Revista Española
de Educación Médica

Informe radiológico estructurado asistido por modelos de lenguaje en residentes de Radiología: piloto de implementación en Urgencias.

Large language model-assisted structured reporting in Radiology residents: an implementation pilot study in Emergency Radiology.

Clemente García-Hidalgo[1*], José Antonio Consentino Hernández[2], José Vicente Cayuela Espí[3], Gonzalo Pagán Vicente[4], Juana María Plasencia Martínez[5], Ana Blanco Barrio[6], Gloria Pérez Hernández[7], Ana Moreno Pastor[8]

[1] Servicio de Radiología, Hospital Universitario Morales Meseguer, Murcia, España. clemente292@gmail.com, https://orcid.org/0009-0001-6672-2714

[2] joseconsentinohernandez@gmail.com; [3] josevicayuela@gmail.com; [4] gonzalopv97@gmail.com; [5] plasen79@gmail.com; [6] anablancowhite@gmail.com; [7] gloriapeh93@gmail.com; [8] anamp78@gmail.com

* Correspondencia: Clemente García-Hidalgo clemente292@gmail.com

Recibido: 3/1/26; Aceptado: 20/1/26; Publicado: 22/1/26

Resumen

Objetivo: Evaluar si un sistema de informe estructurado asistido por Large Language Models puede integrarse de forma práctica en el trabajo de residentes de Radiología durante las guardias. Como objetivos secundarios: describir preferencias de formato mediante evaluación ciega, caracterizar diferencias lingüísticas entre informes manuales y asistidos, e identificar riesgos percibidos para un estudio confirmatorio. **Métodos:** Estudio piloto con dos componentes. En la fase prospectiva, cuatro residentes generaron 480 informes alternando redacción manual y asistida por LLM (Custom GPT-4o). En paralelo, se analizaron 200 informes anonimizados de adjuntos para contextualizar las métricas. Se empleó una encuesta ad hoc tipo Likert (seis dimensiones) y se calcularon métricas de clasificación y perplejidad como indicadores descriptivos. **Resultados:** La herramienta fue bien recibida. Las medianas Likert oscilaron entre 4,75 y 4,90 sobre 5. Los residentes distinguieron con precisión qué informes habían sido asistidos (F1 = 0,92), lo que sugiere una huella formal reconocible. Se observó sesgo de autoatribución en las preferencias ciegas. La perplejidad difirió entre residentes y adjuntos (p = 0,03), apuntando a mayor regularidad en profesionales experimentados. **Conclusiones:** Los hallazgos respaldan la integración inicial del asistente en el circuito de guardias. El interés reside en su función de andamiaje para estandarizar la comunicación entre residentes y médicos peticionarios, no en automatizar el razonamiento diagnóstico.

Palabras clave: Inteligencia artificial; Grandes modelos de lenguaje; Informe estructurado; Educación médica; Radiología; Residentes; Radiología de urgencias; Estudio piloto

Abstract

Objective: To evaluate whether a structured reporting system assisted by Large Language Models (LLMs) can be practically integrated into the work of radiology residents during on-call shifts. Secondary objectives included: describing format preferences through blind evaluation, characterizing linguistic differences between manual and LLM-assisted reports, and identifying perceived risks for a confirmatory study. **Methods**: A two-component pilot study was conducted.

In the prospective phase, four residents generated 480 reports, alternating between manual and LLM-assisted writing (Custom GPT-4o). In parallel, 200 anonymized reports from attending physicians were analyzed to contextualize the metrics. An ad hoc Likert-type survey (six dimensions) was used, and classification and perplexity metrics were calculated as descriptive indicators. **Results:** The tool was well received. Median Likert scores ranged from 4.75 to 4.90 out of 5. Residents accurately distinguished which reports had been assisted (F1 = 0.92), suggesting a recognizable formal signature. Self-attribution bias was observed in blinded preferences. Perplexity differed between residents and attending physicians ($p = 0.03$), suggesting greater regularity among experienced professionals. **Conclusions:** The findings support the initial integration of the assistant into the on-call system. The value lies in its scaffolding function to standardize communication between residents and requesting physicians, not in automating diagnostic reasoning.

Keywords: Artificial intelligence; Large language models; Structured report; Medical education; Radiology; Residents; Emergency radiology; Pilot study

1. Introducción

La inteligencia artificial está transformando la práctica radiológica contemporánea (1-2), con aplicaciones que van más allá de lo interpretativo. Entre ellas, el informe estructurado asistido por grandes modelos de lenguaje (LLM) (3-4) representa una oportunidad para mejorar la comunicación radiológica, especialmente en urgencias, donde la rapidez y precisión son críticas. El informe radiológico es el principal producto comunicativo del acto radiológico. Desde la perspectiva formativa, es un resultado observable de competencias en comunicación clínica. Sin embargo, la enseñanza explícita de su redacción es heterogénea. Muchos residentes aprenden por exposición, imitación y retroalimentación variable (5). Esta brecha justifica intervenciones que aporten estructura, criterios explícitos y retroalimentación durante la práctica. Más allá de las diferencias estilísticas individuales, la estructura del informe codifica el razonamiento subyacente a la interpretación. La radiología de urgencias presenta características que hacen relevante la estandarización: comunicación rápida, toma de decisiones inmediatas y consistencia terminológica.

En educación médica, la calidad del informe no es solo un producto asistencial; constituye evidencia del razonamiento clínico y la comunicación profesional. La literatura propone que el informe estructurado puede emplearse como soporte para la formación basada en competencias, permitiendo evidenciar hitos y EPAs mediante criterios observables (6). También puede reducir omisiones y mejorar la satisfacción de los clínicos (7). Nuestro trabajo evalúa si un asistente basado en LLM puede actuar como andamiaje cognitivo para residentes, y qué señales lingüísticas objetivas emergen de su uso.

Planteamos esta intervención como aprendizaje en el puesto de trabajo. El mecanismo educativo opera a través del andamiaje (un guion estable de secciones reduce la carga cognitiva extrínseca), la práctica deliberada (repetición con criterios explícitos) y la retroalimentación inmediata (señales de estilo que el residente contrasta con la revisión del adjunto). El producto, el informe, se convierte así en evidencia de competencias comunicativas, sin pretender que el sistema realice razonamiento diagnóstico.

Este estudio se plantea como investigación de implementación en guardias reales. El reto no es si el modelo genera texto, sino si puede integrarse de forma segura en la actividad del residente. El objetivo es estimar si la integración es práctica y aceptable, y delimitar riesgos para orientar un ensayo posterior. La incorporación de IA generativa a los servicios obliga a incluir competencias de uso crítico en la residencia: verificación, gestión de sesgos y comunicación segura. Evaluar implementaciones tempranas aporta evidencia para diseñar esa formación. El objetivo principal fue evaluar si el asistente puede integrarse en la práctica de residentes de urgencias. Los objetivos

secundarios, exploratorios, fueron describir preferencias de formato, caracterizar diferencias lingüísticas entre informes manuales y asistidos, e identificar riesgos para un estudio confirmatorio.

2. Métodos

2.1 Marco de evaluación

El estudio priorizó la integración en la práctica real y la identificación de barreras antes de evaluar eficacia. Este enfoque es coherente con la secuencia en ciencia de la implementación: pilotos como prerrequisito para estudios confirmatorios. Los desenlaces corresponden principalmente a Kirkpatrick nivel 1 (reacción) y, de forma limitada, a nivel 2 (aprendizaje percibido). Por diseño, evitamos afirmaciones sobre cambios conductuales o impacto asistencial.

2.2 Diseño del estudio

Estudio piloto con dos componentes. En la fase prospectiva, residentes generaron informes alternando redacción manual y asistida por LLM. En el análisis retrospectivo, se utilizó un corpus independiente de informes de adjuntos anonimizados para contextualizar métricas. Dada la naturaleza de piloto, el estudio describe factibilidad y genera hipótesis, sin establecer relaciones causales. Aunque el corpus incluye múltiples informes por autor, el n efectivo viene dado por el número de autores (cuatro residentes y el conjunto de adjuntos). Los contrastes estadísticos se interpretan como exploratorios.

2.3 Consideraciones éticas y seguridad

Los informes fueron anonimizados mediante un proceso de doble verificación: primero, eliminación automática de identificadores directos y cuasi-identificadores mediante expresiones regulares (nombres, números de historia clínica, fechas de nacimiento, direcciones); segundo, revisión manual para detectar referencias contextuales potencialmente identificativas. La herramienta se usó como asistente de redacción; la responsabilidad diagnóstica permaneció en el residente y el adjunto. La interacción con el LLM excluyó datos identificativos del paciente. Las medidas operativas incluyeron: uso exclusivo como apoyo de redacción, revisión previa a validación y prohibición de introducir identificadores. El asistente no interpreta imágenes ni introduce hallazgos; estructura el texto introducido por el residente.

La responsabilidad del contenido final permaneció en el profesional, coherentemente con el principio de supervisión humana (8). Se explicitaron los parámetros de uso y las salvaguardas del asistente siguiendo recomendaciones emergentes para estudios con modelos generativos, en particular los principios de transparencia, supervisión humana y delimitación del rol del sistema propuestos en TRIPOD-LLM, así como los ítems aplicables de la checklist CLAIM (9). En concreto, se definió el propósito no diagnóstico del modelo, se describió su configuración y contexto de uso, se garantizó la anonimización de los datos, se mantuvo la responsabilidad clínica en el profesional humano y se evitó cualquier automatización de la toma de decisiones diagnósticas.

Este estudio corresponde a una evaluación educativa y de implementación, sin intervención clínica ni modificación de la atención al paciente. Todos los informes fueron anonimizados antes de su análisis. Conforme al Real Decreto 957/2020 y la Ley 14/2007 de Investigación Biomédica española, los estudios que emplean datos completamente anonimizados y no implican intervención sobre pacientes están exentos de aprobación por comité de ética. El trabajo se desarrolló respetando los principios de la Declaración de Helsinki.

2.4 Desarrollo de la herramienta

Se desarrolló una herramienta personalizada basada en Custom GPT-4o (OpenAI, versión mayo 2024), diseñada mediante ingeniería de prompts sin programación. El sistema se configuró con los siguientes parámetros: temperature 0,3 (para favorecer consistencia sobre creatividad), max_tokens 2048, y sin ajuste fino (fine-tuning). La documentación de referencia incluyó las Guías de la European Society of Emergency Radiology (10), plantillas RadReport de la Radiological Society of North America (11), y el documento de Léxico conflictivo de la Sociedad Española de Radiología Médica (12).

Las funcionalidades implementadas fueron: detección automática de modalidad y región anatómica, sugerencias contextuales durante la escritura, recordatorios para reducir omisiones, mejora del reconocimiento de voz, y generación aumentada con documentación de referencia (RAG) (13). La estructura del informe estructurado generado seguía cinco secciones estandarizadas: Técnica (modalidad, contraste, protocolo), Hallazgos (descripción sistemática por regiones anatómicas), Comparación (con estudios previos cuando disponibles), Impresión diagnóstica, y Recomendaciones. La tabla 1 detalla la estructura del informe y su correspondencia con las competencias comunicativas evaluadas.

Tabla 1. Estructura del informe estructurado y competencias comunicativas asociadas.

Sección del informe	Contenido	Competencia comunicativa
Técnica	Modalidad, contraste, protocolo	Precisión terminológica
Hallazgos	Descripción sistemática por regiones	Completitud, organización
Comparación	Cambios respecto a estudios previos	Integración longitudinal
Impresión diagnóstica	Síntesis y diagnóstico diferencial	Razonamiento clínico
Recomendaciones	Seguimiento, estudios adicionales	Orientación clínica

2.5 Población y procedimiento

Cuatro residentes de radiología (tres R2 y un R4) participaron entre junio y diciembre de 2025. El tamaño muestral se determinó por factibilidad operativa: correspondía a los residentes del servicio que aceptaron participar voluntariamente en el piloto (4 de 11 residentes totales, 36%). Esta tasa de participación es razonable para una intervención que requería compromiso adicional durante las guardias. El tamaño es coherente con las recomendaciones para estudios piloto de factibilidad, donde $n \geq 12$ observaciones por condición se considera suficiente para estimar variabilidad (14). Cada residente generó 120 informes (60 asistidos, 60 manuales), totalizando 480. Los casos se asignaron consecutivamente, alternando el método. No se estableció contrabalanceo formal, lo que constituye una limitación reconocida. La tabla 2 resume las características de los participantes y la distribución de informes.

Tabla 2. Características de los participantes y distribución de informes.

Característica	Res. A	Res. B	Res. C	Res. D	Total
Año de residencia	R2	R2	R2	R4	-
Informes asistidos	60	60	60	60	240
Informes manuales	60	60	60	60	240
Total por residente	120	120	120	120	480
Modalidades	TC (65%), Radiografía (30%), Ecografía (5%)				

2.6 Intervención formativa

Se proporcionó una inducción breve (30 minutos) centrada en el objetivo educativo (mejorar estructura, completitud y claridad), uso seguro, y criterios de calidad del informe en urgencias. La

herramienta se empleó durante la redacción; posteriormente, el informe siguió el circuito habitual de supervisión por el radiólogo adjunto de guardia.

2.7 Evaluación multidimensional

Se diseñó una encuesta Likert (escala 1-5) evaluando seis dimensiones: Usabilidad, Utilidad Clínica, Eficiencia, Fiabilidad, Impacto Educativo percibido y Satisfacción General. Los ítems evalúan calidad comunicativa, no capacidad de detección clínica. La encuesta fue desarrollada mediante consenso del equipo investigador, revisando instrumentos similares en la literatura de implementación tecnológica en salud (15), aunque sin validación psicométrica formal dado el carácter exploratorio del estudio. Se realizó una prueba preliminar con dos residentes no participantes para verificar comprensibilidad de los ítems. La encuesta se concibió como retroalimentación formativa para capturar señales tempranas. Dada la naturaleza piloto, no se asume validez psicométrica; los resultados son descripción exploratoria que deberá confirmarse con instrumentos validados en estudios posteriores.

2.8 Fase de clasificación

Para explorar si el LLM deja una huella formal perceptible, se realizó una tarea ciega de identificación (asistido vs. manual) sobre los 480 informes. Los informes se preprocesaron eliminando encabezados y firmas, y se presentaron en orden aleatorizado mediante software dedicado. Cada evaluador clasificó únicamente informes de otros autores para evitar sesgo de reconocimiento. Se calcularon exactitud, precisión, sensibilidad y F1 como métricas descriptivas de la capacidad humana de discriminación, no como indicadores de calidad diferencial.

2.9 Preferencia ciega

Se diseñó un experimento de preferencia ciega para evaluar sesgos de atribución. Para cada caso clínico con doble informe (asistido y manual), se generaron pares anonimizados presentados en orden aleatorizado. Se eliminaron marcadores identificativos de autoría y método. Cada evaluador valoró únicamente pares de informes generados por otros residentes, nunca propios, para evitar reconocimiento estilístico. Se instruyó a elegir el informe preferido atendiendo exclusivamente a claridad, completitud y organización, sin conocimiento de qué método había generado cada uno. Este componente fue exploratorio, orientado a detectar posibles sesgos de atribución más que a establecer superioridad de un método.

2.10 Análisis de perplejidad

Se recuperaron 200 informes de radiólogos adjuntos de 2025, anonimizados y normalizados. La perplejidad se utilizó como métrica exploratoria de regularidad lingüística, no como desenlace educativo ni indicador de calidad clínica. El cálculo se realizó mediante API de OpenAI (gpt-4-1106-preview). Es importante señalar que una menor perplejidad indica únicamente mayor previsibilidad del texto para el modelo de lenguaje empleado, sin implicar necesariamente mejor calidad comunicativa o diagnóstica. Se compararon los valores entre informes de residentes (n = 480) y adjuntos (n = 200) mediante Mann-Whitney, test no paramétrico apropiado para datos no normalmente distribuidos. La perplejidad fue calculada como "Perplejidad" = exp {left (- {1 } over {N } sum from {i=1 } to {N } {log {p} left ({w} rsub {i} right) } right)} , "donde " p left ({w} rsub {i} right) " es la probabilidad asignada por el modelo al token " i.

2.11 Análisis estadístico

Los resultados se presentan de forma descriptiva. Las escalas Likert se resumen mediante mediana y rango intercuartílico, apropiados para datos ordinales. No se emplearon pruebas paramétricas. Las métricas computacionales se interpretan como análisis exploratorios. Los análisis se realizaron con Python 3.11 (bibliotecas pandas, scipy, scikit-learn).

3. Resultados

3.1 Evaluación de aceptabilidad

La tabla 3 muestra los resultados Likert. Todas las dimensiones obtuvieron medianas superiores a 4,7 sobre 5. La figura 1 presenta gráficamente estos resultados. La Satisfacción General alcanzó la mediana más alta (4,90), seguida de Usabilidad y Eficiencia (4,85). La menor fue Fiabilidad (4,75), lo que podría reflejar una cautela apropiada de los residentes respecto a la necesidad de verificación del contenido generado.

3.2 Análisis de clasificación

La tabla 4 muestra las métricas del experimento de identificación de huella lingüística, donde los evaluadores humanos clasificaron los informes. El F1-score de 0,92 indica que los evaluadores humanos pudieron distinguir con alta precisión entre informes asistidos y manuales. Esto sugiere características lingüísticas o estructurales diferenciables, aunque estas diferencias son formales y no implican diferencias en calidad clínica o utilidad diagnóstica. La capacidad de distinguir entre métodos podría deberse a mayor consistencia estructural, uso más sistemático de terminología estandarizada, o patrones de organización característicos del asistente.

Tabla 3. Evaluación de aceptabilidad del informe estructurado asistido por LLM.

Dimensión	Eval. A	Eval. B	Eval. C	Eval. D	Mediana (RIC)
Usabilidad	4,8	4,9	4,7	5,0	4,85 (4,7-5,0)
Utilidad Clínica	4,9	4,7	4,8	4,9	4,80 (4,7-4,9)
Eficiencia	4,7	4,8	4,9	4,9	4,85 (4,7-4,9)
Fiabilidad	4,6	4,8	4,7	4,9	4,75 (4,6-4,9)
Impacto Educativo	4,8	4,9	4,6	4,8	4,80 (4,6-4,9)
Satisfacción General	5,0	4,8	4,9	4,9	4,90 (4,8-5,0)

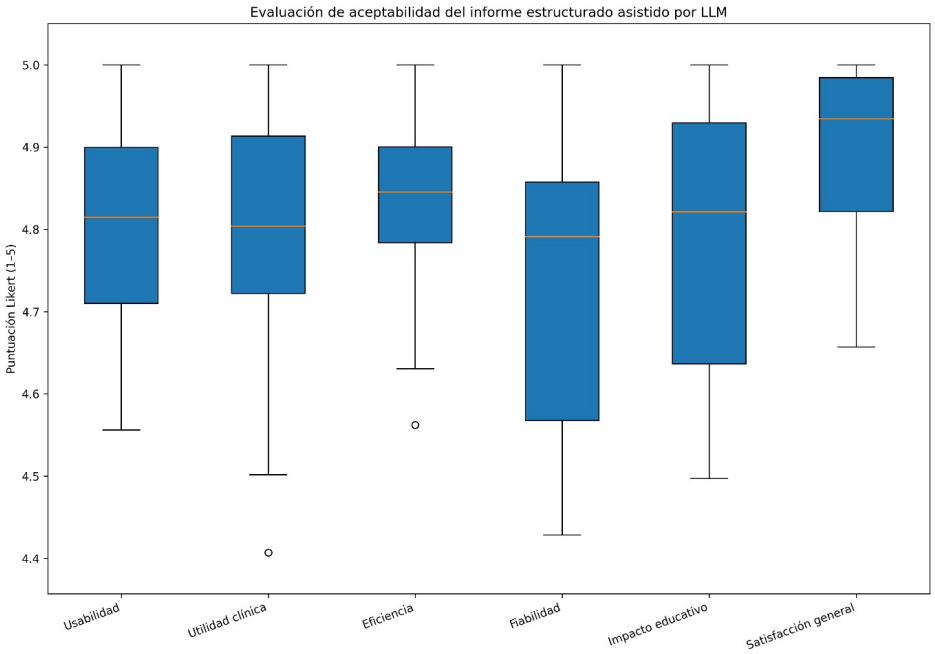

Figura 1. Gráfico de cajas y bigotes con medianas y rangos intercuartílicos Likert por dimensión. Eje Y: puntuación 1-5, con escala ajustada; Eje X: dimensiones evaluadas.

Tabla 4. Métricas del clasificador exploratorio (evaluadores humanos).

Métrica	Valor
Exactitud	0,92
Precisión	0,91
Sensibilidad	0,94
Especificidad	0,91
F1-score	0,92

3.3 Preferencia ciega

Se observó diferencia sistemática según la relación con el autor: los evaluadores seleccionaron con mayor frecuencia informes de colegas (78%) que propios (70%). Este patrón sugiere posibles sesgos de atribución, aunque la interpretación debe ser cautelosa dado el tamaño muestral limitado.

3.4 Perplejidad

Los informes de residentes presentaron perplejidad mayor (mediana 28,5; RIC 25,0-30,5) que los de adjuntos (mediana 26,0; RIC 24,0-29,0; p = 0,03). La figura 2 ilustra esta comparación. Este hallazgo exploratorio sugiere mayor regularidad lingüística en profesionales experimentados, posiblemente reflejando patrones de redacción más estandarizados adquiridos con la experiencia. Es importante reiterar que la perplejidad mide previsibilidad textual para un modelo de lenguaje específico, no calidad clínica del informe.

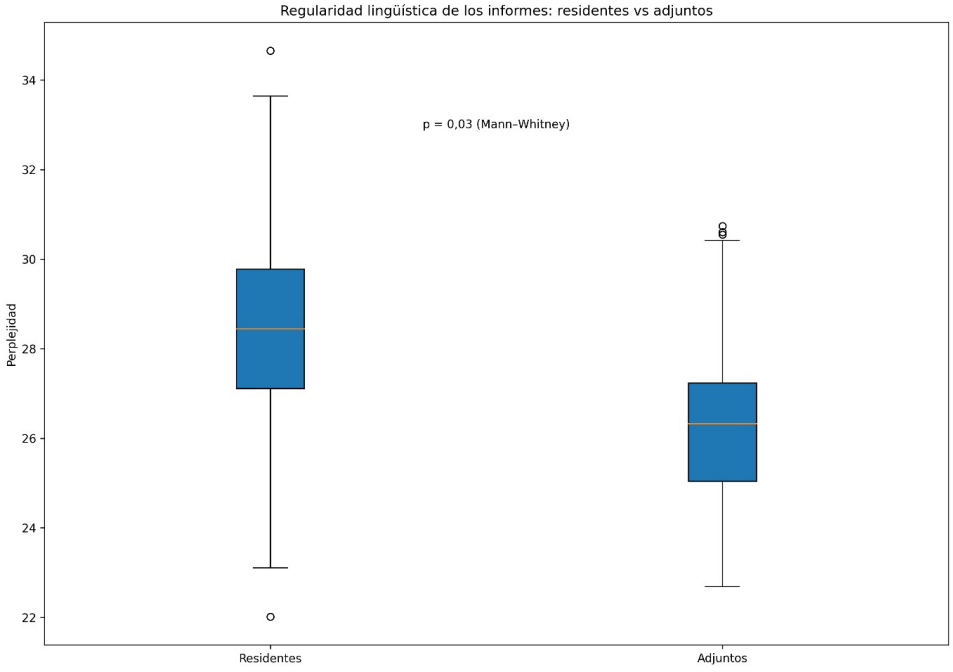

Figura 2. Diagrama de cajas y bigotes comparando perplejidad entre informes de residentes y adjuntos.

4. Discusión

Este estudio piloto explora la implementación de una herramienta de informe estructurado asistido por LLM en urgencias, desarrollada por residentes sin programación avanzada. Los resultados proporcionan señales iniciales favorables sobre factibilidad y aceptación, aunque las limitaciones del diseño impiden establecer relaciones causales o generalizar los hallazgos.

4.1 Interpretación de resultados

Las medianas Likert elevadas (>4,7/5) reflejan buena aceptación inicial. La Satisfacción General alcanzó 4,90, sugiriendo que la experiencia cumplió las expectativas de los participantes. La Usabilidad (4,85) indica que la integración fue percibida como fluida en el flujo de trabajo de guardias. La Fiabilidad, con la mediana más baja (4,75), puede interpretarse favorablemente como reflejo de una actitud crítica apropiada, donde los residentes reconocen que la responsabilidad diagnóstica final recae en el radiólogo y no en la herramienta. Esto subraya la importancia de enmarcar los LLM como herramientas de apoyo (16), no como sistemas autónomos de decisión clínica.

4.2 Conexión entre resultados e impacto formativo

Los resultados obtenidos pueden interpretarse en relación con el impacto educativo propuesto. La alta aceptabilidad (medianas >4,7) sugiere que la herramienta no genera resistencia ni interrumpe el proceso de aprendizaje, condición necesaria para cualquier intervención formativa. La menor perplejidad observada en adjuntos respecto a residentes es coherente con la hipótesis de que la experiencia profesional conduce a patrones comunicativos más estandarizados; si la herramienta facilita la adopción temprana de estos patrones, podría acelerar la curva de aprendizaje, aunque esta hipótesis requiere confirmación longitudinal. La capacidad de los residentes para distinguir informes asistidos (F1 = 0,92) indica que reconocen las características del formato estructurado, paso previo a su internalización. En conjunto, estos indicadores son compatibles con el marco teórico de andamiaje: apoyo externo reconocible que puede retirarse progresivamente a medida que se consolidan las competencias (17).

4.3 Huella lingüística

El F1-score de 0,92 en la tarea de clasificación humana demuestra que existen características formales diferenciables entre informes asistidos y manuales. Sin embargo, conviene interpretar este hallazgo con cautela: indica una diferencia estilística perceptible, pero no implica juicios de valor sobre cuál modalidad produce informes de mayor calidad clínica o comunicativa. La diferencia en perplejidad entre residentes y adjuntos sugiere que la experiencia profesional conduce a mayor estandarización del estilo de redacción, hallazgo coherente con la literatura sobre desarrollo de expertise (cf. 7,20).

4.4 Implicaciones educativas

El valor educativo potencial de esta herramienta no reside en automatizar el informe, sino en entrenar competencias transversales: organización del razonamiento, completitud sistemática, claridad comunicativa. El concepto de andamiaje cognitivo, tomado de la teoría sociocultural del aprendizaje (17), sugiere que estructuras de apoyo temporales pueden facilitar el desarrollo de habilidades que posteriormente se internalizan. En este caso, la plantilla estructurada actúa como un organizador externo que reduce la carga cognitiva extrínseca, permitiendo al residente concentrarse en el contenido diagnóstico.

A modo de ejemplo concreto: el asistente incluye recordatorios sistemáticos para verificar la descripción de hallazgos en todas las regiones anatómicas relevantes, reduciendo omisiones inadvertidas. También sugiere la inclusión de comparaciones con estudios previos cuando están disponibles, un elemento frecuentemente olvidado en informes de urgencias. Estos mecanismos ilustran cómo el andamiaje puede traducirse en mejoras tangibles de completitud, independientemente del diagnóstico.

Este piloto evalúa únicamente la integración operativa; la demostración de aprendizaje genuino requerirá medidas externas y estudios longitudinales. En formación basada en

competencias, la estandarización puede facilitar la evaluación formativa de componentes observables del razonamiento, alineada con EPAs y progresión competencial (6). La literatura de educación médica reconoce el valor de productos observables como el informe estructurado para documentar competencias comunicativas y servir como base para feedback específico (18-19).

En un estudio confirmatorio, el impacto debería medirse con desenlaces observables: completitud del informe (porcentaje de secciones cubiertas), consistencia terminológica (adherencia al léxico estandarizado), tiempo de elaboración, y tasa de discordancias residente-adjunto.

4.5 Contexto en la literatura

Revisiones recientes destacan que, pese al potencial del informe estructurado, su adopción en la práctica clínica sigue siendo limitada. Los LLM se proponen como catalizadores para superar barreras de implementación, aunque persisten retos regulatorios y de validación (21). Una extensión natural de este trabajo sería evaluar discrepancias entre borradores generados y versiones finales validadas, como proponen trabajos que categorizan cambios textuales, semánticos y críticos en informes radiológicos (22).

4.6 Desarrollo descentralizado

Un aspecto distintivo es que la herramienta fue desarrollada por residentes mediante ingeniería de prompts, sin requerir conocimientos de programación. Este modelo de desarrollo descentralizado democratiza el acceso a tecnologías de IA, aunque plantea desafíos de validación y gobernanza que deberán abordarse a medida que proliferen herramientas similares en entornos clínicos.

4.7 Riesgos educativos

Los asistentes generativos pueden producir beneficio operativo inmediato, pero también desplazar el ejercicio cognitivo de describir y redactar. El potencial de eficiencia debe ponderarse con el riesgo de dependencia tecnológica, análogo al descrito para sistemas de navegación GPS (23). En radiología, este riesgo podría afectar competencias nucleares del residente si la herramienta se usa como sustituto del pensamiento diagnóstico en lugar de como andamiaje temporal. Proponemos un uso pedagógicamente seguro con tres salvaguardas: que el residente formule la impresión diagnóstica con sus propias palabras antes de usar la herramienta, que el adjunto ofrezca feedback centrado en estructura y razonamiento, y que se aplique retirada progresiva del apoyo en residentes avanzados.

4.8 Fortalezas y limitaciones

Este estudio aporta una implementación en condiciones reales de urgencias, con alta validez ecológica. El informe se sitúa como evidencia formativa observable, y la IA generativa se enmarca como apoyo no diagnóstico integrado en supervisión habitual. Las limitaciones son sustanciales y deben considerarse al interpretar los resultados. El tamaño muestral (cuatro residentes voluntarios de once posibles, sin grupo control) limita la generalización y la potencia estadística. El uso de una encuesta ad hoc no validada psicométricamente introduce incertidumbre sobre la fiabilidad y validez de las medidas. El carácter perceptivo del "impacto educativo" no permite afirmar aprendizaje real sin evaluación longitudinal con medidas externas. Los riesgos de carry-over inherentes al diseño alternante sin contrabalanceo formal pueden haber introducido sesgos de secuencia. Finalmente, la dependencia del clasificador humano para detectar la huella lingüística introduce subjetividad. Los resultados constituyen evidencia preliminar que justifica, pero no sustituye, estudios multicéntricos controlados con muestras adecuadas e instrumentos validados.

5. Conclusiones

- A la luz de este piloto, los hallazgos sugieren que la integración inicial del informe estructurado asistido por LLM en residentes de radiodiagnóstico es factible y bien aceptada en el contexto de guardias.
- Desde la perspectiva de educación médica, el interés de esta herramienta no radica en automatizar el razonamiento diagnóstico, sino en su potencial como andamiaje que promueve comunicación más estandarizada, reduce variabilidad formal, y puede convertir el informe en evidencia observable para marcos de formación basada en competencias (18, 20, 24).
- Como hallazgo preliminar, estos resultados justifican estudios posteriores con diseños controlados, muestras adecuadas y medidas externas centradas en desempeño comunicativo y calidad objetiva del informe.

Financiación: No ha habido financiación.

Declaración de conflicto de interés: Los autores declaran no tener ningún conflicto de interés.

Contribuciones de los autores: Todos los autores contribuyeron a la conceptualización, metodología, adquisición de datos, análisis formal, investigación, redacción del borrador original, revisión y edición del manuscrito, y aprobación de la versión final.

6. Referencias

1. Kahn CE Jr. Artificial intelligence in radiology: decision support systems. *Radiographics*. **1994**, 14, 849-861. https://doi.org/10.1148/radiographics.14.4.7938772
2. Rajpurkar P, Lungren MP. The current and future state of AI interpretation of medical images. *N Engl J Med*. **2023**, 388, 1981-1990. https://doi.org/10.1056/NEJMra2301725
3. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*. **2020**, 33, 1877-1901. https://doi.org/10.48550/arXiv.2005.14165
4. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Advances in Neural Information Processing Systems*. **2017**, 30, 5998-6008. https://doi.org/10.48550/arXiv.1706.03762
5. Hartung MP, Bickle IC, Gaillard F, Kanne JP. How to Create a Great Radiology Report. *RadioGraphics*. **2020**, 40, 1658-1670. https://doi.org/10.1148/rg.2020200020
6. Castro D, Mishra S, Kwan BY, et al. Structured Reporting in Radiology Residency: A Standardized Approach to Assessing Interpretation Skills and Competence. *Int Med Educ*. **2025**, 4, 40. https://doi.org/10.3390/ime4010002
7. Larson DB, Towbin AJ, Pryor RM, Donnelly LF. Improving consistency in radiology reporting through the use of department-wide standardized structured reporting. *Radiology*. **2013**, 267, 240-250. https://doi.org/10.1148/radiol.12121502
8. Kao JP, Kao HT. Large Language Models in radiology: A technical and clinical perspective. *Eur J Radiol Artif Intell*. **2025**, 2, 100021. https://doi.org/10.1016/j.ejrai.2025.100021
9. Mongan J, Moy L, Kahn CE Jr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. *Radiol Artif Intell*. **2020**, 2, e200029. https://doi.org/10.1148/ryai.2020200029
10. Wirth S, Hebebrand J, Basilico R, et al. European Society of Emergency Radiology (ESER). Guideline on radiological polytrauma imaging and service (full version). Disponible en: https://www.eser-society.org/polytrauma-imaging-guidelines/ (Acceso: enero 2025).
11. Radiological Society of North America. RadReport Templates. Disponible en: https://www.rsna.org/practice-tools/data-tools-and-standards/radreport-templates (Acceso: enero 2025).
12. Sociedad Española de Radiología Médica. *Léxico conflictivo en Radiología*. Madrid: SERAM; 2020. Disponible en: https://static.seram.es/wp-content/uploads/2021/07/lexico_radiologico_conflictivo.pdf (Acceso: enero 2025).
13. Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*. **2020**, 33, 9459-9474. https://doi.org/10.48550/arXiv.2005.11401
14. Lancaster GA, Dodd S, Williamson PR. Design and analysis of pilot studies: recommendations for good practice. *J Eval Clin Pract*. **2004**, 10, 307-312. https://doi.org/10.1111/j..2002.384.doc.x

15. Brooke J. SUS: A 'quick and dirty' usability scale. En: Jordan PW, Thomas B, McClelland IL, Weerdmeester B, eds. *Usability Evaluation in Industry*. London: Taylor & Francis; **1996**. p. 189-194.
16. Patel BN, Rosenberg L, Willcox G, et al. Human-machine partnership with artificial intelligence for chest radiograph diagnosis. *NPJ Digit Med*. **2019**, 2, 111. https://doi.org/10.1038/s41746-019-0189-7
17. Wood D, Bruner JS, Ross G. The role of tutoring in problem solving. *J Child Psychol Psychiatry*. **1976**, 17, 89-100. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x
18. Ten Cate O. Entrustability of professional activities and competency-based training. *Med Educ*. **2005**, 39, 1176-1177. https://doi.org/10.1111/j.1365-2929.2005.02341.x
19. Epstein RM. Assessment in medical education. *N Engl J Med*. **2007**, 356, 387-396. https://doi.org/10.1056/NEJMra054784
20. Schwartz LH, Panicek DM, Berk AR, et al. Improving communication of diagnostic radiology findings through structured reporting. *Radiology*. **2011**, 260, 174-181. https://doi.org/10.1148/radiol.11101913
21. Busch F, Hoffmann L, Pinto dos Santos D, et al. Large language models for structured reporting in radiology: past, present, and future. *Eur Radiol*. **2025**, 35, 2589-2602. https://doi.org/10.1007/s00330-024-11107-6
22. Lindholz M, Burdenski A, Ruppel R, et al. Comparing large language models and text embedding models for automated classification of textual, semantic, and critical changes in radiology reports. *Eur J Radiol*. **2025**, 191, 112316. https://doi.org/10.1016/j.ejrad.2025.112316
23. Martín-Noguerol T, López-Úbeda P, Luna A. From GPS to ChatGPT in Radiology... Dumb and Dumber? *J Am Coll Radiol*. **2025**. https://doi.org/10.1016/j.jacr.2025.09.014
24. European Society of Radiology. ESR paper on structured reporting in radiology. *Insights Imaging*. **2018**, 9, 1-7. https://doi.org/10.1007/s13244-017-0588-8