

NARRATIVE REVIEW

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  

¹General practitioner and information networks engineer, private practice, Ambato, Ecuador.

Suggested citation: Mena Acosta BH. Evaluation of artificial intelligence and large language models (LLMs) in clinical practice: precision, safety, and applicability in middle-income countries. Rev Méd Loja. 2026;1(2):57-66. [doi:10.67370/e53pcq28](https://doi.org/10.67370/e53pcq28)

Received: 25-07-2026

Accepted: 13-10-2026

Published: 22-10-2026

Editor: Dr. Miguel David Alvarez 

Corresponding author: 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.

INTRODUCTION

Noncommunicable diseases represent a major burden for health systems. The World Health Organization estimates that 73% of deaths from these diseases occur in low- and middle-income countries.⁽¹⁾ This burden is compounded by problems in the distribution and availability of the health workforce that hinder timely access to care.⁽²⁾ In this context, artificial intelligence (AI) tools have generated interest as support for clinical work and for communication with patients.

Large language models (LLMs) are a family of AI systems capable of generating and processing text. Their medical applications include generating differential diagnoses, answering clinical questions, and preparing educational information.^(3,4) However, a favorable result on examination

questions or written cases does not by itself demonstrate clinical usefulness in patients. Integration into health services also requires consideration of accuracy, bias, interpretability, and conditions of use.⁽⁵⁾

Unequal representation of populations and contexts in data and evaluations may limit the relevance of responses outside the settings studied. Evidence on biases in the representation of public health attitudes reinforces the need to examine the contextual validity of these models.⁽⁶⁾ In Ecuador, increased internet access has not eliminated the educational, territorial, and digital skills gaps described in national reports.⁽⁷⁾ Likewise, the National Policy for the Digital Transformation of the Health Sector 2024–2034 identifies modernization and interoperability needs that condition the incorporation of emerging technologies.⁽⁸⁾

This narrative review analyzes the performance and clinical safety of LLMs across different medical tasks, incorporates distinct examples of other AI systems, and discusses the conditions necessary for their implementation in middle-income countries, with an emphasis on Ecuador. The analysis of the Andean context is presented as an interpretation of applicability and not as a regional validation of the models.

METHODS

A narrative review with a structured search and thematic synthesis was conducted. The search was performed in May 2026 in PubMed and Google Scholar and targeted publications from 2022 to the search date. In PubMed, the terms “Large Language Models” OR “ChatGPT” were combined with “Clinical Decision-Making” OR “Patient Safety”, using the title and abstract fields and the free full-text filter. The first 200 results sorted by Best Match were examined. In addition, the first 100 Google Scholar results for the same concepts, sorted by relevance, were reviewed.

Publications on diagnostic accuracy, prognosis, communication with patients, safety of responses, or clinical decision support were selected. Empirical evaluations with human comparators, alternative models, clinical scales, or usual care were prioritized. A protocol on diabetes education and self-care and a case study of an oncology assistant were also included in a separate contextual category. Institutional documents, an editorial, and other contextual sources were used for the introduction and discussion, without being considered clinical performance studies.

A single author performed the selection and data extraction by reviewing titles, abstracts, and potentially relevant full texts. Data were organized by publication, clinical task, design, population or unit of analysis, evaluated system, comparator, and findings. The synthesis corpus comprised 12 publications: eight LLM evaluations, two studies of other clinical AI systems, and two implementation documents. The latter were not used to estimate clinical efficacy or safety.

The synthesis was narrative, and no pooled estimates were calculated because of the heterogeneity of tasks, designs, comparators, and outcomes. No prospective protocol was registered, and no formal tools for assessing the risk of bias or the certainty of evidence were applied. The limit on the number of results reviewed, the free-access filter, and selection by a single author restrict the coverage and reproducibility of the search. Consequently, this review is not presented as an exhaustive systematic review or as a meta-analysis. Figure 1 summarizes the counts of the search and selection process using an adaptation of the PRISMA 2020 flow diagram,⁽⁹⁾ used solely to organize their presentation.

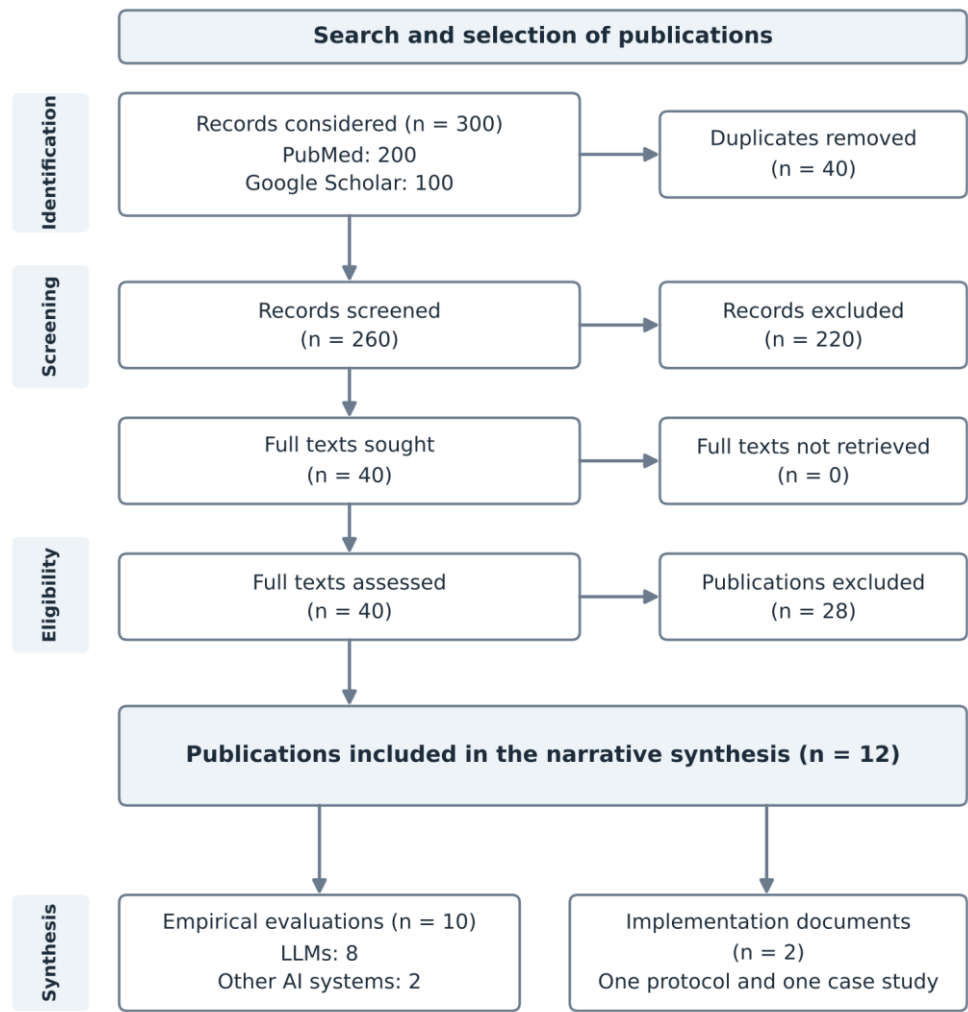


Figure 1. Diagram of the search and selection of publications. Structure adapted from PRISMA 2020,⁽⁹⁾ with the counts from the original search. The 300 records correspond to the results considered within the search limits, not to the total number of matches on the platforms. Its presentation does not imply exhaustiveness or compliance with the requirements of a systematic review.

RESULTS OF THE NARRATIVE SYNTHESIS

The reviewed studies addressed different tasks: generation of differential diagnoses, resolution of clinical scenarios, written communication, outcome prediction, assessment of potential harm in responses, image interpretation, and prevention of hospital events. Their results are described by type of task. Table 1 brings together the 12 publications and distinguishes performance evaluations from implementation documents.

Diagnosis and resolution of clinical scenarios

Ríos-Hoyo et al. evaluated GPT-3.5 and GPT-4 in 75 complex clinical cases. GPT-4 included the correct diagnosis at some point in the differential diagnosis in 68% of cases, compared with 48% for GPT-3.5 (p=0.0063). However, the top-ranked diagnosis was correct far less frequently. This result supports a possible usefulness for broadening the differential diagnosis but does not show that the model can replace the physician’s diagnostic prioritization.⁽⁴⁾

In the PERFORM study, conducted with obstetrics and gynecology scenarios and residents at different years of training, the eight LLMs evaluated achieved an overall accuracy of 73.75% (95% CI: 69.64–77.49), compared with 65.35% (95% CI: 62.85–67.76) for residents (p<0.001); the high-

performing models (o1-preview, GPT-4o, and Claude Sonnet 3.5) achieved an accuracy of 88.33% in both languages evaluated. Differences by language and time constraint were explored. The comparison was made with residents and in standardized tasks; therefore, it cannot be interpreted as general superiority over specialists in real clinical care.⁽¹⁰⁾

The findings were not uniformly favorable. Hager et al. studied open LLMs in a framework based on 2400 cases of four causes of abdominal pain. The models performed worse than physicians and had difficulty following guidelines, interpreting results, and integrating clinical information.⁽¹¹⁾ These results contrast with performance on knowledge tests and underscore the importance of evaluating complete clinical processes.

Communication with patients and preventive recommendations

Rotem et al. compared responses from GPT-4o and from a urogynecologist, as rated by 203 patients. Overall scores were a median of 76 (interquartile range: 67–85) for ChatGPT and 72 (interquartile range: 63–80) for the consultant ($p < 0.01$). The scale assessed understandability, perceived helpfulness, and reassurance, with a maximum of 15 points per response and 90 for the six questions. These scores reflect perceptions of communication and do not constitute a measure of diagnostic accuracy or clinical outcomes.⁽¹²⁾

In preventive cardiology, Sarraju et al. posed 25 questions to ChatGPT and submitted the responses to evaluation by three experts. Responses to 21 questions (84%) were considered appropriate in the two contexts assessed: educational information and draft messages for patients. The study was conducted in December 2022 and published in 2023. Its design assessed the appropriateness of responses, without measuring reductions in cardiovascular events or superiority over medical management.⁽¹³⁾

Prediction of neurological outcomes

Pedro et al. retrospectively analyzed 163 patients with ischemic stroke treated with thrombectomy. Agreement between ChatGPT predictions and the functional outcome at three months was $\kappa = 0.354$ (95% CI: 0.260–0.448) for the exact modified Rankin Scale score and $\kappa = 0.727$ (95% CI: 0.620–0.833) for its dichotomized classification. The article described these levels as fair and good, respectively, and found better performance than MT-DRAGON. This is prognostic prediction, not diagnosis or a direct comparison with specialists.⁽¹⁴⁾

Evaluation of responses and potential harm

Singhal et al. reported a performance of up to 86.5% for Med-PaLM 2 on MedQA, a set of USMLE-style questions. In an evaluation of long-form answers, the proportion rated in the no-potential-harm category was 93.3% for Med-PaLM 2 and 92.9% for physician answers ($p = 0.950$). These percentages represent judgments about answers and not observed rates of adverse events in patients. Nor does the absence of a statistically significant difference establish clinical equivalence or noninferiority. Comparisons with specialists varied according to the task and the evaluation setting.⁽¹⁵⁾

Comparison between open and proprietary models

Sandmann et al. compared language systems in 125 cases from five specialties. They observed no statistically significant differences between GPT-4o and DeepSeek-R1 in diagnosis ($p = 0.3085$) or in treatment recommendations ($p = 0.1522$). Effect sizes based on rank-biserial correlation were 0.27 and 0.26, respectively. These values describe differences in performance scores; they do not represent agreement between models or prove formal equivalence. The results suggest that certain

open models may be alternatives for specific tasks, subject to appropriate validation and infrastructure.⁽¹⁶⁾

Other clinical artificial intelligence systems

Chen et al. developed a 3D convolutional neural network model for temporal bone computed tomography. The internal and external datasets included 1661 and 108 patients; the AUROCs were 0.96 and 0.93 for detecting pathological ears and 0.85 and 0.83 for identifying cholesteatoma. In 121 prospective patients, preoperative diagnostic accuracy was 81.8%, and the system contributed to clinical decisions in 90.1% of cases. This last percentage should not be confused with diagnostic accuracy. The system is imaging AI, not an LLM.⁽¹⁷⁾

Huang et al. evaluated an automated clinical decision support system for preventing hospital-associated venous thromboembolism in 19 785 patients. Events occurred in 27 of 9895 patients in the intervention group and in 49 of 9890 controls (OR: 0.55; 95% CI: 0.34–0.88), with a reported relative reduction of 46%. This clinical outcome belongs to a specific prevention system and cannot be attributed to LLMs or extrapolated to other AI tasks.⁽¹⁸⁾

Implementation documents

The AI-HEALS protocol by Wu et al. describes a WeChat-based intervention for education and self-management in type 2 diabetes, with HbA1c and other indicators planned as outcomes. As a protocol, it provides no efficacy or safety results.⁽¹⁹⁾ In addition, de Marques et al. describe Cuida-AI as a generative AI assistant developed to provide oncology information in public hospitals in Recife, Brazil. It constitutes a contextual development and implementation experience, without contributing to this synthesis a comparative estimate of clinical benefit relative to physicians.⁽²⁰⁾

Table 1. Characteristics of the reviewed publications

Publication	Design and unit of analysis	System and category	Comparator	Findings and interpretation
Ríos-Hoyo et al. ⁽⁴⁾ 2024	Evaluation of 75 complex cases	GPT-3.5 and GPT-4 LLM	Models and reference diagnoses of clinical cases	GPT-4 included the correct diagnosis in the differential in 68% vs. 48% with GPT-3.5; p=0.0063. Inclusion in the list is not equivalent to a correct leading diagnosis.
Martinelli et al. ⁽¹⁰⁾ 2025	Obstetrics and gynecology clinical scenarios	Eight LLMs, including o1-preview, GPT-4o, and Claude Sonnet 3.5 LLM	Obstetrics and gynecology residents	Overall LLM accuracy 73.75% (95% CI: 69.64–77.49) vs. 65.35% (95% CI: 62.85–67.76); p<0.001. High-performing models: 88.33%. Standardized evaluation exploring language and time.
Rotem et al. ⁽¹²⁾ 2025 online	Cross-sectional survey 203 patients; six questions	GPT-4o LLM	One urogynecologist	Median overall score 76 (IQR: 67–85) vs. 72 (IQR: 63–80); p<0.01. Perceived communication, without demonstrating diagnostic superiority.
Pedro et al. ⁽¹⁴⁾ 2025	Retrospective study 163 patients with stroke and thrombectomy	ChatGPT LLM	Observed mRS and MT-DRAGON score	Prognostic agreement κ=0.354 for the exact score and κ=0.727 for the dichotomized outcome. Not a measure of diagnostic accuracy.
Sarraj et al. ⁽¹³⁾ 2023	Evaluation of 25 preventive questions	ChatGPT LLM	Rating by three cardiology experts	21/25 responses (84%) appropriate in both contexts. Content appropriateness study, without cardiovascular outcomes.
Chen et al. ⁽¹⁷⁾ 2024	Internal validation: 1661 patients; external: 108; prospective cohort: 121	3D convolutional network for temporal bone CT Imaging AI	2D model and seven clinicians	Internal/external AUROC: 0.96/0.93 for pathological ear and 0.85/0.83 for cholesteatoma. Prospective accuracy 81.8%; contribution to decisions in 90.1%. Not an LLM.
Wu et al. ⁽¹⁹⁾ 2023	Mixed-methods protocol Type 2 diabetes	AI-HEALS via WeChat Implementation document	Planned usual care	Includes HbA1c and self-management as planned outcomes. The protocol contains no efficacy or safety results.
Hager et al. ⁽¹¹⁾ 2024	Evaluation framework with 2400 cases of four abdominal conditions	Open LLMs, including Llama 2, OASST, WizardLM, Clinical Camel, and Meditron	Physicians and guideline recommendations	Performance inferior to physicians and difficulties integrating information and following guidelines. Relevant evidence on clinical limitations.

Publication	Design and unit of analysis	System and category	Comparator	Findings and interpretation
Huang et al.⁽¹⁸⁾ 2023	Prospective randomized study 19 785 hospitalized patients	Clinical support system for VTE prevention AI other than LLM	Usual care	VTE: 27/9895 vs. 49/9890; OR 0.55 (95% CI: 0.34–0.88). The benefit of the prevention system should not be extrapolated to LLMs.
Singhal et al.⁽¹⁵⁾ 2025	Evaluation of medical questions and long-form answers	Med-PaLM 2 and Med-PaLM LLM	Physician answers and other models	MedQA: up to 86.5%. Answers rated as without potential harm: 93.3% vs. 92.9% for physicians; p=0.950. Does not demonstrate clinical equivalence.
Sandmann et al.⁽¹⁶⁾ 2025	Benchmark of 125 cases across five specialties	DeepSeek-R1/V3, GPT-4o, and other LLMs	Comparison between systems	GPT-4o vs. DeepSeek-R1: diagnosis p=0.3085 and treatment p=0.1522. No significant differences; not a formal test of equivalence.
de Marques et al.⁽²⁰⁾ 2026	Case study in public hospitals in Recife	Cuida-AI Generative AI assistant	No direct clinical comparison used in this synthesis	Development of an oncology information assistant. Considered a Latin American implementation experience, without estimating comparative clinical efficacy.

LLM: large language model; 95% CI: 95% confidence interval; IQR: interquartile range; mRS: modified Rankin Scale; CT: computed tomography; AUROC: area under the ROC curve; VTE: venous thromboembolism; OR: odds ratio. Units of analysis and outcomes are heterogeneous; no pooled effect was estimated.

DISCUSSION

The reviewed evidence shows variable performance depending on the task and the context. Favorable results in standardized scenarios and written communication coexist with limitations in decision processes that require gathering data, integrating findings, and following clinical recommendations. Therefore, it is not appropriate to summarize the body of evidence as a demonstration of LLM superiority over specialists. The distinction between residents, specialists, prognostic scales, and alternative systems is essential to interpret each comparison.^(4,10-16)

Diagnostic accuracy, prognostic agreement, communicative satisfaction, and safety are related but not interchangeable dimensions. A clear answer may be clinically incomplete, and a correct diagnosis included in a long list may not adequately guide management. Similarly, a model that answers medical questions with a high score does not necessarily recognize when data are missing, when it should abstain, or when urgent care is required. The choice of outcomes should correspond to the intended clinical function.^(4,11,12,15)

Safety evaluation requires particular caution. Potential-harm rubrics provide information about the content of responses but do not replace follow-up of patients exposed to recommendations. Claims of equivalence or noninferiority would require specific designs, predefined margins, and sufficiently precise estimates. P values above 0.05 in superiority comparisons do not allow equal performance to be concluded. The benefit of specialized systems, such as automated support for preventing thromboembolism, should also be distinguished from the safety of general-purpose generative models.^(15,16,18)

Applicability in Ecuador and other middle-income countries

Extrapolation to Ecuador depends as much on the available services as on the capabilities of the model. The scenarios studied do not necessarily represent the prevalence of diseases, forms of consultation, clinical language, literacy levels, or diagnostic and therapeutic resources of primary care. In this review, regional applicability derives from a contextual interpretation; no Ecuadorian clinical validation of the models described was identified within the corpus. Evidence on unequal representation and biases reinforces the need to test responses in local populations.^(5,6)

The bulletin of the Escuela Politécnica Nacional reports that household internet access increased from 28% in 2013 to 71.3% in 2025, although territorial and digital skills differences persist.⁽⁷⁾ In turn, the national digital transformation policy includes a 2024 technical assessment according to which 95.55% of primary care facilities had functional connectivity, along with

obsolescence and insufficient interoperability of systems.⁽⁸⁾ These figures come from different years and units of analysis and do not measure institutional readiness to implement LLMs.

The availability of open models may broaden the options for adaptation and deployment, but it does not eliminate the costs of infrastructure, maintenance, security, integration, and training. The comparable performance observed in certain tasks should be weighed against the technical capacity of each institution and the sustainability of the service.⁽¹⁶⁾ In addition, a civil society analysis of the Ecuadorian public system describes problems with the availability of supplies and medicines; it is used here as context and not as a measure of capacity to adopt AI.⁽²¹⁾

For gradual implementation, it is proposed to begin with well-defined and supervised support functions, such as preparing educational information or drafts that the professional reviews before using them. Local evaluation should include the accuracy and relevance of responses, recognition of warning signs, recommendations incompatible with available resources, omissions, and performance by subgroups. Before expanding use, usual clinical work should be compared with assisted work, measuring effects on decisions, time, errors, and outcomes relevant to patients.

Governance should define who reviews responses, how decisions are documented, and how incidents or version changes are managed. It also requires protecting clinical information and evaluating the data processing conditions of each system. These are implementation conditions proposed on the basis of the synthesis; they do not represent benefits demonstrated by the studies reviewed. Regional experiences such as Cuida-AI show lines of development that require additional clinical evaluation before their usefulness can be extrapolated to other services.^(5,8,20)

Limitations of the review

The search was restricted to two sources, to a limited number of results sorted by relevance and, in PubMed, to free full texts. A single author performed selection and extraction; there was no second independent assessment. The lack of exact search dates and of complete platform-specific strategies limits reproducibility. Nor was a protocol registered or the risk of bias or certainty of evidence formally assessed. These decisions may introduce selection and interpretation biases and mean that the synthesis does not represent all of the available literature.

The corpus brought together different designs and units of analysis, including written scenarios, questions, patients, a protocol, and an implementation study. The results refer to the versions and configurations evaluated and cannot be automatically transferred to later versions. The Andean analysis relies mainly on international evidence and Ecuadorian contextual sources, which restricts generalization to other countries in the region. No overall quantitative relationship between benefits, harms, and costs was established.

CONCLUSIONS

LLMs show potential to support specific medical tasks, with favorable results in some diagnostic and communication scenarios, but also with relevant limitations in integrating information and following guidelines. The evidence reviewed does not support claims of general clinical superiority over specialists or of equivalent safety in patients. Results from imaging AI or automated support systems should be interpreted independently of LLMs.

In Ecuador and other middle-income countries, the incorporation of these tools requires contextual and prospective validation, professional supervision, data protection, interoperability, and evaluation of costs and outcomes. Adoption should advance according to the evidence generated for each intended use and preserve the clinical responsibility of the professional.

REFERENCES

1. World Health Organization. Noncommunicable diseases [Internet]. Geneva: WHO; 2025 [cited 2026 May 18]. Available from: <https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases>
2. World Health Organization. Health workforce [Internet]. Geneva: WHO [cited 2026 May 18]. Available from: <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. Available from: <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. Available from: 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 [cited 2026 Oct 22]. Available from: <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. Published online 2025 Oct 23. 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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/10.1007/978-3-032-10929-3_7).
21. Amores Zurita D. Sistema de Salud Pública: una crisis que cobra vidas [Internet]. INREDH; 2026 Apr 23 [cited 2026 May 24]. Available from: <https://inredh.org/sistema-de-salud-publica-una-crisis-que-cobra-vidas/>

FUNDING

The author declares that this research received no external funding.

CONFLICT OF INTEREST

The author declares no financial or non-financial relationships relevant to the content of the manuscript.

AUTHOR CONTRIBUTIONS

Conceptualization: Byron Mena.

Methodology: Byron Mena.

Investigation: Byron Mena.

Formal analysis: Byron Mena.

Writing – original draft: Byron Mena.

Writing – review and editing: Byron Mena.

ETHICAL CONSIDERATIONS

As this is a review of the literature and public documents, without collection of individual participant data, neither ethics committee approval nor informed consent was required.

DATA AVAILABILITY

The synthesized information comes from the publications and documents cited. Table 1 presents the characteristics and results used in the narrative synthesis.

USE OF ARTIFICIAL INTELLIGENCE

The author declares that no artificial intelligence was used in the preparation of this manuscript.

ACKNOWLEDGMENTS

The author thanks Dr. Francisco Isaac Mena Acosta, gynecologist and obstetrician, for his contribution and guidance on gynecological clinical concepts during the development of the manuscript.