

REVISÃO 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 generalista e engenheiro de redes de informação, prática privada, Ambato, Equador.

Citação sugerida: Mena Acosta BH. 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. Rev Méd Loja. 2026;1(2):57-66. doi:10.67370/e53pcq28

Recebido: 25-07-2026

Aceito: 13-10-2026

Publicado: 22-10-2026

Editor: Dr. Miguel David Alvarez 

Autor de correspondente: 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.

INTRODUÇÃO

As doenças não transmissíveis representam uma carga importante para os sistemas de saúde. A Organização Mundial da Saúde estima que 73% das mortes por essas doenças ocorrem em países de baixa e média renda.⁽¹⁾ A essa carga somam-se problemas de distribuição e disponibilidade da força de trabalho em saúde que dificultam o acesso oportuno ao atendimento.⁽²⁾ Nesse contexto, as ferramentas de inteligência artificial (IA) têm despertado interesse como apoio ao trabalho clínico e à comunicação com os pacientes.

Os modelos de linguagem de grande escala (LLM, na sigla em inglês) constituem uma família de sistemas de IA capazes de gerar e processar texto. Suas aplicações médicas abrangem a elaboração

de diagnósticos diferenciais, a resposta a perguntas clínicas e a preparação de informações educativas.^(3,4) Contudo, um resultado favorável em questões de exame ou casos escritos não demonstra, por si só, utilidade clínica em pacientes. A integração aos serviços de saúde também exige considerar acurácia, viés, interpretabilidade e condições de uso.⁽⁵⁾

A representação desigual de populações e contextos nos dados e nas avaliações pode limitar a pertinência das respostas fora dos ambientes estudados. As evidências sobre vieses na representação de atitudes de saúde pública reforçam a necessidade de examinar a validade contextual dos modelos.⁽⁶⁾ No Equador, o aumento do acesso à internet não eliminou as lacunas educacionais, territoriais e de competências digitais descritas em relatórios nacionais.⁽⁷⁾ Da mesma forma, a Política Nacional de Transformação Digital do Setor Saúde 2024–2034 identifica necessidades de modernização e interoperabilidade que condicionam a incorporação de tecnologias emergentes.⁽⁸⁾

Esta revisão narrativa analisa o desempenho e a segurança clínica dos LLM em diferentes tarefas médicas, incorpora exemplos diferenciados de outros sistemas de IA e discute as condições necessárias para sua implementação em países de renda média, com ênfase no Equador. A análise do contexto andino é apresentada como uma interpretação de aplicabilidade, e não como uma validação regional dos modelos.

METODOLOGIA

Realizou-se uma revisão narrativa com busca estruturada e síntese temática. A busca foi efetuada em maio de 2026 no PubMed e no Google Scholar e orientou-se a publicações de 2022 até a data da busca. No PubMed, combinaram-se os termos “Large Language Models” OR “ChatGPT” com “Clinical Decision-Making” OR “Patient Safety”, utilizando os campos de título e resumo e o filtro de texto completo gratuito. Foram examinados os primeiros 200 resultados ordenados pelo critério Best Match. Complementarmente, revisaram-se os primeiros 100 resultados do Google Scholar com os mesmos conceitos e ordenação por relevância.

Foram selecionadas publicações sobre acurácia diagnóstica, prognóstico, comunicação com pacientes, segurança das respostas ou apoio às decisões clínicas. Priorizaram-se avaliações empíricas com comparadores humanos, modelos alternativos, escalas clínicas ou atendimento convencional. Também foram incorporados, em uma categoria contextual distinta, um protocolo sobre educação e autocuidado em diabetes e um estudo de caso de um assistente oncológico. Os documentos institucionais, um editorial e outras fontes de contexto foram utilizados para a introdução e a discussão, sem serem considerados estudos de desempenho clínico.

Um único autor realizou a seleção e a extração dos dados mediante revisão de títulos, resumos e textos completos potencialmente pertinentes. Os dados foram organizados por publicação, tarefa clínica, desenho, população ou unidade de análise, sistema avaliado, comparador e achados. O corpus da síntese compreendeu 12 publicações: oito avaliações de LLM, dois estudos de outros sistemas de IA clínica e dois documentos sobre implementação. Estes últimos não foram utilizados para estimar eficácia ou segurança clínica.

A síntese foi narrativa e não se calcularam estimativas combinadas devido à heterogeneidade de tarefas, desenhos, comparadores e desfechos. Não se registrou um protocolo prospectivo nem se aplicaram ferramentas formais de avaliação do risco de viés ou da certeza da evidência. O limite de resultados revisados, o filtro de acesso gratuito e a seleção por um único autor restringem a cobertura e a reprodutibilidade da busca. Consequentemente, a revisão não se apresenta como uma revisão sistemática exaustiva nem como uma metanálise. A Figura 1 resume as contagens do

processo de busca e seleção mediante uma adaptação do fluxograma PRISMA 2020,⁽⁹⁾ utilizada unicamente para organizar sua apresentação.

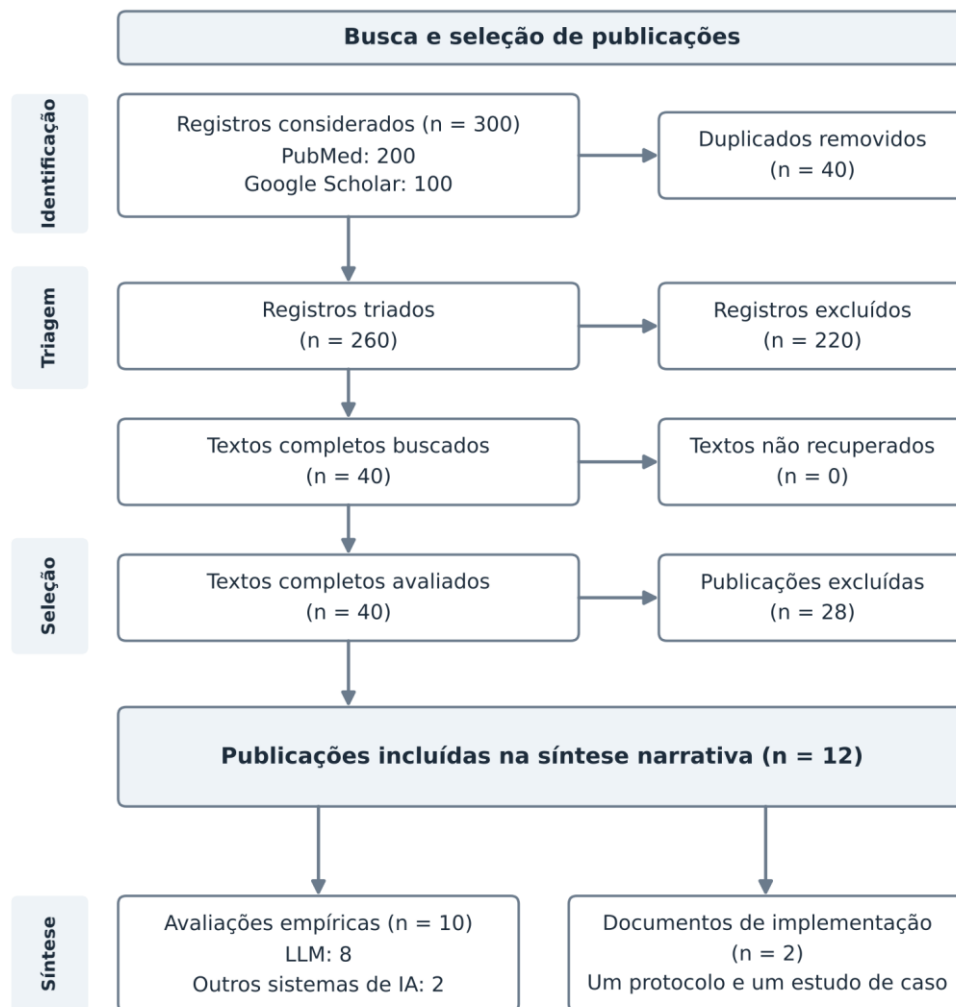


Figura 1. Diagrama de busca e seleção de publicações. Estrutura adaptada do PRISMA 2020,⁽⁹⁾ com as contagens da busca original. Os 300 registros correspondem aos resultados considerados dentro dos limites da busca, e não ao total de correspondências das plataformas. Sua apresentação não implica exaustividade nem cumprimento dos requisitos de uma revisão sistemática.

RESULTADOS DA SÍNTESE NARRATIVA

Os trabalhos revisados abordaram tarefas diferentes: geração de diagnósticos diferenciais, resolução de cenários clínicos, comunicação escrita, predição de desfechos, avaliação do dano potencial das respostas, interpretação de imagens e prevenção de eventos hospitalares. Seus resultados são descritos por tipo de tarefa. A Tabela 1 reúne as 12 publicações e distingue as avaliações de desempenho dos documentos de implementação.

Diagnóstico e resolução de cenários clínicos

Ríos-Hoyo et al. avaliaram o GPT-3.5 e o GPT-4 em 75 casos clínicos complexos. O GPT-4 incluiu o diagnóstico correto em algum ponto do diagnóstico diferencial em 68% dos casos, frente a 48% com o GPT-3.5 ($p=0,0063$). Entretanto, o diagnóstico colocado em primeiro lugar foi correto com frequência muito menor. Esse resultado apoia uma possível utilidade para ampliar o diagnóstico diferencial, mas não demonstra que o modelo substitua a priorização diagnóstica do médico.⁽⁴⁾

No estudo PERFORM, desenvolvido com cenários de obstetrícia e ginecologia e residentes de diferentes anos de formação, os oito LLM avaliados alcançaram, em conjunto, acurácia de 73,75% (IC de 95%: 69,64–77,49), frente a 65,35% (IC de 95%: 62,85–67,76) dos residentes ($p<0,001$); os modelos de alto desempenho (o1-preview, GPT-4o e Claude Sonnet 3.5) alcançaram acurácia de 88,33% nos dois idiomas avaliados. Foram exploradas diferenças por idioma e por restrição de tempo. A comparação foi realizada com residentes e em tarefas padronizadas; por isso, não pode ser interpretada como superioridade geral frente a especialistas no atendimento clínico real.⁽¹⁰⁾

Os achados não foram uniformemente favoráveis. Hager et al. estudaram LLM abertos em uma estrutura baseada em 2400 casos de quatro causas de dor abdominal. Os modelos tiveram desempenho inferior ao dos médicos e mostraram dificuldades para seguir diretrizes, interpretar resultados e integrar informações clínicas.⁽¹¹⁾ Esses resultados contrastam com o desempenho em testes de conhecimento e ressaltam a importância de avaliar processos clínicos completos.

Comunicação com pacientes e recomendações preventivas

Rotem et al. compararam respostas do GPT-4o e de um uroginecologista, avaliadas por 203 pacientes. As pontuações globais foram mediana de 76 (intervalo interquartil: 67–85) para o ChatGPT e de 72 (intervalo interquartil: 63–80) para o consultor ($p<0,01$). A escala avaliou compreensibilidade, ajuda percebida e tranquilização, com máximo de 15 pontos por resposta e 90 para as seis perguntas. Essas pontuações refletem a percepção da comunicação e não constituem uma medida de acurácia diagnóstica nem de desfechos clínicos.⁽¹²⁾

Em cardiologia preventiva, Sarraju et al. formularam 25 perguntas ao ChatGPT e submeteram as respostas à avaliação de três especialistas. As respostas a 21 perguntas (84%) foram consideradas apropriadas nos dois contextos avaliados: informação educativa e rascunhos de mensagens para pacientes. O estudo foi realizado em dezembro de 2022 e publicado em 2023. Seu desenho avaliou a adequação das respostas, sem medir redução de eventos cardiovasculares nem superioridade frente ao manejo médico.⁽¹³⁾

Predição de desfechos neurológicos

Pedro et al. analisaram retrospectivamente 163 pacientes com acidente vascular cerebral isquêmico tratados por trombectomia. A concordância entre as predições do ChatGPT e o resultado funcional em três meses foi $\kappa=0,354$ (IC de 95%: 0,260–0,448) para a pontuação exata da escala de Rankin modificada e $\kappa=0,727$ (IC de 95%: 0,620–0,833) para sua classificação dicotômica. O artigo descreveu esses níveis como fair e good, respectivamente, e encontrou desempenho superior ao do MT-DRAGON. Trata-se de predição prognóstica, e não de diagnóstico nem de comparação direta com especialistas.⁽¹⁴⁾

Avaliação de respostas e dano potencial

Singhal et al. relataram desempenho de até 86,5% do Med-PaLM 2 no MedQA, um conjunto de questões no estilo do USMLE. Em uma avaliação de respostas extensas, a proporção classificada na categoria de ausência de dano potencial foi de 93,3% para o Med-PaLM 2 e de 92,9% para as respostas médicas ($p=0,950$). Essas porcentagens representam julgamentos sobre respostas, e não taxas observadas de eventos adversos em pacientes. A ausência de diferença estatisticamente significativa também não estabelece equivalência nem não inferioridade clínica. As comparações com especialistas variaram segundo a tarefa e o ambiente de avaliação.⁽¹⁵⁾

Comparação entre modelos abertos e proprietários

Sandmann et al. compararam sistemas de linguagem em 125 casos de cinco especialidades. Não observaram diferenças estatisticamente significativas entre o GPT-4o e o DeepSeek-R1 no

diagnóstico ($p=0,3085$) nem nas recomendações terapêuticas ($p=0,1522$). Os tamanhos de efeito pela correlação bisserial de postos foram 0,27 e 0,26, respectivamente. Esses valores descrevem diferenças nas pontuações de desempenho; não representam concordância entre modelos nem provam equivalência formal. Os resultados sugerem que certos modelos abertos podem ser alternativas para tarefas concretas, sujeitos a validação e infraestrutura apropriadas.⁽¹⁶⁾

Outros sistemas de inteligência artificial clínica

Chen et al. desenvolveram um modelo de redes neurais convolucionais 3D para tomografias do osso temporal. Os conjuntos interno e externo incluíram 1661 e 108 pacientes; as AUROC foram de 0,96 e 0,93 para detectar orelhas patológicas e de 0,85 e 0,83 para identificar colesteatoma. Em 121 pacientes prospectivos, a acurácia diagnóstica pré-operatória foi de 81,8%, e o sistema contribuiu para as decisões clínicas em 90,1% dos casos. Esta última porcentagem não deve ser confundida com acurácia diagnóstica. O sistema é uma IA de imagem, e não um LLM.⁽¹⁷⁾

Huang et al. avaliaram um sistema automatizado de apoio clínico para prevenir o tromboembolismo venoso hospitalar em 19 785 pacientes. Os eventos ocorreram em 27 de 9895 pacientes do grupo de intervenção e em 49 de 9890 controles (OR: 0,55; IC de 95%: 0,34–0,88), com redução relativa relatada de 46%. Esse desfecho clínico pertence a um sistema específico de prevenção e não pode ser atribuído aos LLM nem extrapolado para outras tarefas de IA.⁽¹⁸⁾

Documentos sobre implementação

O protocolo AI-HEALS de Wu et al. descreve uma intervenção por meio do WeChat para educação e autocuidado em diabetes tipo 2, com HbA1c e outros indicadores previstos como desfechos. Por ser um protocolo, não fornece resultados de eficácia ou segurança.⁽¹⁹⁾ Por sua vez, de Marques et al. descrevem o Cuida-AI como um assistente de IA generativa desenvolvido para fornecer informação oncológica em hospitais públicos do Recife, Brasil. Constitui uma experiência contextual de desenvolvimento e implementação, sem fornecer a esta síntese uma estimativa comparativa de benefício clínico frente aos médicos.⁽²⁰⁾

Tabela 1. Características das publicações revisadas

Publicação	Desenho e unidade de análise	Sistema e categoria	Comparador	Achados e interpretação
Ríos-Hoyo et al.⁽⁴⁾ 2024	Avaliação de 75 casos complexos	GPT-3.5 e GPT-4 LLM	Modelos e diagnósticos de referência dos casos clínicos	O GPT-4 incluiu o diagnóstico correto no diferencial em 68% vs. 48% com o GPT-3.5; $p=0,0063$. A inclusão na lista não equivale a um diagnóstico principal correto.
Martinelli et al.⁽¹⁰⁾ 2025	Cenários clínicos de obstetrícia e ginecologia	Oito LLM, incluindo o1-preview, GPT-4o e Claude Sonnet 3.5 LLM	Residentes de ginecologia e obstetrícia	Acurácia global dos LLM de 73,75% (IC95%: 69,64–77,49) vs. 65,35% (IC95%: 62,85–67,76); $p<0,001$. Modelos de alto desempenho: 88,33%. Avaliação padronizada com exploração de idioma e tempo.
Rotem et al.⁽¹²⁾ 2025 online	Inquérito transversal 203 pacientes; seis perguntas	GPT-4o LLM	Um uroginecologista	Pontuação global mediana 76 (IIQ: 67–85) vs. 72 (IIQ: 63–80); $p<0,01$. Percepção comunicativa, sem demonstrar superioridade diagnóstica.
Pedro et al.⁽¹⁴⁾ 2025	Estudo retrospectivo 163 pacientes com AVC e trombectomia	ChatGPT LLM	mRS observada e escore MT-DRAGON	Concordância prognóstica $\kappa=0,354$ para a pontuação exata e $\kappa=0,727$ para o resultado dicotômico. Não corresponde à acurácia diagnóstica.
Sarraj et al.⁽¹³⁾ 2023	Avaliação de 25 perguntas preventivas	ChatGPT LLM	Avaliação por três especialistas em cardiologia	21/25 respostas (84%) apropriadas em ambos os contextos. Estudo de adequação de conteúdo, sem desfechos cardiovasculares.
Chen et al.⁽¹⁷⁾ 2024	Validação interna: 1661 pacientes; externa: 108; coorte prospectiva: 121	Rede convolucional 3D para TC do osso temporal IA de imagem	Modelo 2D e sete clínicos	AUROC interna/externa: 0,96/0,93 para orelha patológica e 0,85/0,83 para colesteatoma. Acurácia prospectiva de 81,8%; contribuição para decisões em 90,1%. Não é um LLM.
Wu et al.⁽¹⁹⁾ 2023	Protocolo de métodos mistos Diabetes tipo 2	AI-HEALS via WeChat Documento de implementação	Atendimento habitual previsto	Inclui HbA1c e autocuidado como desfechos previstos. O protocolo não contém resultados de eficácia ou segurança.

Publicação	Desenho e unidade de análise	Sistema e categoria	Comparador	Achados e interpretação
Hager et al. ⁽¹¹⁾ 2024	Estrutura de avaliação com 2400 casos de quatro doenças abdominais	LLM abertos, incluindo Llama 2, OASST, WizardLM, Clinical Camel e Meditron	Médicos e recomendações de diretrizes	Desempenho inferior ao do médico e dificuldades de integração de informações e de seguimento de diretrizes. Evidência relevante sobre limitações clínicas.
Huang et al. ⁽¹⁸⁾ 2023	Estudo prospectivo randomizado 19 785 pacientes internados	Sistema de apoio clínico para prevenção de TEV IA distinta de LLM	Atendimento convencional	TEV: 27/9895 vs. 49/9890; OR 0,55 (IC95%: 0,34–0,88). Não extrapolar o benefício do sistema de prevenção para os LLM.
Singhal et al. ⁽¹⁵⁾ 2025	Avaliação de perguntas médicas e respostas extensas	Med-PaLM 2 e Med-PaLM LLM	Respostas médicas e outros modelos	MedQA: até 86,5%. Respostas classificadas sem dano potencial: 93,3% vs. 92,9% das médicas; p=0,950. Não demonstra equivalência clínica.
Sandmann et al. ⁽¹⁶⁾ 2025	Benchmark de 125 casos em cinco especialidades	DeepSeek-R1/V3, GPT-4o e outros LLM	Comparação entre sistemas	GPT-4o vs. DeepSeek-R1: diagnóstico p=0,3085 e tratamento p=0,1522. Sem diferenças significativas; não constitui prova formal de equivalência.
de Marques et al. ⁽²⁰⁾ 2026	Estudo de caso em hospitais públicos do Recife	Cuida-AI Assistente de IA generativa	Sem comparação clínica direta utilizada nesta síntese	Desenvolvimento de um assistente de informação oncológica. Considerado uma experiência de implementação latino-americana, sem estimar eficácia clínica comparativa.

LLM: modelo de linguagem de grande escala; IC95%: intervalo de confiança de 95%; IIQ: intervalo interquartil; AVC: acidente vascular cerebral; mRS: escala de Rankin modificada; TC: tomografia computadorizada; AUROC: área sob a curva ROC; TEV: tromboembolismo venoso; OR: odds ratio. As unidades de análise e os desfechos são heterogêneos; não se estimou um efeito conjunto.

DISCUSSÃO

As evidências revisadas mostram desempenho variável segundo a tarefa e o contexto. Os resultados favoráveis em cenários padronizados e na comunicação escrita coexistem com limitações em processos de decisão que exigem coletar dados, integrar achados e seguir recomendações clínicas. Por isso, não é adequado resumir o conjunto como uma demonstração de superioridade dos LLM frente a especialistas. A distinção entre residentes, especialistas, escalas prognósticas e sistemas alternativos é essencial para interpretar cada comparação.^(4,10-16)

A acurácia diagnóstica, a concordância prognóstica, a satisfação comunicativa e a segurança são dimensões relacionadas, mas não intercambiáveis. Uma resposta clara pode ser clinicamente incompleta, e um diagnóstico correto incluído em uma lista extensa pode não orientar adequadamente a conduta. Da mesma forma, um modelo que responde a perguntas médicas com pontuação alta não necessariamente identifica quando faltam dados, quando deve abster-se ou quando é necessário atendimento urgente. A seleção dos desfechos deve corresponder à função clínica prevista.^(4,11,12,15)

A avaliação da segurança requer cautela especial. As rubricas de dano potencial fornecem informações sobre o conteúdo das respostas, mas não substituem o seguimento de pacientes expostos às recomendações. Para sustentar equivalência ou não inferioridade, seriam necessários desenhos específicos, margens predefinidas e estimativas suficientemente precisas. Valores de p superiores a 0,05 em comparações de superioridade não permitem concluir igualdade de desempenho. Também se deve diferenciar o benefício de sistemas especializados, como o apoio automatizado para prevenir o tromboembolismo, da segurança dos modelos generativos de uso geral.^(15,16,18)

Aplicabilidade no Equador e em outros países de renda média

A extrapolação para o Equador depende tanto dos serviços disponíveis quanto das capacidades do modelo. Os cenários estudados não representam necessariamente a prevalência de doenças, as formas de consulta, a linguagem clínica, os níveis de alfabetização nem os recursos diagnósticos ou terapêuticos da atenção primária. Nesta revisão, a aplicabilidade regional deriva de uma interpretação contextual; não se identificou no corpus uma validação clínica equatoriana dos

modelos descritos. As evidências sobre representação desigual e vieses reforçam a necessidade de confrontar as respostas em populações locais.^(5,6)

O boletim da Escuela Politécnica Nacional aponta que o acesso dos domicílios à internet aumentou de 28% em 2013 para 71,3% em 2025, embora persistam diferenças territoriais e de competências digitais.⁽⁷⁾ Por sua vez, a política nacional de transformação digital reúne um diagnóstico técnico de 2024 segundo o qual 95,55% dos estabelecimentos de atenção primária tinham conectividade funcional, além de obsolescência e interoperabilidade insuficiente dos sistemas.⁽⁸⁾ Esses números provêm de anos e unidades de análise diferentes e não medem a preparação institucional para implementar LLM.

A disponibilidade de modelos abertos pode ampliar as opções de adaptação e implantação, mas não elimina os custos de infraestrutura, manutenção, segurança, integração e capacitação. O desempenho comparável observado em determinadas tarefas deve ser confrontado com a capacidade técnica de cada instituição e com a sustentabilidade do serviço.⁽¹⁶⁾ Além disso, uma análise da sociedade civil sobre o sistema público equatoriano descreve problemas de disponibilidade de insumos e medicamentos; ela é utilizada aqui como contexto, e não como medida da capacidade de adotar IA.⁽²¹⁾

Para uma implementação gradual, propõe-se começar com funções de apoio delimitadas e supervisionadas, como a preparação de informações educativas ou de rascunhos que o profissional revise antes de utilizá-los. A avaliação local deveria incluir a acurácia e a pertinência das respostas, o reconhecimento de sinais de alarme, recomendações incompatíveis com os recursos disponíveis, omissões e desempenho por subgrupos. Antes de ampliar o uso, convém comparar o trabalho clínico habitual com o assistido e medir os efeitos sobre decisões, tempos, erros e resultados relevantes para os pacientes.

A governança deve definir quem revisa as respostas, como se documentam as decisões e como se gerenciam incidentes ou mudanças de versão. Também requer proteger a informação clínica e avaliar as condições de tratamento de dados de cada sistema. Essas são condições de implementação propostas a partir da síntese; não representam benefícios demonstrados pelos estudos revisados. Experiências regionais como o Cuida-AI mostram linhas de desenvolvimento que requerem avaliação clínica adicional antes de extrapolar sua utilidade para outros serviços.^(5,8,20)

Limitações da revisão

A busca restringiu-se a duas fontes, a um número limitado de resultados ordenados por relevância e, no PubMed, a textos completos gratuitos. Um único autor realizou a seleção e a extração; não houve uma segunda avaliação independente. A falta de datas exatas e de estratégias completas específicas para cada plataforma limita a reprodutibilidade. Tampouco se registrou um protocolo nem se avaliaram formalmente o risco de viés ou a certeza da evidência. Essas decisões podem introduzir vieses de seleção e de interpretação e fazem com que a síntese não represente toda a literatura disponível.

O corpus reuniu desenhos e unidades de análise diferentes, incluindo cenários escritos, perguntas, pacientes, um protocolo e um estudo de implementação. Os resultados referem-se às versões e configurações avaliadas e não podem ser transferidos automaticamente para versões posteriores. A análise andina apoia-se principalmente em evidências internacionais e em fontes contextuais equatorianas, o que restringe a generalização para outros países da região. Não se estabeleceu uma relação quantitativa global entre benefícios, danos e custos.

CONCLUSÕES

Os LLM mostram potencial para apoiar tarefas médicas específicas, com resultados favoráveis em alguns cenários de diagnóstico e comunicação, mas também com limitações relevantes na integração de informações e no seguimento de diretrizes. As evidências revisadas não permitem afirmar superioridade clínica geral frente a especialistas nem segurança equivalente em pacientes. Os resultados de sistemas de IA de imagem ou de suporte automatizado devem ser interpretados de forma independente dos LLM.

No Equador e em outros países de renda média, a incorporação dessas ferramentas requer validação contextual e prospectiva, supervisão profissional, proteção de dados, interoperabilidade e avaliação de custos e resultados. A adoção deve avançar de acordo com as evidências geradas para cada uso previsto e preservar a responsabilidade clínica do profissional.

REFERÊNCIAS

1. World Health Organization. Noncommunicable diseases [Internet]. Geneva: WHO; 2025 [citado 18 maio 2026]. Disponível em: <https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases>
2. World Health Organization. Health workforce [Internet]. Geneva: WHO [citado 18 maio 2026]. Disponível em: <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. Disponível em: <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. Disponível em: 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]. Disponível em: <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 online em 23 de outubro 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 maio 2026]. Disponível em: <https://inredh.org/sistema-de-salud-publica-una-crisis-que-cobra-vidas/>

FINANCIAMENTO

O autor declara que esta pesquisa não recebeu financiamento externo.

CONFLITO DE INTERESSES

O autor declara não ter relações financeiras nem não financeiras relevantes para o conteúdo do manuscrito.

CONTRIBUIÇÃO DE AUTORIA

Conceitualização: Byron Mena.

Metodologia: Byron Mena.

Investigação: Byron Mena.

Análise formal: Byron Mena.

Redação do rascunho original: Byron Mena.

Redação, revisão e edição: Byron Mena.

CONSIDERAÇÕES ÉTICAS

Por se tratar de uma revisão de literatura e de documentos públicos, sem coleta de dados individuais de participantes, não foram necessários aprovação por comitê de ética nem consentimento informado.

DISPONIBILIDADE DE DADOS

As informações sintetizadas provêm das publicações e documentos citados. A Tabela 1 apresenta as características e os resultados utilizados na síntese narrativa.

USO DE INTELIGÊNCIA ARTIFICIAL

O autor declara que não foi utilizada inteligência artificial na elaboração deste manuscrito.

AGRADECIMENTOS

Ao Dr. Francisco Isaac Mena Acosta, ginecologista e obstetra, por sua contribuição e orientação nos conceitos clínicos ginecológicos durante a elaboração do manuscrito.