales como una alternativa más flexible y eficaz que los sistemas tradicionales basados en reglas, contribuyendo al desarrollo de soluciones de teleasistencia cognitiva de precisión.

Keywords: Deterioro cognitivo · Agentes · Procesamiento de lenguaje natural.

1. Introducción

El envejecimiento demográfico de la población mundial ha consolidado a las patologías neurodegenerativas como uno de los desafíos sociosanitarios más complejos del siglo XXI, de acuerdo con los informes globales de salud [2]. El núcleo

de esta problemática radica en la desarticulación de la vida cotidiana del paciente a causa de la pérdida de memoria de fijación o memoria episódica de corto plazo. Este síntoma sumerge al individuo en bucles cognitivos recurrentes, manifestados a través de preguntas iterativas y estados de angustia existencial al no ser capaz de reconciliar su percepción interna con la realidad objetiva temporal o espacial del entorno.

La monitorización de estas crisis tradicionalmente ha dependido de los cuidadores informales, quienes asumen un rol de supervisión continua que suele desembocar en el denominado síndrome del cuidador quemado [3]. Las alternativas tecnológicas existentes en el mercado, agrupadas bajo el espectro de la teleasistencia clásica, confían en que el usuario interactúe voluntariamente con un dispositivo (pulsadores de pánico, pantallas táctiles o aplicaciones móviles) o en la supervisión continuada del cuidador. No obstante, la manifestación clínica de una crisis de desorientación anula las capacidades ejecutivas del sujeto, inhabilitando el uso de interfaces de usuario tradicionales.

Para solventar este vacío, este trabajo introduce un paradigma asistencial proactivo e invisible, superando las limitaciones tradicionales documentadas en la literatura sobre ayuda asistencial [4]. A través del despliegue de un entorno de escucha pasiva en el hogar, el sistema intercepta las manifestaciones verbales espontáneas del paciente sin requerir interacción física. Mediante un procesamiento semántico local y avanzado, el ecosistema analiza si las expresiones emitidas denotan confusión, interviniendo acústicamente en tiempo real con locuciones personalizadas destinadas a mitigar la agitación y reorientar al individuo de forma inmediata.

En resumen, las contribuciones principales de este trabajo son:

- Una arquitectura cliente-servidor modular y pro-privacidad para asistencia proactiva basada en audio ambiental, que delega el procesamiento pesado en backend y minimiza el consumo de batería del dispositivo.
- Un detector de episodios que combina reglas configurables por el cuidador (frases gatillo y expresiones regulares) con razonamiento contextual mediante un LLM apoyado en memoria a corto plazo y diario resumido.
- Un estudio que combina una encuesta a 105 participantes y una entrevista semiestructurada a dos psicólogos clínicos, validando la pertinencia del enfoque y delimitando sus límites éticos y clínicos.
- Una validación funcional del prototipo mediante pruebas automáticas, pruebas experimentales con audio real y análisis de KPIs operativos.

2. Estado del Arte

La monitorización tecnológica de la agitación en personas con demencia ha sido abordada tradicionalmente mediante sensores ambientales y fisiológicos no lingüísticos, cuyos datos alimentan modelos de aprendizaje automático clásico (LSTM, gradient boosting, regresión logística) para inferir episodios de agitación [15], [16]. Au-Yeung et al. combinaron sensores de infrarrojos con actigrafía de

muñeca para la detección de agitación nocturna [16], mientras que otros estudios recogen variables ambientales durante semanas para entrenar modelos que se centran en detectar la agitación mientras ocurre, más que en anticiparla [15]. Estos trabajos comparten con nuestra la propuesta el interés por la monitorización pasiva y no invasiva, pero difieren en un punto esencial: no analizan el contenido del habla, sino correlatos indirectos de la conducta motriz o el entorno físico, por lo que no pueden discriminar la causa concreta de una crisis.

Una segunda línea de trabajo emplea LLMs para la interacción directa con personas con demencia, aunque con objetivos distintos al soporte ante crisis agudas. MindTalker, construido sobre GPT-4 y co-diseñado con terapeutas, permitió a ocho pacientes con demencia en fase temprana mantener un vínculo conversacional sostenido durante un mes o más, mitigando el aislamiento social [17]. Otros trabajos exploran agentes conversacionales para el cribado o diagnóstico temprano [18] o para el entrenamiento de cuidadores mediante pacientes simulados [19]. Estos sistemas requieren una interacción deliberada e iniciada por el usuario o el profesional, lo que los diferencia del enfoque propuesto. La línea más cercana a la propuesta de este trabajo combina audio ambiental continuo con modelos de lenguaje para la interpretación contextual en el hogar. DESAMO integra un LLM de audio embebido en un dispositivo doméstico para personas mayores, detectando emergencias como caídas o gritos de auxilio a partir de la monitorización pasiva de segmentos sonoros cortos [20]. En una línea similar, un pipeline reciente integra activación de voz, diarización de hablante, transcripción y reconocimiento de emociones para alimentar a un LLM que genera un informe diario al cuidador sobre orientación, temas delirantes, agitación y señales de seguridad [21]. Ambos trabajos validan la viabilidad de emplear LLMs sobre audio doméstico y comparten con la propuesta la arquitectura de percepción pasiva y la verificación de hablante. Sin embargo, se centran en la detección de emergencias explícitas mediante prompts genéricos (DESAMO) o en la generación de informes diferidos y no interactivos para el cuidador, sin cerrar el bucle con el paciente en el momento de la crisis. Ninguno de los dos incorpora una memoria biográfica estructurada del paciente como ancla de verdad contextual, elemento que en nuestro caso permite distinguir una pseudo-justificación de un hecho objetivamente correcto

En conjunto, la literatura revisada revela una brecha clara: los sistemas basados en sensores ambientales carecen de comprensión semántica del habla; los agentes conversacionales dependen de una interacción deliberada incompatible con el estado ejecutivo del paciente durante una crisis; y los sistemas de audio con LLM más recientes, aunque técnicamente próximos, se limitan a la detección de emergencias explícitas o a la generación de informes diferidos, sin un mecanismo de contraste contra la biografía del paciente ni una respuesta de reorientación verbal inmediata.

3. Metodología

Para fundamentar empíricamente la construcción del ecosistema, se optó por una metodología de investigación mixta. En primera instancia, se realizó una entrevista estructurada con dos psicólogos sanitarios de la Confederación Andaluza de Alzheimer (ConFEAFA). En segunda instancia, se distribuyó un cuestionario de caracterización cuantitativa que recopiló datos de 105 participantes anónimos, muchos de ellos cuidadores y familiares directos vinculados a patologías neurodegenerativas.

Cuadro 1. Jerarquía de requisitos funcionales extraídos del estudio demográfico.

Funcionalidad Solicitada por los Cuidadores	**Media**
Detección automatizada de confusión temporoespacial y alerta remota	3.15
Modelado analítico de las rutinas biográficas e historial del paciente	2.84
Canal de interacción conversacional natural por voz con el agente	2.74
Interfaz de personalización ágil del contexto del entorno clínico	2.51

Un hallazgo analítico fundamental de las encuestas reveló que una parte relevante de la muestra tenía contacto directo con personas con deterioro cognitivo (66,6 %) y que la repetición de preguntas y la desorientación fueron identificadas como situaciones especialmente relevantes para el diseño del sistema. Los profesionales clínicos enfatizaron que las intervenciones tempranas basadas en la validación de la emoción del paciente (en lugar de la confrontación directa con su error de memoria) reducen significativamente la ansiedad y evitan la evolución de la crisis hacia cuadros de agresividad o deambulación errática. A partir de este estudio, las prioridades funcionales del software se ponderaron estadísticamente, estableciendo el orden de relevancia técnica reflejado en el Cuadro 1.

El modelado conceptual de las interacciones queda formalizado en el diagrama de casos de uso de la Figura 1, donde se separan nítidamente las acciones no invasivas aplicadas sobre el Paciente de las herramientas de gestión e instrumentación clínica conferidas al Cuidador.

3.1. Arquitectura del Sistema y Privacidad por Diseño

La infraestructura adopta un enfoque topológico sustentado en los pilares de la Privacidad por Diseño (*Privacy by Design*) descritos por Cavoukian [6]. El canal de comunicación entre el dispositivo ubicado en la estancia y el nodo de cómputo local se establece de forma aislada del Internet público mediante Tailscale [12], una solución de red privada de confianza cero. De este modo, el tráfico entre ambos nodos queda encapsulado en un túnel cifrado, como se muestra en la Figura 2.

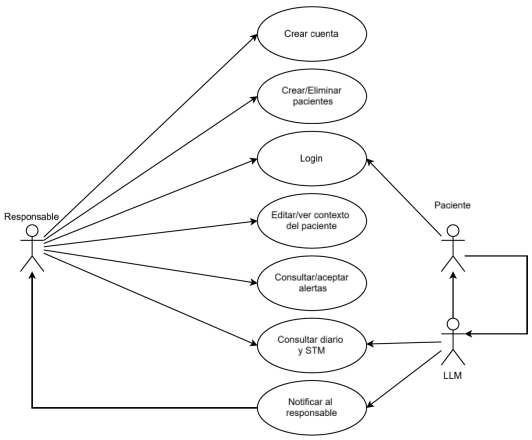


Figura 1. Diagrama de Casos de Uso del ecosistema asistencial.

El Cliente Ligero Móvil El componente frontal ha sido desarrollado utilizando React Native y el framework Expo. Su rol es actuar como un sensor acústico continuo de bajo consumo, capturando fragmentos de audio con lógica de activación por voz, reduciendo así el envío innecesario de audio al backend. La app registra fragmentos de audio mediante expo-av, calcula la amplitud en decibelios relativos a la escala completa (dBFS), segmenta el audio en ficheros temporales y envía al backend ficheros *.m4a/audio-mp4* únicamente cuando detecta un incremento sostenido por encima del ruido ambiental de fondo.

El Servidor (FastAPI Backend) El procesamiento de los flujos de audio es recibido en el servidor local a través de un backend asíncrono construido sobre FastAPI [10]. La elección de este framework responde a su arquitectura basada en el estándar ASGI y su soporte nativo para la programación basada en corrutinas nativas de Python (`async/await`), esenciales para evitar el bloqueo del bucle de eventos principal (*event loop*) cuando múltiples flujos temporales de audio ingresan concurrentemente. La orquestación y el catálogo completo de las tecnologías usadas se resumen en el Cuadro 2.

3.2. Pipeline Secuencial de Procesamiento de Audio

El flujo analítico que transforma el sonido de la persona en una decisión asistencial estructurada se ejecuta en el backend mediante un pipeline determinista compuesto por cinco pasos secuenciales y síncronos (Figura 3).

Estadio 1: Filtrado de Actividad de Voz (VAD) Al recibir un fragmento de audio, el backend lo procesa con faster-whisper, que integra Silero VAD [11], una red neuronal profunda optimizada para discernir la presencia de voz humana

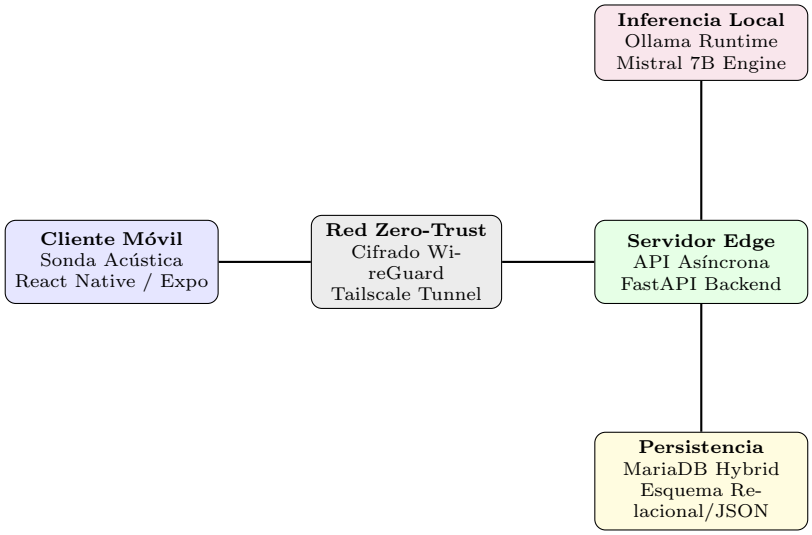


Figura 2. Diagrama de arquitectura técnica de la infraestructura distribuida Edge.

Cuadro 2. Catálogo tecnológico de la pila de software del servidor local.

Módulo del Sistema	**Solución Tecnológica**	**Tipo**
API Backend	FastAPI (Python 3.11)	MIT License
Persistencia Datos	MariaDB 11.2 (JSON Extensions)	GPL v2
Detección Silencios (VAD)	Silero VAD Runtime	MIT License
Transcripción (STT)	faster-whisper (CTranslate2)	MIT License
Extracción Embeddings	SpeechBrain ECAPA-TDNN	Apache 2.0
Inferencia Cognitiva (LLM)	Mistral 7B Instruct v0.2 (GGUF Q4) Modelo de inferencia configurable mediante variables de entorno	Apache 2.0

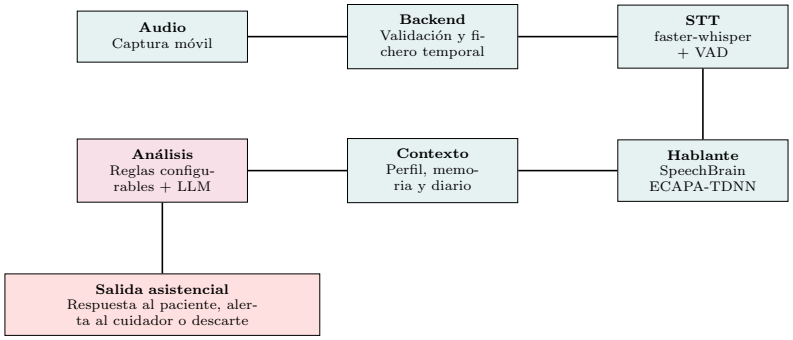


Figura 3. Pipeline analítico secuencial desde la captura acústica hasta la respuesta.

frente al ruido ambiente de fondo. Si la probabilidad matemática intrínseca de contener habla es inferior al umbral adaptativo del 75 %, el proceso descarta el archivo de la memoria RAM del servidor de inmediato, evitando el consumo innecesario de ciclos de cómputo en la GPU.

Estadio 2: Transcripción por Atención (Speech-to-Text) Los segmentos validados son procesados por el motor *faster-whisper*, una reimplementación del modelo Whisper de OpenAI [7]. Para mitigar el fenómeno de las alucinaciones repetitivas de texto en muestras con elevado silencio, se restringe la entropía de compresión a valores inferiores a 2,2 y se obliga al decodificador a utilizar búsquedas de tipo *Beam Search* con un ancho de haz parametrizado en 5 candidatos.

Estadio 3: Diarización y Verificación Biométrica Para blindar el sistema frente a falsos positivos derivados de conversaciones emitidas por visitas, cuidadores informales o dispositivos multimedia presentes en la sala (televisión, radio), se integró el modelo preentrenado *SpeechBrain ECAPA-TDNN* (Time-Delay Neural Network) [8]. Este componente extrae un vector embebido de identidad acústica de 192 dimensiones a partir del espectrograma de la señal. El sistema contrasta este vector en tiempo real contra un embedding de referencia $\boldsymbol{R}$ previamente capturado de la voz del paciente. La validación se rige por el cálculo de la similitud del coseno:

$$\text{Similitud}(\boldsymbol{R}, \boldsymbol{A}) = \frac{\boldsymbol{R} \cdot \boldsymbol{A}}{\|\boldsymbol{R}\|\|\boldsymbol{A}\|} = \frac{\sum_{i=1}^{n} R_i A_i}{\sqrt{\sum_{i=1}^{n} R_i^2}\sqrt{\sum_{i=1}^{n} A_i^2}}$$

El sistema utiliza umbrales configurables para etiquetar la voz detectada. Por defecto,
`SPEAKER_DIARIZATION_THRESHOLD = 0.40` marca un fragmento como `PACIENTE` cuando la similitud supera dicho umbral, mientras que `SPEAKER_UNCERTAIN_THR = 0.30` permite clasificar casos intermedios como `PACIENTE?`. Por debajo de este segundo umbral, el habla se etiqueta como `OTRO` o se considera que no existe voz fiable del paciente. Adicionalmente, si se detecta eco del TTS, se etiqueta esta transcripción como `ASISTENTE`. Esta clasificación permite evitar alertas no justificadas o trasladar la incertidumbre al flujo posterior, según el caso.

El responsable puede registrar varias muestras de voz del paciente, hasta un máximo de 10. Durante la diarización, el sistema compara el fragmento analizado con las muestras activas disponibles y emplea la referencia que obtiene mayor similitud.

Estadio 4: Procesamiento Semántico Híbrido La cadena de texto depurada se somete a una doble capa lógica:

1. **Capa Determinista de Emergencias:** Se evalúa mediante expresiones regulares precompiladas dirigidas a interceptar exclamaciones de auxilio o

accidentes físicos explícitos (p. ej., `/(auxilio|socorro|ayuda|me caí|me duele)`). De dar positivo, el backend elude el paso por el modelo de lenguaje, reduciendo a cero la latencia de inferencia cognitiva y despachando una alerta push crítica inmediata al cuidador.

2. **Capa de Inferencia Cognitiva Locacional:** Si el enunciado no contiene términos explícitos de emergencia pero denota posible desorientación, se transfiere a *Mistral 7B Instruct* [9] a través de la interfaz de ejecución local *Ollama* [5]. Mediante un prompt estructural, se inyectan variables del sistema (hora actual, día de la semana) junto con las directrices del JSON biográfico. Se restringe formalmente la salida del LLM exigiéndole un objeto estructurado válido bajo formato JSON como el detallado en el Listado 1.1.

Listing 1.1. Estructura JSON estricta de salida para la validación de la capa cognitiva

```
{
  "crisis_detectada": true,
  "nivel_severidad": 4,
  "motivo_analisis": "El paciente manifiesta intenciones de acudir a su
      puesto de trabajo de carpinteria a pesar de estar jubilado desde
      2004 y ser las 02:45 horas.",
  "respuesta_calmante": "Manuel, no te preocupes, ya has terminado todos
      tus trabajos de carpinteria por hoy. Son las tres de la manana y tu
       taller esta cerrado y seguro. Puedes volver a descansar."
}
```

Estadio 5: Síntesis de Voz y Mitigación (Text-to-Speech) Cuando el atributo `crisis_detectada` evalúa como `true`, el servidor despacha en paralelo la notificación remota al terminal del cuidador y devuelve la cadena `respuesta_calmante` al dispositivo móvil del paciente. Este último ejecuta el motor nativo `expo-speech`, configurando una velocidad de locución ralentizada y una entonación calmada para proyectar el mensaje de reorientación en la sala.

3.3. Metodología de Evaluación y Dataset de Control

Ante la inexistencia de repositorios de datos acústicos públicos de libre acceso que recojan grabaciones ambientales reales de pacientes con demencia en fases críticas de agitación, se diseñó un entorno metodológico de simulación controlada. Se elaboró una colección de pruebas basadas en las indicaciones del equipo clínico de ConFEAFA. Esta colección de datos de control se segmentó en dos categorías operacionales bien diferenciadas:

- **Subconjunto de Control Neutro:** Oraciones emitidas de manera recurrente en el hogar que carecen de indicadores de amnesia, tales como peticiones funcionales (p. ej., "¿Me traes un vaso de agua?", "Hace frío aquí"). El comportamiento esperado de la arquitectura ante estas muestras era el filtrado pasivo y la ausencia de disparos de alarma.

- **Subconjunto de Crisis de Desorientación:** Construcciones complejas que reproducían el patrón clínico de la pseudo-justificación biográfica y el desfase temporal. Se programaron combinaciones específicas, alterando dinámicamente los metadatos temporales inyectados en la cabecera HTTP de la petición de FastAPI. Un caso representativo consistió en la frase "Tengo que salir a abrir la escuela, los niños están llegando", procesada bajo una estampa de tiempo artificial de las 23:45 horas, habiendo cargado previamente en el objeto biográfico del paciente su antiguo rol laboral como conserje escolar durante su juventud.

Para simular de forma fidedigna las condiciones acústicas reales de una vivienda de dimensiones estándar, las oraciones del dataset fueron locutadas en un laboratorio acústico en presencia de ruidos parasitarios superpuestos, incluyendo sonidos de menaje de cocina, pasos y la emisión de fondo de un aparato de televisión a un volumen medio de 65 dB.

4. Resultados

La infraestructura física dispuesta para albergar el nodo de computación local consistió en un ordenador anfitrión de gama media provisto de un procesador Intel Core i7-13700H, 16 GB de memoria RAM en arquitectura DDR5 y una tarjeta gráfica comercial *NVIDIA RTX 4060* dotada de 8 GB de VRAM dedicada.

4.1. Métricas de Latencia Temporal

Para que el sistema asistencial sea clínicamente viable, la latencia de respuesta de extremo a extremo (T_{total}) no debe exceder un límite tolerable de 20 segundos. Si la respuesta calmante tarda más de este umbral, el estado de confusión del paciente corre el riesgo de escalar o podría abandonar físicamente la habitación, invalidando el propósito de la intervención proactiva. La descomposición analítica de la latencia total responde a la siguiente función:

$$T_{\text{total}} = T_{\text{VAD}} + T_{\text{STT}} + T_{\text{Diarización}} + T_{\text{LLM}} + T_{\text{Red}}$$

Los resultados cronometrados en las pruebas de estrés reportaron una latencia media global controlada de 3.8 segundos ($T_{\text{total}} = 3{,}8$ s). El desglose empírico por módulo arrojó que $T_{\text{VAD}} = 0{,}12$ s, $T_{\text{STT}} = 0{,}85$ s, $T_{\text{Diarización}} = 0{,}43$ s, y la inferencia del modelo de lenguaje consumió $T_{\text{LLM}} = 2{,}40$ s. Estos datos confirman la idoneidad del hardware de consumo para soportar arquitecturas de IA distribuida en tiempo real.

4.2. Precisión Semántica de Clasificación

La efectividad de la clasificación semántica realizada por el motor Mistral 7B se contrastó mediante el cálculo de la matriz de confusión sobre el conjunto

de datos de control simulado. El modelo alcanzó una precisión global del 85 % en la discriminación de crisis frente a conversaciones neutras. La tasa de falsos positivos se situó en un 4.5 %, mitigando de esta forma el riesgo de saturación de alarmas o fatiga por alertas en el terminal móvil del cuidador. Los indicadores clave de rendimiento (KPI) se resumen en el Cuadro 3.

Cuadro 3. Matriz de Indicadores Clave de Rendimiento (KPI) y resultados del sistema.

KPI Evaluado	**Objetivo Clínico**	**Valor Logrado**
Latencia de Respuesta Completa	$< 20\,\text{s}$	3,8 s
Precisión de Clasificación Semántica	$\geq 80\,\%$	85 %
Tasa de Falsos Positivos Acústicos	$< 10\,\%$	4,5 %
Autonomía de Batería (Sonda Móvil)	$\geq 12\,\text{h}$	16,5 h

4.3. Análisis de Riesgos Técnicos

Durante la fase de integración, se identificaron contingencias relevantes asociadas al despliegue local de modelos. Para combatirlas, el backend incorpora un semáforo asíncrono para acotar el número de peticiones de audio procesadas simultáneamente y bloqueos internos para serializar el acceso a modelos cargados en memoria, como el transcriptor y el codificador de voz. Esta estrategia reduce la probabilidad de colisiones, saturación de recursos y errores por inferencia concurrente (Cuadro 4).

Cuadro 4. Matriz de Riesgos e instrumentación de contingencias técnicas.

Riesgo	**Prob.**	**Impacto**	**Mitigación**
Concurrencia de audio	Media	Crítico	Semáforo asíncrono y bloqueos de modelo
Voces externas / TV	Media	Mayor	SpeechBrain y etiquetas de hablante
Eco TTS del asistente	Media	Medio	Filtro de eco y etiqueta `ASISTENTE`

5. Conclusiones y Trabajo Futuro

El presente trabajo valida que la combinación de STT, verificación de hablante, memoria contextual, reglas configurables y LLM local permite construir un sistema asistencial proactivo para el cuidado de personas con deterioro cognitivo. El sistema muestra que es viable priorizar privacidad y procesamiento local sin renunciar a una intervención contextualizada ante episodios de desorientación. Todo esto con un hardware relativamente modesto.

A pesar de los buenos resultados obtenidos, el sistema presenta limitaciones técnicas y éticas para su despliegue en entornos reales. Entre los principales riesgos destaca el solapamiento acústico, donde conversaciones del paciente y audios procedentes de televisores u otras fuentes pueden mezclarse, generando falsos positivos en la interpretación del LLM [13]. Asimismo, la precisión del modelo depende de la calidad de la información clínica proporcionada por el cuidador, ya que un contexto incompleto puede incrementar los falsos negativos. Finalmente, la monitorización continua plantea desafíos en materia de privacidad, por lo que resulta imprescindible garantizar el consentimiento informado, la transparencia en el tratamiento de los datos y mecanismos que permitan desactivar temporalmente el sistema.

Como líneas preferentes de trabajo futuro se contemplan dos vías de investigación complementarias:

- **Hardware asistencial dedicado:** Trasladar la captura de audio a un dispositivo dedicado, preferentemente un wearable o colgante (similar a proyectos open-source ya existentes como Friend [14]), que acompañe al paciente de forma discreta y requiera una interacción mínima. En este escenario, el dispositivo actuaría como cliente ligero encargado de la adquisición y envío de audio, mientras que las tareas computacionalmente intensivas continuarían delegándose en un nodo local de procesamiento. La separación entre cliente y backend, así como la flexibilidad de configuración incorporada en el sistema, permitirían abordar esta migración sin rediseñar el núcleo de inferencia ni comprometer la autonomía energética del dispositivo portado por el paciente.
- **Ensayos clínicos controlados:** Iniciar una fase de validación clínica regulada en entornos reales de convivencia tutorizada en colaboración con facultativos del sistema sanitario público y asociaciones de pacientes, con el fin de medir estadísticamente el impacto del asistente en la reducción de los niveles de cortisol en sangre del cuidador informal y la atenuación de los brotes de ansiedad en los usuarios afectados.

Agradecimientos. Esta publicación es parte del proyecto de I+D+i PID2023 -146037OB-C21, financiado por MICIU/AEI/10.13039/501100011033/.

Referencias

1. Álvaro Luque Buzón. *Integración de agentes de Inteligencia Artificial distribuidos para la asistencia proactiva y la atenuación del deterioro cognitivo en el entorno doméstico.* Trabajo Fin de Grado, Escuela Técnica Superior de Ingeniería Informática, Universidad de Sevilla, 2026.
2. Organización Mundial de la Salud (OMS). *Global status report on the public health response to dementia.* World Health Organization, Ginebra, 2021.
3. Schulz R, Beach SR. *Caregiving as a Risk Factor for Mortality: The Caregiver Health Effects Study.* JAMA, Vol. 282, No. 23, pp. 2215–2219, 1999.
4. D. J. Cook y S. K. Das. *Smart Environments: Technology, Protocols and Applications.* John Wiley & Sons, New Jersey, 2005.

5. Ollama Team. *Ollama: Local library manager for large language models*. URL: `https://ollama.com`, 2024.
6. A. Cavoukian. *Privacy by Design: The 7 Foundational Principles*. Information and Privacy Commissioner of Ontario, Canadá, 2009.
7. A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey e I. Sutskever. *Robust Speech Recognition via Large-Scale Weak Supervision*. arXiv preprint arXiv:2212.04356, 2022.
8. M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch et al. *SpeechBrain: A General-Purpose Speech Toolkit*. arXiv preprint arXiv:2106.04624, 2021.
9. A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas et al. *Mistral 7B*. arXiv preprint arXiv:2310.06825, 2023.
10. S. Ramírez. *FastAPI framework, high performance, easy to learn, fast to code, ready for production*. URL: `https://fastapi.tiangolo.com/`, 2018.
11. Silero Team. *Silero VAD: pre-trained enterprise-grade Voice Activity Detector*. GitHub Repository, URL: `https://github.com/snakers4/silero-vad`, 2021.
12. Tailscale Inc. *Tailscale: WireGuard-based zero-configuration mesh VPN*. URL: `https://tailscale.com`, 2020.
13. F. Pérez e I. Ribeiro. *Ignore Previous Instructions: Prompt Injection Attacks on Large Language Models*. arXiv preprint arXiv:2211.09527, 2022.
14. itsliupeng. *Friend: the world's leading open-source AI wearable*. GitHub Repository, URL: `https://github.com/itsliupeng/Friend`, 2025.
15. Benchmarking Early Agitation Prediction in Community-Dwelling People with Dementia Using Multimodal Sensors and Machine Learning. arXiv preprint arXiv:2506.06306, 2025.
16. W.-T. Au-Yeung, L. Miller, Z. Beattie, J. Kaye. Passive Infrared Motion Sensors Improved the Detection Accuracy of Nocturnal Agitation. *Innovation in Aging*, Vol. 5, Suppl. 1, p. 908, 2021.
17. MindTalker: Navigating the Complexities of AI-Enhanced Social Engagement for People with Early-Stage Dementia. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*, Honolulu, HI, USA, mayo de 2024.
18. J. Favela, D. Cruz-Sandoval, M. O. Parra. Conversational agents for dementia using large language models. En *2023 Mexican International Conference on Computer Science (ENC)*, pp. 1–7, IEEE, 2023.
19. An Interactive LLM-Based Simulator for Dementia-Related Activities of Daily Living. arXiv preprint arXiv:2603.29856, 2026.
20. DESAMO: A Device for Elder-Friendly Smart Homes Powered by Embedded LLM with Audio Modality. arXiv preprint arXiv:2508.18918, 2025.
21. Affordable Audio Hardware and Artificial Intelligence Can Transform the Dementia Care Pipeline. *Algorithms*, 18(12):787, 2025. DOI: 10.3390/a18120787.