

REVISIÓN NARRATIVA

Evaluation of Artificial Intelligence and Large Language Models (LLMs) in Clinical Practice: Precision, Safety, and Applicability in Middle-Income Countries

Evaluación de inteligencia artificial y modelos de lenguaje grande (LLM) en la práctica clínica: precisión, seguridad y aplicabilidad en países de ingresos medios

Avaliação da inteligência artificial e dos modelos de linguagem de grande escala (LLM) na prática clínica: precisão, segurança e aplicabilidade em países de renda média

Byron Hernán Mena Acosta¹  

¹Médico general e ingeniero en redes de información, práctica privada, Ambato, Ecuador.

Citación sugerida: Mena Acosta BH. Evaluación de inteligencia artificial y modelos de lenguaje grande (LLM) en la práctica clínica: precisión, seguridad y aplicabilidad en países de ingresos medios. Rev Méd Loja. 2026;1(2):57-66. doi:[10.67370/e53pcq28](https://doi.org/10.67370/e53pcq28)

Recibido: 25-07-2026

Aceptado: 13-10-2026

Publicado: 22-10-2026

Editor: Dr. Miguel David Alvarez 

Autor de correspondencia: Byron Hernán Mena Acosta 

ABSTRACT

Introduction: Large language models (LLMs) are being investigated to support diagnosis, communication, and clinical decision-making. Performance depends on the task, model, and evaluation setting, limiting extrapolation to middle-income countries. **Objective:** To analyze evidence on the performance and clinical safety of LLMs and other artificial intelligence systems and discuss their applicability in middle-income countries, with an emphasis on Ecuador. **Methods:** A narrative review with a structured search of PubMed and Google Scholar conducted in May 2026, targeting publications from 2022 through that date. One author selected sources and extracted data. Twelve publications addressing performance or implementation were synthesized, distinguishing LLM evaluations, other AI systems, and documents without comparative clinical results. No meta-analysis or formal risk-of-bias assessment was performed. **Results:** In obstetrics and gynecology scenarios, the evaluated LLMs as a group achieved an overall accuracy of 73.75%, compared with 65.35% among residents. Conversely, an evaluation of abdominal pain cases found performance inferior to physicians. Ratings of potential harm in Med-PaLM 2 answers did not establish equivalent safety in patient care. Studies of communication, prognosis, and automated decision support measured different outcomes and could not support a pooled estimate. Connectivity, interoperability, local validation, and costs influence applicability in Ecuador. **Conclusions:** LLMs show potential as supervised support for specific tasks, but the reviewed evidence does not establish general clinical superiority or safety equivalent to specialists. Implementation requires prospective local evaluation, data protection, training, and performance monitoring.

Keywords: Artificial Intelligence; Clinical Decision-Making; Patient Safety; Language Models; Developing Countries.

RESUMEN

Introducción: Los modelos de lenguaje grande (LLM) se investigan como apoyo al diagnóstico, la comunicación y la toma de decisiones clínicas. Su rendimiento depende de la tarea, el modelo y el entorno de evaluación, lo que limita la extrapolación a países de ingresos medios. **Objetivo:** Analizar la evidencia sobre el rendimiento y la seguridad clínica de los LLM y otros sistemas de inteligencia artificial, y discutir su aplicabilidad en países de ingresos medios, con énfasis en Ecuador. **Métodos:** Revisión narrativa con búsqueda estructurada en PubMed y Google Scholar realizada en mayo de 2026, orientada a publicaciones de 2022 hasta esa fecha. Un autor seleccionó las fuentes y extrajo los datos. Se sintetizaron 12 publicaciones sobre rendimiento o implementación, diferenciando evaluaciones de LLM, otros sistemas de IA y documentos sin resultados clínicos comparativos. No se efectuaron metaanálisis ni evaluación formal del riesgo de sesgo. **Resultados:** En escenarios de obstetricia y ginecología, el conjunto de LLM evaluados alcanzó una exactitud global del 73,75 %, frente al 65,35 % de residentes. En contraste, una evaluación de dolor abdominal encontró un rendimiento inferior al de médicos. Las valoraciones de daño potencial de respuestas de Med-PaLM 2 no demostraron seguridad equivalente en la atención de pacientes. Los estudios de comunicación, pronóstico y soporte automatizado midieron desenlaces distintos y no permiten una estimación conjunta. La conectividad, la interoperabilidad, la validación local y los costos condicionan la aplicabilidad en Ecuador. **Conclusiones:** Los LLM muestran potencial como apoyo supervisado en tareas específicas, pero la evidencia revisada no establece superioridad clínica general ni seguridad equivalente a especialistas. Su implementación requiere evaluación prospectiva local, protección de datos, capacitación y vigilancia del desempeño.

Palabras clave: Inteligencia Artificial; Toma de Decisiones Clínicas; Seguridad del Paciente; Modelos de Lenguaje; Países en Desarrollo.

RESUMO

Introdução: Os modelos de linguagem de grande escala (LLM) vêm sendo investigados como apoio ao diagnóstico, à comunicação e à tomada de decisões clínicas. Seu desempenho depende da tarefa, do modelo e do ambiente de avaliação, o que limita a extrapolação para países de renda média. **Objetivo:** Analisar as evidências sobre o desempenho e a segurança clínica dos LLM e de outros sistemas de inteligência artificial e discutir sua aplicabilidade em países de renda média, com ênfase no Equador. **Métodos:** Revisão narrativa com busca estruturada no PubMed e no Google Scholar realizada em maio de 2026, orientada a publicações de 2022 até essa data. Um autor selecionou as fontes e extraiu os dados. Foram sintetizadas 12 publicações sobre desempenho ou implementação, diferenciando avaliações de LLM, outros sistemas de IA e documentos sem resultados clínicos comparativos. Não foram realizadas metanálises nem avaliação formal do risco de viés. **Resultados:** Em cenários de obstetrícia e ginecologia, o conjunto de LLM avaliados alcançou acurácia global de 73,75%, frente a 65,35% dos residentes. Em contraste, uma avaliação de dor abdominal encontrou desempenho inferior ao dos médicos. As avaliações de dano potencial das respostas do Med-PaLM 2 não demonstraram segurança equivalente no atendimento a pacientes. Os estudos de comunicação, prognóstico e suporte automatizado mediram desfechos distintos e não permitem uma estimativa conjunta. A conectividade, a interoperabilidade, a validação local e os custos condicionam a aplicabilidade no Equador. **Conclusões:** Os LLM mostram potencial como apoio supervisionado em tarefas específicas, mas as evidências revisadas não estabelecem superioridade clínica geral nem segurança equivalente à de especialistas. Sua implementação requer avaliação prospectiva local, proteção de dados, capacitação e monitoramento do desempenho.

Palavras-chave: Inteligência Artificial; Tomada de Decisão Clínica; Segurança do Paciente; Modelos de Linguagem; Países em Desenvolvimento.

INTRODUCCIÓN

Las enfermedades no transmisibles representan una carga importante para los sistemas sanitarios. La Organización Mundial de la Salud estima que el 73 % de las muertes por estas enfermedades ocurre en países de ingresos bajos y medios.⁽¹⁾ A esta carga se suman problemas de distribución y disponibilidad del personal sanitario que dificultan el acceso oportuno a la atención.⁽²⁾ En este contexto, las herramientas de inteligencia artificial (IA) han generado interés como apoyo al trabajo clínico y a la comunicación con los pacientes.

Los modelos de lenguaje grande (LLM, por sus siglas en inglés) constituyen una familia de sistemas de IA capaces de generar y procesar texto. Sus aplicaciones médicas abarcan la

elaboración de diagnósticos diferenciales, la respuesta a preguntas clínicas y la preparación de información educativa.^(3,4) Sin embargo, un resultado favorable en preguntas de examen o casos escritos no demuestra por sí mismo utilidad clínica en pacientes. La integración en los servicios de salud también exige considerar exactitud, sesgo, interpretabilidad y condiciones de uso.⁽⁵⁾

La representación desigual de poblaciones y contextos en los datos y evaluaciones puede limitar la pertinencia de las respuestas fuera de los entornos estudiados. La evidencia sobre sesgos en la representación de actitudes de salud pública refuerza la necesidad de examinar la validez contextual de los modelos.⁽⁶⁾ En Ecuador, el aumento del acceso a internet no ha eliminado las brechas educativas, territoriales y de competencias digitales descritas en informes nacionales.⁽⁷⁾ Asimismo, la Política Nacional de Transformación Digital del Sector Salud 2024–2034 identifica necesidades de modernización e interoperabilidad que condicionan la incorporación de tecnologías emergentes.⁽⁸⁾

Esta revisión narrativa analiza el rendimiento y la seguridad clínica de los LLM en distintas tareas médicas, incorpora ejemplos diferenciados de otros sistemas de IA y discute las condiciones necesarias para su implementación en países de ingresos medios, con énfasis en Ecuador. El análisis del contexto andino se plantea como una interpretación de aplicabilidad y no como una validación regional de los modelos.

METODOLOGÍA

Se realizó una revisión narrativa con búsqueda estructurada y síntesis temática. La búsqueda se efectuó durante mayo de 2026 en PubMed y Google Scholar y se orientó a publicaciones desde 2022 hasta la fecha de búsqueda. En PubMed se combinaron los términos “Large Language Models” OR “ChatGPT” con “Clinical Decision-Making” OR “Patient Safety”, utilizando los campos de título y resumen y el filtro de texto completo gratuito. Se examinaron los primeros 200 resultados ordenados mediante Best Match. De manera complementaria se revisaron los primeros 100 resultados de Google Scholar con los mismos conceptos y ordenamiento por relevancia.

Se seleccionaron publicaciones sobre exactitud diagnóstica, pronóstico, comunicación con pacientes, seguridad de las respuestas o apoyo a las decisiones clínicas. Se priorizaron evaluaciones empíricas con comparadores humanos, modelos alternativos, escalas clínicas o atención convencional. También se incorporaron, en una categoría contextual distinta, un protocolo sobre educación y autocuidado en diabetes y un estudio de caso de un asistente oncológico. Los documentos institucionales, un editorial y otras fuentes de contexto se utilizaron para la introducción y la discusión, sin considerarlos estudios de rendimiento clínico.

Un único autor realizó la selección y la extracción de datos mediante revisión de títulos, resúmenes y textos completos potencialmente pertinentes. Se organizaron los datos por publicación, tarea clínica, diseño, población o unidad de análisis, sistema evaluado, comparador y hallazgos. El corpus de síntesis comprendió 12 publicaciones: ocho evaluaciones de LLM, dos estudios de otros sistemas de IA clínica y dos documentos sobre implementación. Estos últimos no se utilizaron para estimar eficacia o seguridad clínica.

La síntesis fue narrativa y no se calcularon estimaciones combinadas debido a la heterogeneidad de tareas, diseños, comparadores y desenlaces. No se registró un protocolo prospectivo ni se aplicaron herramientas formales de evaluación del riesgo de sesgo o de certeza de la evidencia. El límite de resultados revisados, el filtro de acceso gratuito y la selección por un solo autor restringen la cobertura y reproducibilidad de la búsqueda. En consecuencia, la revisión no se presenta como una revisión sistemática exhaustiva ni como un metaanálisis. La Figura 1 resume los recuentos del

proceso de búsqueda y selección mediante una adaptación del esquema de flujo PRISMA 2020,⁽⁹⁾ utilizada únicamente para organizar su presentación.

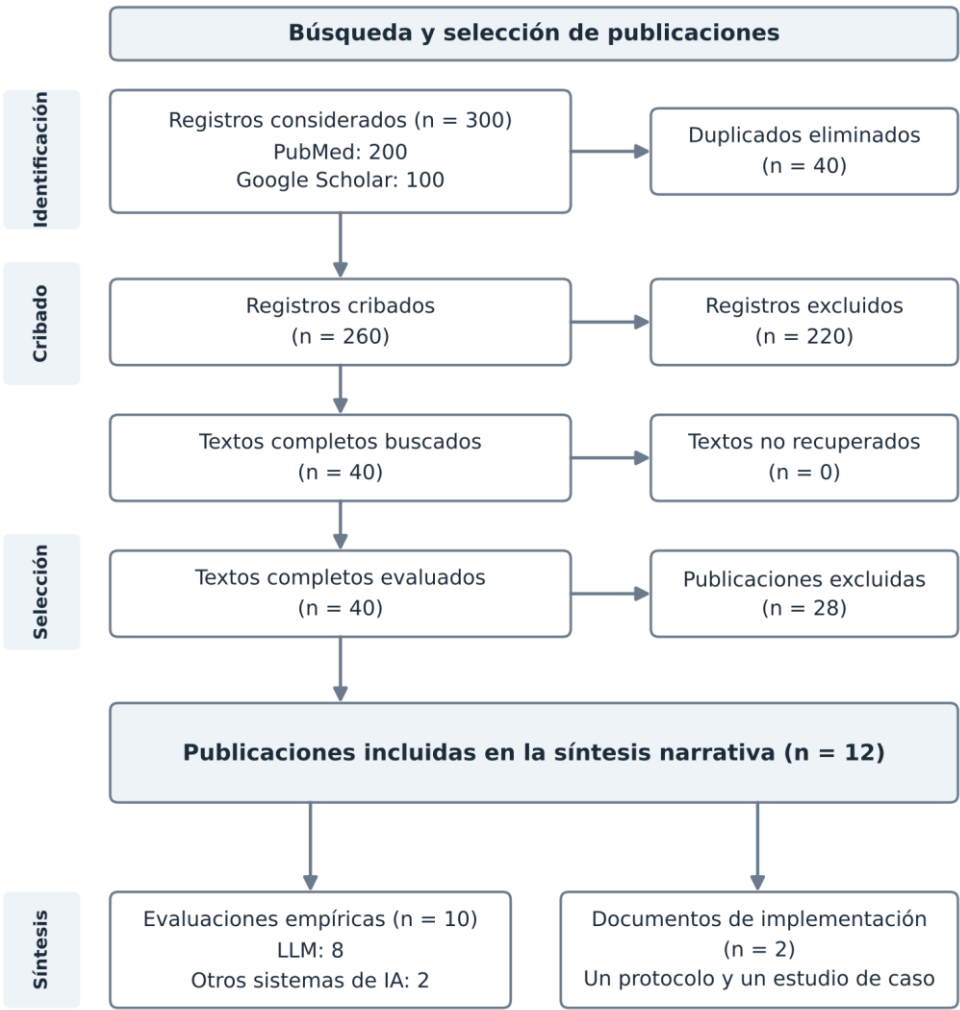


Figura 1. Diagrama de búsqueda y selección de publicaciones. Estructura adaptada de PRISMA 2020,⁽⁹⁾ con los recuentos de la búsqueda original. Los 300 registros corresponden a los resultados considerados dentro de los límites de búsqueda, no al total de coincidencias de las plataformas. Su presentación no implica exhaustividad ni cumplimiento de los requisitos de una revisión sistemática.

RESULTADOS DE LA SÍNTESIS NARRATIVA

Los trabajos revisados abordaron tareas diferentes: generación de diagnósticos diferenciales, resolución de escenarios clínicos, comunicación escrita, predicción de desenlaces, valoración del daño potencial de respuestas, interpretación de imágenes y prevención de eventos hospitalarios. Sus resultados se describen por tipo de tarea. La Tabla 1 reúne las 12 publicaciones y distingue las evaluaciones de rendimiento de los documentos de implementación.

Diagnóstico y resolución de escenarios clínicos

Ríos-Hoyo et al. evaluaron GPT-3.5 y GPT-4 en 75 casos clínicos complejos. GPT-4 incluyó el diagnóstico correcto en algún punto del diagnóstico diferencial en el 68 % de los casos, frente al 48 % con GPT-3.5 ($p=0,0063$). Sin embargo, el diagnóstico colocado en primer lugar fue correcto con mucha menor frecuencia. Este resultado apoya una posible utilidad para ampliar el diagnóstico diferencial, pero no demuestra que el modelo sustituya la priorización diagnóstica del médico.⁽⁴⁾

En el estudio PERFORM, desarrollado con escenarios de obstetricia y ginecología y residentes de distintos años de formación, los ocho LLM evaluados alcanzaron en conjunto una exactitud del 73,75 % (IC del 95 %: 69,64–77,49), frente al 65,35 % (IC del 95 %: 62,85–67,76) de los residentes ($p<0,001$); los modelos de alto rendimiento (o1-preview, GPT-4o y Claude Sonnet 3.5) alcanzaron una exactitud del 88,33 % en ambos idiomas evaluados. Se exploraron diferencias por idioma y restricción de tiempo. La comparación se realizó con residentes y en tareas estandarizadas; por ello, no puede interpretarse como superioridad general frente a especialistas en atención clínica real.⁽¹⁰⁾

Los hallazgos no fueron uniformemente favorables. Hager et al. estudiaron LLM abiertos en un marco basado en 2400 casos de cuatro causas de dolor abdominal. Los modelos rindieron peor que los médicos y mostraron dificultades para seguir guías, interpretar resultados e integrar información clínica.⁽¹¹⁾ Estos resultados contrastan con el desempeño en pruebas de conocimiento y subrayan la importancia de evaluar procesos clínicos completos.

Comunicación con pacientes y recomendaciones preventivas

Rotem et al. compararon respuestas de GPT-4o y de un uroginecólogo, valoradas por 203 pacientes. Las puntuaciones globales fueron mediana 76 (rango intercuartílico: 67–85) para ChatGPT y 72 (rango intercuartílico: 63–80) para el consultante ($p<0,01$). La escala evaluó comprensibilidad, ayuda percibida y tranquilidad, con un máximo de 15 puntos por respuesta y 90 para las seis preguntas. Estas puntuaciones reflejan percepción de la comunicación y no constituyen una medida de exactitud diagnóstica ni de desenlaces clínicos.⁽¹²⁾

En cardiología preventiva, Sarraju et al. formularon 25 preguntas a ChatGPT y sometieron las respuestas a evaluación por tres expertos. Las respuestas a 21 preguntas (84 %) se consideraron apropiadas en los dos contextos valorados: información educativa y borradores de mensajes para pacientes. El estudio se realizó en diciembre de 2022 y se publicó en 2023. Su diseño evaluó adecuación de respuestas, sin medir reducción de eventos cardiovasculares ni superioridad frente al manejo médico.⁽¹³⁾

Predicción de desenlaces neurológicos

Pedro et al. analizaron retrospectivamente 163 pacientes con accidente cerebrovascular isquémico tratados mediante tromboectomía. La concordancia entre las predicciones de ChatGPT y el resultado funcional a tres meses fue $\kappa=0,354$ (IC del 95 %: 0,260–0,448) para la puntuación exacta de la escala de Rankin modificada y $\kappa=0,727$ (IC del 95 %: 0,620–0,833) para su clasificación dicotómica. El artículo describió estos niveles como fair y good, respectivamente, y encontró un mejor desempeño que MT-DRAGON. Se trata de predicción pronóstica, no de diagnóstico ni de comparación directa con especialistas.⁽¹⁴⁾

Evaluación de respuestas y daño potencial

Singhal et al. comunicaron un rendimiento de hasta el 86,5 % de Med-PaLM 2 en MedQA, un conjunto de preguntas de estilo USMLE. En una evaluación de respuestas extensas, la proporción calificada en la categoría de ausencia de daño potencial fue 93,3 % para Med-PaLM 2 y 92,9 % para respuestas médicas ($p=0,950$). Estos porcentajes representan juicios sobre respuestas y no tasas observadas de eventos adversos en pacientes. La ausencia de una diferencia estadísticamente significativa tampoco establece equivalencia o no inferioridad clínica. Las comparaciones con especialistas variaron según la tarea y el entorno de evaluación.⁽¹⁵⁾

Comparación entre modelos abiertos y propietarios

Sandmann et al. compararon sistemas de lenguaje en 125 casos de cinco especialidades. No observaron diferencias estadísticamente significativas entre GPT-4o y DeepSeek-R1 en diagnóstico (p=0,3085) ni en recomendaciones terapéuticas (p=0,1522). Los tamaños del efecto mediante correlación biserial por rangos fueron 0,27 y 0,26, respectivamente. Estos valores describen diferencias en las puntuaciones de desempeño; no representan concordancia entre modelos ni prueban equivalencia formal. Los resultados sugieren que ciertos modelos abiertos pueden ser alternativas para tareas concretas, sujetos a validación e infraestructura apropiadas.⁽¹⁶⁾

Otros sistemas de inteligencia artificial clínica

Chen et al. desarrollaron un modelo de redes neuronales convolucionales 3D para tomografías de hueso temporal. Los conjuntos interno y externo incluyeron 1661 y 108 pacientes; las AUROC fueron 0,96 y 0,93 para detectar oídos patológicos, y 0,85 y 0,83 para identificar colesteatoma. En 121 pacientes prospectivos, la exactitud diagnóstica preoperatoria fue del 81,8 % y el sistema contribuyó a las decisiones clínicas en el 90,1 % de los casos. Este último porcentaje no debe confundirse con exactitud diagnóstica. El sistema es IA de imagen, no un LLM.⁽¹⁷⁾

Huang et al. evaluaron un sistema automatizado de apoyo clínico para prevenir tromboembolismo venoso hospitalario en 19 785 pacientes. Los eventos ocurrieron en 27 de 9895 pacientes del grupo de intervención y en 49 de 9890 controles (OR: 0,55; IC del 95 %: 0,34–0,88), con una reducción relativa comunicada del 46 %. Este desenlace clínico pertenece a un sistema específico de prevención y no puede atribuirse a los LLM ni extrapolarse a otras tareas de IA.⁽¹⁸⁾

Documentos sobre implementación

El protocolo AI-HEALS de Wu et al. describe una intervención mediante WeChat para educación y autocuidado en diabetes tipo 2, con HbA1c y otros indicadores previstos como desenlaces. Al ser un protocolo, no aporta resultados de eficacia o seguridad.⁽¹⁹⁾ Por otra parte, de Marques et al. describen Cuida-AI como un asistente de IA generativa desarrollado para información oncológica en hospitales públicos de Recife, Brasil. Constituye una experiencia contextual de desarrollo e implementación, sin aportar en esta síntesis una estimación comparativa de beneficio clínico frente a médicos.⁽²⁰⁾

Tabla 1. Características de las publicaciones revisadas

Publicación	Diseño y unidad de análisis	Sistema y categoría	Comparador	Hallazgos e interpretación
Ríos-Hoyo et al. ⁽⁴⁾ 2024	Evaluación de 75 casos complejos	GPT-3.5 y GPT-4 LLM	Modelos y diagnósticos de referencia de casos clínicos	GPT-4 incluyó el diagnóstico correcto en el diferencial en 68 % vs. 48 % con GPT-3.5; p=0,0063. La inclusión en la lista no equivale a un diagnóstico principal correcto.
Martinelli et al. ⁽¹⁰⁾ 2025	Escenarios clínicos de obstetricia y ginecología	Ocho LLM, incluidos o1-preview, GPT-4o y Claude Sonnet 3.5 LLM	Residentes de ginecología y obstetricia	Exactitud global de los LLM 73,75 % (IC95 %: 69,64–77,49) vs. 65,35 % (IC95 %: 62,85–67,76); p<0,001. Modelos de alto rendimiento: 88,33 %. Evaluación estandarizada con exploración de idioma y tiempo.
Rotem et al. ⁽¹²⁾ 2025 en línea	Encuesta transversal 203 pacientes; seis preguntas	GPT-4o LLM	Un uroginecólogo	Puntuación global mediana 76 (RIC: 67–85) vs. 72 (RIC: 63–80); p<0,01. Percepción comunicativa, sin demostrar superioridad diagnóstica.
Pedro et al. ⁽¹⁴⁾ 2025	Estudio retrospectivo 163 pacientes con ACV y trombectomía	ChatGPT LLM	mRS observado y escala MT-DRAGON	Concordancia pronóstica κ=0,354 para puntuación exacta y κ=0,727 para resultado dicotómico. No corresponde a exactitud diagnóstica.
Sarraj et al. ⁽¹³⁾ 2023	Evaluación de 25 preguntas preventivas	ChatGPT LLM	Valoración por tres expertos en cardiología	21/25 respuestas (84 %) apropiadas en ambos contextos. Estudio de adecuación de contenido, sin desenlaces cardiovasculares.

Publicación	Diseño y unidad de análisis	Sistema y categoría	Comparador	Hallazgos e interpretación
Chen et al. ⁽¹⁷⁾ 2024	Validación interna: 1661 pacientes; externa: 108; cohorte prospectiva: 121	Red convolucional 3D para TC temporal IA de imagen	Modelo 2D y siete clínicos	AUROC interna/externa: 0,96/0,93 para oído patológico y 0,85/0,83 para colesteatoma. Exactitud prospectiva 81,8 %; contribución a decisiones en 90,1 %. No es un LLM.
Wu et al. ⁽¹⁹⁾ 2023	Protocolo de métodos mixtos Diabetes tipo 2	AI-HEALS mediante WeChat Documento de implementación	Atención habitual prevista	Incluye HbA1c y autocuidado como desenlaces previstos. El protocolo no contiene resultados de eficacia o seguridad.
Hager et al. ⁽¹¹⁾ 2024	Marco de evaluación con 2400 casos de cuatro patologías abdominales	LLM abiertos, incluidos Llama 2, OASST, WizardLM, Clinical Camel y Meditron	Médicos y recomendaciones de guías	Rendimiento inferior al médico y dificultades de integración de información y seguimiento de guías. Evidencia relevante sobre limitaciones clínicas.
Huang et al. ⁽¹⁸⁾ 2023	Estudio prospectivo aleatorizado 19 785 pacientes hospitalizados	Sistema de apoyo clínico para prevención de TEV IA distinta de LLM	Atención convencional	TEV: 27/9895 vs. 49/9890; OR 0,55 (IC95 %: 0,34–0,88). No extrapolar el beneficio del sistema de prevención a LLM.
Singhal et al. ⁽¹⁵⁾ 2025	Evaluación de preguntas médicas y respuestas extensas	Med-PaLM 2 y Med-PaLM LLM	Respuestas médicas y otros modelos	MedQA: hasta 86,5 %. Respuestas calificadas sin daño potencial: 93,3 % vs. 92,9 % médicas; p=0,950. No demuestra equivalencia clínica.
Sandmann et al. ⁽¹⁶⁾ 2025	Benchmark de 125 casos en cinco especialidades	DeepSeek-R1/V3, GPT-4o y otros LLM	Comparación entre sistemas	GPT-4o vs. DeepSeek-R1: diagnóstico p=0,3085 y tratamiento p=0,1522. Sin diferencias significativas; no constituye prueba formal de equivalencia.
de Marques et al. ⁽²⁰⁾ 2026	Estudio de caso en hospitales públicos de Recife	Cuida-AI Asistente de IA generativa	Sin comparación clínica directa utilizada en esta síntesis	Desarrollo de un asistente de información oncológica. Se considera como experiencia de implementación latinoamericana, sin estimar eficacia clínica comparativa.

LLM: modelo de lenguaje grande; IC95 %: intervalo de confianza del 95 %; RIC: rango intercuartílico; ACV: accidente cerebrovascular; mRS: escala de Rankin modificada; TC: tomografía computarizada; AUROC: área bajo la curva ROC; TEV: tromboembolismo venoso; OR: odds ratio. Las unidades de análisis y los desenlaces son heterogéneos; no se estimó un efecto conjunto.

DISCUSIÓN

La evidencia revisada muestra un desempeño variable según la tarea y el contexto. Los resultados favorables en escenarios estandarizados y comunicación escrita coexisten con limitaciones en procesos de decisión que requieren recopilar datos, integrar hallazgos y seguir recomendaciones clínicas. Por ello, no es adecuado resumir el conjunto como una demostración de superioridad de los LLM frente a especialistas. La distinción entre residentes, especialistas, escalas pronósticas y sistemas alternativos es esencial para interpretar cada comparación.^(4,10-16)

La exactitud diagnóstica, la concordancia pronóstica, la satisfacción comunicativa y la seguridad son dimensiones relacionadas, pero no intercambiables. Una respuesta clara puede ser clínicamente incompleta, y un diagnóstico correcto incluido en una lista extensa puede no orientar adecuadamente la conducta. Del mismo modo, un modelo que responde preguntas médicas con un alto puntaje no necesariamente identifica cuándo faltan datos, cuándo debe abstenerse o cuándo se requiere atención urgente. La selección de desenlaces debe corresponder a la función clínica prevista.^(4,11,12,15)

La evaluación de seguridad requiere especial cautela. Las rúbricas de daño potencial aportan información sobre el contenido de las respuestas, pero no sustituyen el seguimiento de pacientes expuestos a recomendaciones. Para sostener equivalencia o no inferioridad serían necesarios diseños específicos, márgenes predefinidos y estimaciones suficientemente precisas. El uso de valores p superiores a 0,05 en comparaciones de superioridad no permite concluir igualdad de desempeño. También debe diferenciarse el beneficio de sistemas especializados, como el apoyo automatizado para prevenir tromboembolismo, de la seguridad de los modelos generativos de propósito general.^(15,16,18)

Aplicabilidad en Ecuador y otros países de ingresos medios

La extrapolación a Ecuador depende tanto de los servicios disponibles como de las capacidades del modelo. Los escenarios estudiados no representan necesariamente la prevalencia de enfermedades, las formas de consulta, el lenguaje clínico, los niveles de alfabetización y los recursos diagnósticos o terapéuticos del primer nivel de atención. En esta revisión, la aplicabilidad regional se deriva de una interpretación contextual; no se identificó dentro del corpus una validación clínica ecuatoriana de los modelos descritos. La evidencia sobre representación desigual y sesgos refuerza la necesidad de contrastar las respuestas en poblaciones locales.^(5,6)

El boletín de la Escuela Politécnica Nacional señala que el acceso a internet de los hogares aumentó del 28 % en 2013 al 71,3 % en 2025, aunque persisten diferencias territoriales y de competencias digitales.⁽⁷⁾ Por su parte, la política nacional de transformación digital recoge un diagnóstico técnico de 2024 según el cual el 95,55 % de los establecimientos de primer nivel tenía conectividad funcional, junto con obsolescencia e insuficiente interoperabilidad de sistemas.⁽⁸⁾ Estas cifras proceden de años y unidades de análisis diferentes y no miden preparación institucional para implementar LLM.

La disponibilidad de modelos abiertos puede ampliar las opciones de adaptación y despliegue, pero no elimina los costos de infraestructura, mantenimiento, seguridad, integración y capacitación. El desempeño comparable observado en determinadas tareas debe contrastarse con la capacidad técnica de cada institución y con la sostenibilidad del servicio.⁽¹⁶⁾ Además, un análisis de la sociedad civil sobre el sistema público ecuatoriano describe problemas de disponibilidad de insumos y medicamentos; se utiliza aquí como contexto y no como medición de capacidad para adoptar IA.⁽²¹⁾

Para una implementación gradual, se propone comenzar con funciones de apoyo delimitadas y supervisadas, como la preparación de información educativa o borradores que el profesional revise antes de utilizarlos. La evaluación local debería incluir exactitud y pertinencia de las respuestas, reconocimiento de signos de alarma, recomendaciones incompatibles con los recursos disponibles, omisiones y desempeño por subgrupos. Antes de ampliar el uso, conviene comparar el trabajo clínico habitual con el asistido y medir efectos en decisiones, tiempos, errores y resultados relevantes para los pacientes.

La gobernanza debe definir quién revisa las respuestas, cómo se documentan las decisiones y cómo se gestionan incidentes o cambios de versión. También requiere proteger la información clínica y evaluar las condiciones de tratamiento de datos de cada sistema. Estas son condiciones de implementación propuestas a partir de la síntesis; no representan beneficios demostrados por los estudios revisados. Experiencias regionales como Cuida-AI muestran líneas de desarrollo que requieren evaluación clínica adicional antes de extrapolar su utilidad a otros servicios.^(5,8,20)

Limitaciones de la revisión

La búsqueda se restringió a dos fuentes, a un número limitado de resultados ordenados por relevancia y, en PubMed, a texto completo gratuito. Un único autor realizó la selección y extracción; no hubo una segunda evaluación independiente. La falta de fechas diarias y estrategias completas específicas para cada plataforma limita la reproducibilidad. Tampoco se registró un protocolo ni se evaluaron formalmente el riesgo de sesgo o la certeza de la evidencia. Estas decisiones pueden introducir sesgos de selección y de interpretación y hacen que la síntesis no represente toda la literatura disponible.

El corpus reunió diseños y unidades de análisis diferentes, incluidos escenarios escritos, preguntas, pacientes, un protocolo y un estudio de implementación. Los resultados se refieren a las

versiones y configuraciones evaluadas, y no pueden trasladarse automáticamente a versiones posteriores. El análisis andino se apoya principalmente en evidencia internacional y fuentes contextuales ecuatorianas, lo que restringe la generalización a otros países de la región. No se estableció una relación cuantitativa global entre beneficios, daños y costos.

CONCLUSIONES

Los LLM muestran potencial para apoyar tareas médicas específicas, con resultados favorables en algunos escenarios de diagnóstico y comunicación, pero también con limitaciones relevantes en la integración de información y el seguimiento de guías. La evidencia revisada no permite afirmar superioridad clínica general frente a especialistas ni seguridad equivalente en pacientes. Los resultados de sistemas de IA de imagen o soporte automatizado deben interpretarse de forma independiente de los LLM.

En Ecuador y otros países de ingresos medios, la incorporación de estas herramientas requiere validación contextual y prospectiva, supervisión profesional, protección de datos, interoperabilidad y evaluación de costos y resultados. La adopción debe avanzar según la evidencia generada en cada uso previsto y conservar la responsabilidad clínica del profesional.

REFERENCIAS BIBLIOGRÁFICAS

1. World Health Organization. Noncommunicable diseases [Internet]. Geneva: WHO; 2025 [citado 18 may 2026]. Disponible en: <https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases>
2. World Health Organization. Health workforce [Internet]. Geneva: WHO [citado 18 may 2026]. Disponible en: <https://www.who.int/health-topics/health-workforce>
3. Teles AS, Abd-Alrazaq A, Heston TF, Damseh R, Ruback L. Editorial: Large Language Models for medical applications. *Front Med (Lausanne)*. 2025;12:1625293. doi:10.3389/fmed.2025.1625293.
4. Ríos-Hoyo A, Shan NL, Li A, Pearson AT, Pusztai L, Howard FM. Evaluation of large language models as a diagnostic aid for complex medical cases. *Front Med (Lausanne)*. 2024;11:1380148. doi:10.3389/fmed.2024.1380148.
5. Artsi Y, Sorin V, Glicksberg BS, Korfiatis P, Freeman R, Nadkarni GN, et al. Challenges of Implementing LLMs in Clinical Practice: Perspectives. *J Clin Med*. 2025;14(17):6169. doi:10.3390/jcm14176169.
6. Gao Y, Feng Y, Jahng SG. Implicit bias in digital health: systematic biases in large language models' representation of global public health attitudes and challenges to health equity. *Front Public Health*. 2025;13:1705082. doi:10.3389/fpubh.2025.1705082.
7. Santorum Gaibor M, Carrión Toro M. ¿Conectados pero no incluidos? Brecha digital e inclusión en el Ecuador (2013-2025) [Internet]. Quito: Escuela Politécnica Nacional, ÁPICE y LudoLab; 2025. Disponible en: <https://apice.epn.edu.ec/images/apice/noticias/boletin/Conectados%20pero%20no%20incluidos%20vf.pdf>
8. Ministerio de Salud Pública del Ecuador. Política Nacional de Transformación Digital del Sector Salud 2024-2034 [Internet]. Quito: Ministerio de Salud Pública; 2024. Disponible en: https://www.salud.gob.ec/wp-content/uploads/2025/07/politica_nacional_de_transformaciOn_digital_del_sector_salud_14nov2024_.pdf
9. PRISMA. PRISMA 2020 flow diagram [Internet]. PRISMA Statement [citado 22 oct 2026]. Disponible en: <https://www.prisma-statement.org/prisma-2020-flow-diagram>

10. Martinelli C, Giordano A, Carnevale V, Burk SR, Porto L, Vizzielli G, et al. The PERFORM study: Artificial intelligence versus human residents in cross-sectional obstetrics-gynecology scenarios across languages and time constraints. *Mayo Clin Proc Digit Health*. 2025;3(2):100206. doi:10.1016/j.mcpdig.2025.100206.
11. Hager P, Jungmann F, Holland R, Bhagat K, Hubrecht I, Knauer M, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat Med*. 2024;30(9):2613-2622. doi:10.1038/s41591-024-03097-1.
12. Rotem R, Simon C, Rottenstreich M, O'Reilly BA, Weintraub AY, O'Sullivan OE. ChatGPT in urogynecology: Comparing large language model responses to human experts. *Acta Obstet Gynecol Scand*. 2026;105(8):1629-1635. Publicado en línea el 23 de octubre de 2025. doi:10.1111/aogs.70085.
13. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. *JAMA*. 2023;329(10):842-844. doi:10.1001/jama.2023.1044.
14. Pedro T, Sousa JM, Fonseca L, Gama MG, Moreira G, Pintalhão M, et al. Exploring the use of ChatGPT in predicting anterior circulation stroke functional outcomes after mechanical thrombectomy: a pilot study. *J Neurointerv Surg*. 2025;17(3):261-265. doi:10.1136/jnis-2024-021556.
15. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Amin M, et al. Toward expert-level medical question answering with large language models. *Nat Med*. 2025;31(3):943-950. doi:10.1038/s41591-024-03423-7.
16. Sandmann S, Hegselmann S, Fujarski M, Bickmann L, Wild B, Eils R, et al. Benchmark evaluation of DeepSeek large language models in clinical decision-making. *Nat Med*. 2025;31(8):2546-2549. doi:10.1038/s41591-025-03727-2.
17. Chen B, Li Y, Sun Y, Sun H, Wang Y, Lyu J, et al. A 3D and explainable artificial intelligence model for evaluation of chronic otitis media based on temporal bone computed tomography: Model development, validation, and clinical application. *J Med Internet Res*. 2024;26:e51706. doi:10.2196/51706.
18. Huang X, Zhou S, Ma X, Jiang S, Xu Y, You Y, et al. Effectiveness of an artificial intelligence clinical assistant decision support system to improve the incidence of hospital-associated venous thromboembolism: a prospective, randomised controlled study. *BMJ Open Qual*. 2023;12(4):e002267. doi:10.1136/bmjopen-2023-002267.
19. Wu Y, Min H, Li M, Shi Y, Ma A, Han Y, et al. Effect of Artificial Intelligence-based Health Education Accurately Linking System (AI-HEALS) for Type 2 diabetes self-management: protocol for a mixed-methods study. *BMC Public Health*. 2023;23:1325. doi:10.1186/s12889-023-16066-z.
20. de Marques DG, Summer H, Barreto GLL, de Miranda AFS, da Silva MRB, de Vasconcelos G, et al. Cuida-AI: Virtual assistant for optimizing oncology care in public hospitals in Recife Brazil. In: *Proceedings of 20th Iberian Conference on Information Systems and Technologies (CISTI 2025)*. Lecture Notes in Networks and Systems. Vol. 1716. Cham: Springer; 2026. p. 73-85. doi:10.1007/978-3-032-10929-3_7.
21. Amores Zurita D. Sistema de Salud Pública: una crisis que cobra vidas [Internet]. INREDH; 23 abr 2026 [citado 24 may 2026]. Disponible en: <https://inredh.org/sistema-de-salud-publica-una-crisis-que-cobra-vidas/>

FINANCIACIÓN

El autor declara que esta investigación no recibió financiamiento externo.

CONFLICTO DE INTERESES

El autor declara no tener relaciones financieras ni no financieras relevantes para el contenido del manuscrito.

CONTRIBUCIÓN DE AUTORÍA

Conceptualización: Byron Mena.

Metodología: Byron Mena.

Investigación: Byron Mena.

Análisis formal: Byron Mena.

Redacción – borrador original: Byron Mena.

Redacción – revisión y edición: Byron Mena.

CONSIDERACIONES ÉTICAS

Por tratarse de una revisión de literatura y documentos públicos, sin recolección de datos individuales de participantes, no se requirieron aprobación por un comité de ética ni consentimiento informado.

DISPONIBILIDAD DE DATOS

La información sintetizada procede de las publicaciones y documentos citados. La Tabla 1 presenta las características y resultados utilizados en la síntesis narrativa.

USO DE INTELIGENCIA ARTIFICIAL

El autor declara que no se utilizó inteligencia artificial en la elaboración de este manuscrito.

AGRADECIMIENTOS

Al Dr. Francisco Isaac Mena Acosta, ginecólogo y obstetra, por su aporte y orientación en los conceptos clínicos ginecológicos durante el planteamiento del manuscrito.