IIAR JOURNALS

Innovating Knowledge. Advancing Research.

IIAR Journal of Health, Medicine, and Biomedical Research

Vol 1, No 1, 2026

Clinical Decision Support with Large Language Models: A Theoretical Framework for Safe Integration into Resource-Constrained Health Systems

IGWE Emmanuel Sunday[1], NWOGBOJI Emmanuel Usulor[2], ATOLAGBE Umar Kawu[3], OGAH Sunday Ogbo[4], BAMIGBOLA Paul Idowu[5]

[1,3,4,5] Veristzon Global Research Lab; [2] International Institute of AI and Research

ABSTRACT

Large language models (LLMs) can now summarise records, suggest differential diagnoses, support triage and converse with patients, and the health systems with the fewest clinicians face the strongest pressure to adopt them. Yet the evidence on their clinical use comes mostly from high-income settings, and existing guidance treats LLM-based clinical decision support as one undifferentiated technology rather than a set of tasks with different risks. This study aimed to specify where in the clinical workflow an LLM may safely sit in resource-constrained health systems, the failure modes characteristic of each position, and the safeguards each position requires. A theoretical, integrative theory-building design was used. The study population comprised peer-reviewed journal articles published from 2023 to 2026 on LLMs and tool-using agentic systems, clinical decision support, patient safety, governance, implementation and the health workforce in low-resource settings. Purposive, relevance-ranked sampling across eight search strings in PubMed/MEDLINE yielded 265 unique records; 213 were screened against predefined criteria, and 42 were retained, together with five landmark articles added by digital object identifier, giving a final sample of 47 articles. Every included article was verified against its Crossref publisher deposit, and none had been retracted. Data were extracted into a structured matrix and analysed by deductive and inductive thematic synthesis. Only 13 of the 47 articles were set in Africa, and only two reported prospective use of an LLM by African clinicians. The synthesis produced three outputs. The first is a taxonomy of four workflow positions of rising risk: documentation support, triage and referral support, differential diagnosis and management planning, and direct patient communication. The second is a five-domain safeguard architecture whose intensity scales with the risk of the position. The third is six testable propositions and seven design principles for bounded clinical autonomy. The central argument is that risk depends less on model capability than on interception distance, defined as the number of independent human checks between a model's output and the patient. The study recommends that health ministries, regulators and developers choose the workflow position before choosing the model, ground outputs in nationally adapted guidelines, default to referral under uncertainty, and grant greater autonomy only on the evidence of post-deployment monitoring.

Keywords: clinical decision support; large language models; patient safety; human oversight; bounded autonomy

Clinical Decision Support with Large Language Models: A Theoretical Framework for Safe Integration into Resource-Constrained Health Systems

1. INTRODUCTION

Sub-Saharan Africa has fewer medical workers per capita than any other region of the world (Sheffel et al., 2024), yet in many of its countries trained health workers remain unemployed because posts are unfunded (Pieterse, 2025). In much of the region, first-contact decisions are made by clinical officers, nurses, midwives and community health workers who work with limited supervision and intermittent connectivity, and whose guidelines are often outdated or unavailable at the point of care (Cornick et al., 2025; Rath et al., 2025). Large language models (LLMs) appear to offer a way to narrow this gap: they can summarise records, retrieve and explain guidelines, propose differential diagnoses and converse in natural language at a marginal cost far below that of an additional clinician (Chen et al., 2025; Thirunavukarasu et al., 2023). Commentators on the World Health Organization (WHO) guidance on large multimodal models have argued that such tools could support universal health coverage, but only if their ethical and safety risks are addressed during design and deployment (Malpani & Moodley, 2025).

The evidence on capability is substantial. LLMs encode considerable clinical knowledge (Singhal et al., 2023); open-source models perform as well as or better than leading proprietary models on clinical decision tasks (Sandmann et al., 2025); and one recent model outperformed physician baselines across a series of diagnostic-reasoning experiments, including second opinions in an emergency department (Brodeur et al., 2026). Agentic systems, which allow an LLM to call external tools, have improved accuracy further on specialist tasks (Ferber et al., 2025; Zhao et al., 2026). The evidence on safety is equally substantial. On full decision-making tasks drawn from real patient cases, LLMs performed worse than clinicians, did not follow guidelines and were sensitive to the order in which information was presented (Hager et al., 2024). Physicians given access to GPT-4 did not reason significantly better, even though the model alone outscored both study arms (Goh et al., 2024). In maternal and neonatal care scenarios for low- and middle-income countries (LMICs), GPT-4o presented invented statistics without signalling uncertainty in 44% of cases and produced incorrect citations 46.7% of the time (Yu et al., 2026).

The first field evidence from Africa sharpens these questions rather than resolving them. A pragmatic cluster-randomised trial in 16 Kenyan primary care facilities found LLM assistance for clinical officers to be safe but without a statistically significant effect on 14-day treatment failure (Agweyu et al., 2026). A pilot on a neurosurgery ward in the Gambia reported fewer critical errors in medical officers' plans for hypothetical case dilemmas, but it involved only nine users and no control arm (Gupta et al., 2026). The wider literature is also unevenly distributed: in a scoping review of clinician acceptance of artificial intelligence (AI), 70% of studies came from high-income countries and only 20% from LMICs (Scipion et al., 2025).

Existing frameworks address this field at the level of policy and governance, through roadmaps for equitable adoption (Chen et al., 2025), multidimensional governance frameworks (Ferreira & Rosa, 2026) and accountability charters (Mourão & Juliasse, 2026). They do not, however, theorise *where* in the clinical workflow an LLM is placed or how that placement changes risk. Drafting a discharge note and advising a mother about danger signs in her newborn are both forms of LLM-based clinical decision support (LLM-CDS), yet they differ in the harm an error can cause and in the likelihood that a human will detect the error first. Without a position-specific theory, health systems cannot match safeguards to risk and are liable either to over-regulate low-risk uses or to under-protect high-risk ones.

This paper addresses that gap by developing a theoretical framework for the safe integration of LLM-CDS into resource-constrained health systems. It asks three research questions. (RQ1) What distinct workflow positions can an LLM occupy in clinical care, and how do they differ in risk? (RQ2) What failure modes are characteristic of each position in

© IIARJournal. All rights reserved.

resource-constrained settings? (RQ3) What safeguards, and what conditions for extending autonomy, does each position require? The answers take the form of a taxonomy, a safeguard architecture, a set of testable propositions and design principles for bounded clinical autonomy.

2. LITERATURE REVIEW

2.1 CONCEPTUAL REVIEW

Large language models and clinical decision support

Clinical decision support (CDS) comprises tools that provide clinicians with patient-specific information or recommendations at the point of care. Earlier CDS systems in African primary care were rule-based, encoding guideline logic for maternal care or antimicrobial prescribing (Peiffer-Smadja et al., 2024; Tegenaw et al., 2024). LLMs differ in kind: trained on large text corpora to predict language, they produce fluent, context-sensitive output for tasks they were never explicitly programmed to perform (Thirunavukarasu et al., 2023). This flexibility is the source of both their value and their hazard. It has prompted calls to redesign CDS around generative models rather than alerts (Liu et al., 2026), but it also allows general-purpose chatbots to produce output resembling that of a regulated medical device without ever having been authorised as one (Weissman et al., 2025).

Tool-using and agentic systems

A newer class of agentic systems allows an LLM to divide a task into steps and call external tools, such as literature search, guideline and knowledge bases, image-analysis models or data-analysis pipelines, rather than answering from its training data alone (Ferber et al., 2025; Low et al., 2025; Zhao et al., 2026). Tool use alters both the strengths and the risks of CDS. Outputs can be grounded in retrievable, traceable evidence, but every step the system takes without review is a further point at which an error can enter unobserved. Agentic designs therefore bear directly on how much autonomy a model should be granted in the clinical workflow.

Resource-constrained health systems

In this paper, ‘resource-constrained’ denotes concurrent limitations in five areas: infrastructure, workforce, language, knowledge, and regulation and financing. Unreliable electricity and internet access constrain cloud-based tools, especially in rural facilities (Shidende & Mwogosi, 2025). Health workers are too few and unevenly deployed, and care is task-shifted to non-physician cadres (Hussein & Yusuf, 2026; Okoroafor & Dela Christmals, 2024; Rath et al., 2025; Sheffel et al., 2024). Patients and many health workers communicate in languages under-represented in model training data (Kimera et al., 2026; Lima et al., 2025). National guidelines require continual adaptation and curation (Cornick et al., 2025; Muliokela et al., 2025). Data are fragmented, prospective validation is scarce and regulatory capacity is weak (Guan et al., 2026). Together, these constraints alter both the benefit a tool can deliver and the ways in which it can fail.

Patient safety concepts for generative models

Four error concepts recur in this literature. *Hallucination* is fluent output unsupported by the input or by fact, and *omission* is the failure to include clinically material information; both have been quantified in clinical summarisation (Asgari et al., 2025). *Robustness* is the stability of output under realistic variation in input, yet most robustness tests rely on misleading or adversarial prompts, and only a third simulate plausible clinical scenarios (Chang et al., 2026). *Bias* operates in two directions: LLMs can reproduce demographic and cost-related biases in diagnosis (Su et al., 2025), and they can inherit or amplify clinicians’ cognitive biases, although they may also counter some of them (Mahajan et al., 2025).

2.2 THEORETICAL REVIEW

Five theoretical perspectives inform the framework. The **Unified Theory of Acceptance and Use of Technology (UTAUT)** explains clinicians’ uptake of AI through performance expectancy, effort expectancy, social influence and facilitating conditions. A scoping review that mapped barriers and facilitators to UTAUT identified further factors outside its core constructs, namely physician involvement as a facilitator and clinician hesitancy and legal and ethical considerations as barriers, and found little evidence from LMICs (Scipion et al., 2025). Consistent with this, an acceptance study of point-of-care CDS in Ethiopia found that system quality and perceived benefit drove acceptance,

© IIARJournal. All rights reserved.

whereas ease of use had no direct effect (Tegenaw et al., 2024).

The **Consolidated Framework for Implementation Research** locates implementation outcomes in individual, organisational and outer-setting factors. Applied to an antimicrobial CDS in West and Central Africa, it identified barriers at the individual, system and national levels and showed that a tool must suit rural nurses and midwives, not only academic physicians (Peiffer-Smadja et al., 2024).

Cognitive-bias and automation-bias theory holds that users of decision aids may over-rely on them, anchor on their suggestions or overlook their errors. Applied to LLMs, it predicts that model accuracy alone does not determine outcomes, because the human–AI interface mediates whether accurate output is used well (Goh et al., 2024; Mahajan et al., 2025).

Task-shifting theory holds that care can be safely redistributed to less specialised cadres when tasks are well defined, supported by protocols and backed by referral (Rath et al., 2025). It implies that the safety of a decision aid depends on the scope of practice of its user. Across 38 studies in 23 African countries, task-shifting was supported by in-service training, onsite supervision and mentoring, supportive supervision, job aids and preservice education (Okoroafor & Dela Christmals, 2024). In these terms, an LLM-based tool is a new form of job aid that carries the same need for supervision as its predecessors, and AI is increasingly framed as support for task-sharing rather than as a replacement for clinicians (Hussein & Yusuf, 2026).

Finally, **earned delegation** is an accountability concept for LLM-mediated harm. It holds that a clinical task should be delegated to a model only in proportion to the evidence of its performance in a defined use, under a predeployment charter specifying intended and excluded uses, oversight, escalation, audit and subgroup review (Mourão & Juliasse, 2026).

Together, these perspectives explain *whether* a tool is adopted, *how* it is implemented and *how* clinicians interact with it. None, however, explains how risk varies with the position a model occupies in the clinical workflow, or how safeguards should scale with that risk. Earned delegation comes closest but operates at the level of the organisation rather than the task. The framework developed here integrates the five perspectives at the level of the workflow position.

2.3 EMPIRICAL REVIEW

Capability benchmarks

Benchmark studies establish what models can do under controlled conditions. Flan-PaLM reached state-of-the-art accuracy on medical question-answering benchmarks, including 67.6% on MedQA, and the instruction-prompt-tuned Med-PaLM narrowed, but did not close, the gap with clinicians in human evaluation (Singhal et al., 2023). Across 2,000 cases derived from intensive care records, LLMs produced useful diagnoses, specialist suggestions and urgency assessments, and a retrieval-augmented workflow was tested as an add-on (Gaber et al., 2025). Open-source DeepSeek models matched or exceeded proprietary models on 125 cases, a finding of practical importance because open models can be deployed locally in compliance with privacy requirements (Sandmann et al., 2025). A scoping review of 95 diagnostic studies found that GPT-4 was used in 74% of them and reported high accuracy in several specialties, alongside biases by race, gender and cost (Su et al., 2025). On 2,400 real cases, however, benchmarks of autonomous decision-making found models unready for unsupervised use (Hager et al., 2024).

Tool-using and agentic systems

Agentic studies evaluate the model connected to external tools. A GPT-4 agent equipped with precision-oncology tools (histopathology models, image segmentation, OncoKB and literature search) was evaluated on 20 realistic patient cases. It selected appropriate tools with 87.5% accuracy, reached correct clinical conclusions in 91.0% of cases and cited relevant guidelines correctly 75.5% of the time, and its overall decision-making accuracy rose from 30.3% for GPT-4 alone to 87.2% with tools (Ferber et al., 2025). A multi-agent system using more than 40 tools for rare disease diagnosis exceeded the next best method by 23.79% in average Recall@1 (the proportion of cases in which the correct diagnosis is ranked first) on phenotype-based tasks. Experts agreed with 95.4% of its reasoning chains, which were linked to verifiable evidence (Zhao et al., 2026).

© IIARJournal. All rights reserved.

On 50 real-world clinical questions rated by nine physicians, general-purpose LLMs gave relevant, evidence-based answers to only 2–10% of questions, compared with 24% for a retrieval-augmented system and 58% for an agentic system that analysed real-world data (Low et al., 2025). None of these systems, however, was tested in Africa or in primary care, and all were evaluated on cases or questions rather than in prospective use.

Human–AI interaction studies

Studies that keep a clinician in the loop report more mixed results. In a randomised trial with 50 physicians, access to GPT-4 did not improve diagnostic reasoning (Goh et al., 2024). By contrast, across 91 prescribing-error scenarios in 16 specialties, a ‘co-pilot’ mode in which pharmacists worked with an LLM achieved the highest accuracy (61%) and raised the detection of errors with potential for serious harm 1.5-fold compared with pharmacists alone (Ong et al., 2025). In primary diabetes care, assistance from the image-based module of an integrated image and language system raised primary care physicians’ accuracy for referable retinopathy from 81.0% to 92.3% (Li et al., 2024). These results suggest that benefit depends on the design of the interaction and on the knowledge gap between user and model.

Evidence from resource-constrained settings

These findings must be read against the African workforce context. Facility surveys in 10 African countries found that many facilities were barely staffed once absences were taken into account and that large proportions of health workers had very low knowledge of standard maternal and child health conditions. Low caseloads in many facilities further suggested that deployment, and not only numbers, limits care (Sheffel et al., 2024).

Evidence on LLMs in LMICs is recent and sparse. In the Kenyan cluster-randomised trial (9,691 patients; 103 clinical officers), crude treatment failure rates were 2.2% in the LLM-assisted arm and 2.0% in the control arm. After adjustment, the odds ratio was 0.77 (95% CI 0.55–1.08; p = .13), and no serious adverse events were attributed to the LLM (Agweyu et al., 2026). In the Gambian pilot, which involved nine users (four medical officers and five nurses), plan quality improved in 60% of 75 interactions. In hypothetical case dilemmas, the proportion of medical officers’ plans containing critical errors fell from 33.3% to 0%, and no unsafe outputs were recorded (Gupta et al., 2026). A Rwandan protocol will compare LLM-generated referral decisions with expert consensus across 800 community health worker consultations (Menon et al., 2025). Other studies reveal contextual weaknesses. LLM answers on maternal health varied in quality by language and showed gender bias (Lima et al., 2025), and GPT-4o answered 79% of examination questions correctly yet erred on locally prevalent conditions (Yu et al., 2026). In-language resources, such as a 3,694-pair maternal health dataset in English, Luganda, Runyankore and Swahili, are only beginning to appear (Kimera et al., 2026).

Perception studies identify the main barriers to use. In Tanzania, health professionals were more accepting than patients, and unreliable internet and electricity, privacy and bias were the principal obstacles (Shidende & Mwogosi, 2025). In South Africa, experts raised data security, reliability and over-reliance as concerns (Chuma & Sibiya, 2026). Reviews of AI for diabetes care and for primary care equity in LMICs report promise offset by infrastructure gaps, unrepresentative data, digital exclusion and weak regulation (Guan et al., 2026; Osonuga et al., 2025).

Safety engineering, evaluation and governance

A growing literature builds safety into LLM workflows. A clinical safety framework for summarisation found a hallucination rate of 1.47% and an omission rate of 3.45%, and reduced major errors below reported human rates through prompt and workflow refinement (Asgari et al., 2025). An uncertainty-calibrated, retrieval-augmented pipeline flagged 18% of cases as low-certainty for expert review and produced 67% fewer unsafe antibiotic suggestions than GPT-5 (Hasan et al., 2026). With regard to evaluation, a systematic review of 137 chatbot studies found that almost none identified the model version, settings or prompt engineering, and that most used subjective success criteria (Huo et al., 2025). This gap motivated the CHART reporting guideline, developed through a Delphi process with 531 stakeholders (CHART Collaborative et al., 2025). With regard to governance, Meskó and Topol (2023) argued that LLMs require dedicated

© IIARJournal. All rights reserved.

regulatory oversight, possibly through a new regulatory category, because they differ from previously regulated AI tools. Others have proposed multidimensional governance frameworks (Ferreira & Rosa, 2026) and specified the competencies clinicians need (Cao et al., 2026).

Synthesis and gap

The empirical literature supports four conclusions. First, capability no longer appears to be the binding constraint, and tool use extends it further. Second, integration with human users determines whether capability translates into benefit. Third, errors in resource-constrained settings frequently arise from contextual mismatch. Fourth, safety can be engineered through grounding, traceable tool use, calibrated deferral and monitoring. What is missing is a framework that links these conclusions to the specific clinical tasks LLMs perform, limits the autonomy granted to agentic systems, and is grounded in the workforce and service conditions of African primary care.

3. MATERIALS AND METHODS

3.1 RESEARCH DESIGN

This study used a theoretical, non-empirical design based on integrative theory-building synthesis, which develops new constructs and propositions by systematically combining theoretical and empirical literature rather than by estimating pooled effects. The design suits a young field whose evidence is heterogeneous in design and outcome, and research questions that concern the structure of risk rather than the size of an effect. No human participants were involved, and ethical approval was therefore not required.

3.2 POPULATION, SAMPLE AND SAMPLING TECHNIQUE

The study population comprised peer-reviewed journal articles published between 1 January 2023 and 20 October 2026 on LLMs, tool-using agents or generative AI in clinical decision-making, patient safety, governance or implementation in low-resource settings, together with articles on the health workforce of African health systems. Sampling was purposive and relevance-ranked. Eight search strings were run in PubMed/MEDLINE on 20 October 2026, covering LMIC context, LLM-CDS performance, safety and hallucination, governance and ethics, implementation and adoption, LLM use in African and other LMIC settings, tool-using agents, and the African health workforce. The 25–40 most relevant records for each string were screened, and one further agent study, counted among the 42 retained records, was located by author search. Five widely cited landmark articles from 2023–2024 were added by digital object identifier (DOI). The full search strings, hit counts and screening flow are reported in the supplementary search log.

3.3 INCLUSION AND EXCLUSION CRITERIA

Articles were included if they were peer-reviewed journal articles published from 2023 to 2026 that addressed LLMs, generative AI or related CDS and at least one of the following: clinical performance, safety, human–AI interaction, governance, implementation or resource-constrained contexts. Pre-LLM CDS studies from African settings were included when they informed implementation in those settings, as were workforce studies without an AI component that described the African staffing context. Editorials, letters, news items, preprints, conference papers, unverifiable articles and studies outside health care were excluded.

3.4 DATA COLLECTION AND VERIFICATION

Of 270 records retrieved, 265 remained after de-duplication, and 213 returned full PubMed metadata and were screened; 171 were excluded and 42 included. With the five landmark articles, the final sample comprised 47 articles. Each DOI was checked against its Crossref publisher record for title, authors, journal, volume, issue, pagination or article number, dates and update notices. All fields matched for 44 articles. For the remaining three, the Crossref deposit was incomplete or erroneous (a non-standard volume field, a first page only, and a transposed volume and issue with incorrectly split author names); these were corrected against PubMed and the publisher's page, and each correction was documented. For one article, Crossref records the online-first year (2023), and the year of the print issue (2024) is cited, in accordance with APA 7. No retractions or expressions of concern were found. One article carries a publisher correction issued in July 2023 (Singhal et al., 2023), and the corrected version of record is cited.

© IIARJournal. All rights reserved.

3.5 DATA ANALYSIS AND BASES OF JUDGEMENT

Data were extracted into a 20-field Microsoft Excel matrix recording setting, design, aim, key findings, limitations and relevance to the framework, and bibliographic verification used the Crossref REST API and PubMed. Analysis proceeded in three stages. First, deductive coding assigned each article to provisional safeguard domains derived from the theoretical review. Second, inductive coding of reported failure modes generated the workflow-position taxonomy, in which each position was defined by the clinical task, the user and the number of independent human checks before the output reaches the patient. Third, relationships between positions, failure modes and safeguards were formulated as propositions. A proposition was retained only if it was supported by at least two independent articles, at least one of them empirical, and contrary evidence was reported where it existed.

3.6 RELIABILITY AND VALIDITY

Reliability was supported by a documented search log, verification of every reference against the publisher record, and a complete extraction matrix that allows the synthesis to be audited and repeated. Construct validity was supported by triangulation across five theoretical perspectives and across designs ranging from randomised trials to reviews. External validity was supported by giving priority to evidence from African and LMIC settings and by stating boundary conditions. The principal threats are that the search covered a single database and that screening was limited to the most relevant records for each string; both may have led to the omission of relevant studies, particularly from regional journals.

3.7 CONSTRUCTS AND THEIR MEASUREMENT

The framework uses four constructs. *Workflow position* is the clinical task the model performs and the user who receives its output; it is a categorical construct with four levels. *Interception distance* is the number of independent, competent human checks between the model's output and an action affecting the patient. It ranges from one or more (for example, author sign-off and countersignature) to zero (direct output to the patient). *Safeguard intensity* is the strength of the controls applied within each of five domains, graded as baseline, enhanced, stringent or maximal. *Safe assistance*, the outcome construct, denotes net clinical benefit without an increase in serious harm, assessed through process quality, critical-error rates, appropriate referral and patient outcomes.

3.8 DISCLOSURE OF ARTIFICIAL INTELLIGENCE USE

The authors used a generative AI assistant (Claude, Anthropic) to support database searching, Crossref metadata checks and language drafting. All sources were retrieved from bibliographic databases and verified against publisher records. The authors reviewed and edited all content, checked every citation against its source, and take full responsibility for the manuscript. No AI tool is listed as an author.

3.9 DATA AVAILABILITY

The literature matrix, search strings and screening log are provided as a supplementary file. No new patient data were generated.

4. RESULTS AND DISCUSSION

4.1 RESULTS

4.1.1 Characteristics of the evidence base

The 47 articles were published in 2023 (n = 3), 2024 (n = 7), 2025 (n = 22) and 2026 (n = 15), reflecting the recency of the field. Table 1 summarises their design and setting. Only ten reported primary empirical data from clinicians, health workers or patients. Thirteen were set in Africa, of which only two reported prospective use of an LLM by African clinicians (Agweyu et al., 2026; Gupta et al., 2026). More than half (24 of 47) were set in high-income countries or had no specific setting.

Table. 1. *Characteristics of the included articles (N = 47)*

Dimension	Category	n	%
Design	Primary empirical study with health workers or patients	10	21
	Model or agent evaluation, benchmark or dataset	14	30
	Systematic, scoping or narrative review	10	21
	Perspective, viewpoint or policy analysis	9	19
	Protocol, guideline or implementation report	4	9
Setting	African country or countries	13	28

© IIARJournal. All rights reserved.

Dimension	Category	n	%
	Other LMIC or LMIC-focused	10	21
	High-income or unspecified	24	51

Note. Percentages are rounded and may not sum to 100. LMIC = low- and middle-income country.

4.1.2 A taxonomy of workflow positions

The synthesis identified four workflow positions, ordered by rising risk (Table 2). Risk depends on two properties. The first is the directness with which an error translates into harm to the patient. The second is interception distance, together with the quality of the check it provides: at WP1 the check is made by the clinician who knows the encounter, at WP2 and WP3 by a single user who may lack the knowledge to detect the error, and at WP4 there is no check at all. WP2 and WP3 share an interception distance of one but differ in directness of harm, because a management plan acts on the patient, whereas a triage decision changes where care takes place.

In **documentation support (WP1)**, errors are visible to the signing clinician and can be intercepted; hallucination and omission rates are low and can be reduced further by workflow refinement (Asgari et al., 2025). In **triage and referral support (WP2)**, the characteristic failure is under-triage: false reassurance delays referral, and the error is invisible when made, because a patient who is not referred does not return to reveal it. In **differential diagnosis and management planning (WP3)**, the characteristic failure is a plausible but wrong plan arising from guideline non-adherence, sensitivity to information order, errors on locally prevalent conditions or anchoring by the user (Hager et al., 2024; Mahajan et al., 2025; Yu et al., 2026). In **direct patient communication (WP4)**, interception distance falls to zero, so invented statistics, language-dependent quality and device-like advice reach the patient with no clinician in between (Lima et al., 2025; Weissman et al., 2025; Yu et al., 2026).

Agentic designs cut across the four positions. They can raise accuracy and make reasoning traceable at WP3 (Ferber et al., 2025; Zhao et al., 2026), but each tool call an agent makes on its own is a step that no human sees. The framework therefore counts only human checks in interception distance, however many automated steps an agent takes, and applies to the whole agent the safeguards of the position that its final output feeds.

Table. 2. *Workflow-position taxonomy for LLM clinical decision support in resource-constrained settings*

Workflow position	Typical task and user	Characteristic failure modes	Interception distance	Minimum oversight
WP1 Documentation support	Drafting notes, referral letters and handovers for the treating clinician	Hallucinated or omitted facts entering the record (Asgari et al., 2025)	One or more (author sign-off; countersignature where required)	Clinician sign-off before filing; periodic sample audit
WP2 Triage and referral support	Urgency and refer-or-manage decisions by nurses and community health workers	Under-triage and false reassurance; errors invisible when made	One (the user), often unsupervised	National danger-sign rules that override the model; default to referral
WP3 Differential diagnosis and management planning	Diagnosis, investigation and treatment plans by clinical and medical officers	Plausible but wrong plans; guideline non-adherence; order sensitivity; local-disease errors; automation bias (Hager et al., 2024; Mahajan et al., 2025; Yu et al., 2026)	One (the user), whose expertise may be below the model's	Co-pilot mode; guideline grounding; calibrated deferral; cadre-scoped outputs; regulation as medical-purpose software
WP4 Direct patient communication	Health advice to patients and caregivers in their own language	Unsafe or invented advice; language-dependent quality; unregulated device-like output (Lima et al., 2025; Weissman et al., 2025)	Zero	Closed content scope; escalation to a human on any red flag; regulation as medical-purpose software

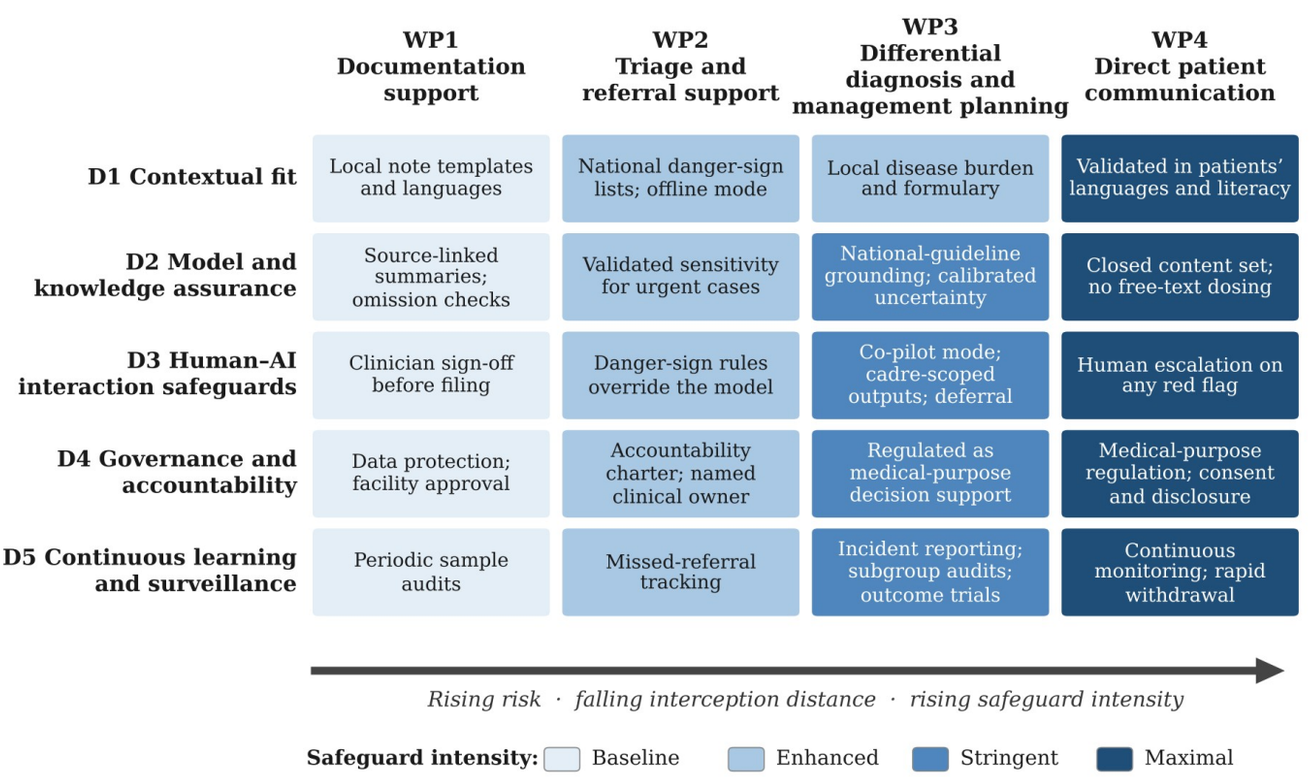
Note. Interception distance is the number of independent, competent human checks between model output and an action affecting the patient. WP = workflow position.

© IIARJournal. All rights reserved.

4.1.3 A position-calibrated safeguard architecture

The coding produced five safeguard domains, each supported by multiple articles. **Contextual fit (D1)** covers alignment with the local disease burden, languages, infrastructure and user cadre (21 articles). **Model and knowledge assurance (D2)** covers grounding in nationally adapted guidelines, local validation, calibrated uncertainty and transparent reporting of model versions (19 articles). **Human–AI interaction safeguards (D3)** cover co-pilot design, scope limited by cadre, escalation triggers and user competencies (22 articles). **Governance and accountability (D4)** covers regulatory classification, accountability charters, data protection and procurement (14 articles). **Continuous learning and surveillance (D5)** covers post-deployment monitoring, incident reporting, subgroup audits and outcome trials (9 articles). The central claim of the architecture is that all five domains apply at every position, but that their required intensity scales with the risk of the position (Figure 1).

Figure. 1. *The position-calibrated safeguard architecture: minimum safeguards by domain and workflow position*



Note. Shading shows the required safeguard intensity (baseline, enhanced, stringent, maximal). Risk and safeguard intensity rise from left to right as directness of harm rises and interception distance falls.

4.1.4 Propositions

Six testable propositions relate workflow position, interception distance and safeguard intensity to safe assistance, the outcome construct defined in Section 3.7 (Table 3).

P1 (interception). For a given model, the safeguards required for safe assistance increase as interception distance decreases. This would explain why documentation errors that are tolerable when reviewed (Asgari et al., 2025) become unacceptable when similar output reaches patients directly (Weissman et al., 2025; Yu et al., 2026).

P2 (complementarity). LLM assistance improves decisions most when its output is reviewed rather than adopted, and when users have a meaningful knowledge gap relative to the model but sufficient training to act on its output. Co-pilot designs outperformed both human-alone and model-alone

© IIARJournal. All rights reserved.

arms (Ong et al., 2025), and assistance improved the accuracy of primary care physicians (Li et al., 2024). In the Gambian pilot, however, medical officers gained more than nurses (Gupta et al., 2026), suggesting that a minimum level of competence is needed to use assistance well; where integration is poorly designed, access alone does not help (Goh et al., 2024). In African primary care, low clinical knowledge among many health workers makes such a knowledge gap common (Sheffel et al., 2024), whereas the competence to act on assistance depends on the training and supervision that already support task-shifting (Okoroafor & Dela Christmals, 2024).

P3 (local grounding). In resource-constrained settings, contextual mismatch in disease burden, guidelines and language is a principal source of error, and grounding outputs in nationally adapted guidelines reduces this error more than gains in general model capability do (Cornick et al., 2025; Lima et al., 2025; Muliokela et al., 2025; Yu et al., 2026). The marked improvement of an oncology agent once the same model was connected to tools is consistent with this proposition (Ferber et al., 2025).

P4 (calibrated deferral). Systems that express calibrated uncertainty and defer low-certainty cases to human review produce fewer unsafe outputs without a proportionate loss of usefulness (Hasan et al., 2026; Mahajan et al., 2025).

P5 (outcome sensitivity). Where baseline rates of adverse outcomes are low, safety can be demonstrated before benefit, and detecting benefit requires measures sensitive to the quality of the care process, such as critical-error rates and appropriate referral (Agweyu et al., 2026; Gupta et al., 2026).

P6 (earned delegation). Accountability arrangements specified before deployment are a precondition for safe scale-up beyond pilots, and their absence predicts unsafe drift or abandonment (Guan et al., 2026; Mourão & Juliasse, 2026; Peiffer-Smadja et al., 2024; Weissman et al., 2025).

Table. 3. *Propositions, supporting evidence and how each could be tested*

Proposition	Statement (abridged)	Key supporting evidence (contrary evidence where found)	Illustrative test
P1 Interception	Required safeguards rise as interception distance falls	Asgari et al. (2025); Weissman et al. (2025); Yu et al. (2026)	Compare how often critical errors reach patients across WP1–WP4 for the same model
P2 Complementarity	Benefit needs reviewed output and a user gap the user is trained to close	Li et al. (2024); Ong et al. (2025); mixed: Gupta et al. (2026); contrary: Goh et al. (2024)	Factorial trial of cadre × interaction design (co-pilot vs. advisory)
P3 Local grounding	Local guideline grounding reduces contextual error more than capability gains	Cornick et al. (2025); Ferber et al. (2025); Lima et al. (2025); Muliokela et al. (2025); Yu et al. (2026)	Grounded vs. ungrounded models on local-disease vignettes in local languages
P4 Calibrated deferral	Calibrated deferral lowers unsafe output with little loss of usefulness	Hasan et al. (2026); Mahajan et al. (2025)	Deferral rate vs. unsafe-output rate in field deployment
P5 Outcome sensitivity	Safety precedes demonstrable benefit where event rates are low	Agweyu et al. (2026); Gupta et al. (2026)	Trials with process-quality co-primary outcomes
P6 Earned delegation	Predeployment accountability is a precondition for scale-up	Guan et al. (2026); Mourão & Juliasse (2026); Peiffer-Smadja et al. (2024); Weissman et al. (2025)	Prospective comparison of deployments with and without charters

4.1.5 Design principles for bounded clinical autonomy

The propositions yield seven design principles. (1) *Position before model*: choose and justify the workflow position before selecting a model. (2) *Never the last barrier*: at WP2 and WP3, the model is never the final check, and hard-coded danger-sign rules override it; at WP4, any red flag escalates to a human (Menon et al., 2025; Rath et al., 2025). (3) *Ground locally and cite visibly*: link outputs to nationally adapted guideline content that the user can inspect, and, where an agent retrieves evidence through tools, show its sources and reasoning chain (Cornick et al., 2025; Muliokela et al., 2025; Zhao et al., 2026). (4) *Fail safe towards referral*: under uncertainty, the default action is escalation (Hasan et al., 2026). (5) *Scope by cadre*: limit outputs to actions

© IIARJournal. All rights reserved.

the user is trained and permitted to take (Peiffer-Smadja et al., 2024). (6) *Earn autonomy*: move a tool to a higher-risk position only after monitored performance at the lower one (Mourão & Juliasse, 2026), and extend autonomy only incrementally to agentic systems that act in several unreviewed steps. (7) *Design for the constraint*: prefer models that can run offline or locally, and interfaces that work on low bandwidth and in local languages (Chen et al., 2025; Kimera et al., 2026; Sandmann et al., 2025).

4.2 DISCUSSION

In answer to RQ1, the synthesis suggests that LLM-CDS is better understood as four workflow positions than as a single technology, and that the position, more than the model, determines risk. This reading reframes existing findings. The null primary result of the Kenyan trial (Agweyu et al., 2026) and the process gains in the Gambian pilot (Gupta et al., 2026) become consistent once outcome sensitivity is considered (P5). Both studies placed the model at WP3 with a clinician in the loop, and neither reported serious model-attributed harm, which is the pattern the framework predicts where interception distance is one and the user's knowledge gap is meaningful. Likewise, the failure of physicians to benefit from GPT-4 (Goh et al., 2024), contrasted with the co-pilot gains among pharmacists (Ong et al., 2025), supports the view that integration design is a safety variable in its own right (P2).

In answer to RQ2, the dominant failure modes in resource-constrained settings are contextual rather than generic. Errors on locally prevalent conditions, uneven quality across languages and the need for continual guideline curation (Cornick et al., 2025; Lima et al., 2025; Yu et al., 2026) imply that a model validated in a high-income setting does not carry its safety record with it. This extends acceptance theory: whereas UTAUT and related models treat system quality and facilitating conditions as drivers of uptake (Scipion et al., 2025; Tegenaw et al., 2024), the framework treats contextual fit as a driver of *safety*. A tool that is adopted enthusiastically but fits its context poorly is therefore a particularly hazardous combination.

In answer to RQ3, safeguards should be calibrated to workflow position rather than applied uniformly. This refines existing governance proposals: equity roadmaps (Chen et al., 2025), multidimensional governance frameworks (Ferreira & Rosa, 2026) and the WHO guidance discussed by Malpani and Moodley (2025) specify *what* must be governed, whereas the present framework specifies *how much* control each use requires. The framework gives the principle of earned delegation (Mourão & Juliasse, 2026) a concrete path, in which a tool earns movement from WP1 towards WP3 through monitored performance. It also places direct patient communication (WP4) last, to be opened only under regulation suited to LLMs directed towards medical purposes (Meskó & Topol, 2023; Weissman et al., 2025). The framework further cautions against substitution: where trained health workers are unemployed because posts are unfunded (Pieterse, 2025), deploying an LLM instead of hiring staff would remove the human checks on which its safety depends. For evaluation, the framework complements the CHART guideline (CHART Collaborative et al., 2025); reports should state the workflow position and interception distance alongside the model version and prompts, so that findings can be compared across studies.

The framework also has practical application. Consider a primary health care centre in Nigeria, where a community health extension worker assesses a febrile child. A WP2 deployment would run offline on a mobile phone and draw on nationally adapted guideline content of the kind maintained for Nigerian primary care (Cornick et al., 2025). National danger-sign rules would override the model, referral would be the default under uncertainty, outputs would be limited to the worker's scope of practice, and the local government health authority would track missed referrals. The same model would not advise the child's mother directly (WP4) until it had been validated locally in Hausa, Yoruba, Igbo and Nigerian Pidgin and had received regulatory approval. Facility managers, regulators and developers can use the framework to scope a tool in this way.

Four limitations apply. First, the framework is theoretical, and its propositions are untested. Second, the search limits set out in Section 3.6 may have excluded studies in regional journals. Third, only two prospective African field studies were available, and none of the agentic systems was evaluated in Africa,

© IIARJournal. All rights reserved.

so the framework draws heavily on high-income evidence interpreted for low-resource settings. Fourth, model capability is changing rapidly, although the framework's emphasis on position rather than model is intended to make it durable. Future research should test P1 to P6 empirically. Priorities include factorial trials of user cadre and interaction design, comparisons of grounded and ungrounded models in African languages, field evaluations of agentic systems in African primary care, and prospective studies of predeployment accountability charters, particularly in Nigeria and other West African health systems, where no LLM-CDS field evidence was identified.

5. CONCLUSION AND RECOMMENDATION

LLMs can strengthen clinical decision-making in health systems with too few clinicians, but only if they are placed where their errors will be caught. This paper has argued that the safety of an LLM depends less on what the model can do than on where it sits in the clinical workflow and on how many competent human checks lie between its output and the patient. To operationalise this argument, it has proposed a taxonomy of four workflow positions, a five-domain safeguard architecture calibrated to their risk, six testable propositions and seven design principles for bounded clinical autonomy.

Five recommendations follow. (1) Ministries of health and regulators should classify LLM-CDS by workflow position, require the position to be chosen before the model, and regulate WP3 and WP4 uses as medical-purpose software. (2) Developers and implementers should ground outputs in nationally adapted guidelines, support local languages and offline use, and build in calibrated deferral and default referral. (3) Facility managers should deploy LLMs first at WP1, then at WP2 only with sign-off, override rules and missed-referral tracking, and extend their autonomy only on the evidence of monitoring. (4) Training institutions should build clinicians' competencies in evaluating and supervising AI outputs. (5) Researchers and funders should prioritise prospective trials in African primary care that use outcomes sensitive to process quality and are reported according to CHART, with the workflow position stated.

6. CONTRIBUTION TO EXISTING KNOWLEDGE

The paper makes five contributions. Theoretically, it introduces *interception distance* and *workflow position* as constructs that explain variation in the risk of LLM-CDS, and it integrates acceptance, implementation, cognitive-bias, task-shifting and accountability theories at the level of the clinical task. Conceptually, it offers, to our knowledge, the first workflow-position taxonomy and position-calibrated safeguard architecture designed for resource-constrained health systems. For research, it sets out six testable propositions with suggested study designs, providing the field with a research agenda. Practically, it gives health ministries, regulators, developers and facility managers in Africa and other LMICs design principles with which to scope, procure and monitor LLM tools. Methodologically, it demonstrates a citation standard in which every source is verified against its publisher record.

7. REFERENCES

Agweyu, A., Mwaniki, P., Menon, V., Korom, R., Isaaka, L., Wanyama, C., Gill, J., Kiptinness, S., Adan, N., Emmanuel-Fabula, M., Riley, R. D., Archer, L., Denniston, A. K., Liu, X., & Mateen, B. A. (2026). Generative AI-enabled clinical decision support system in primary care: A pragmatic, cluster-randomized trial. *Nature Medicine, 32*(8), 3032–3039. https://doi.org/10.1038/s41591-026-04503-6

Asgari, E., Montaña-Brown, N., Dubois, M., Khalil, S., Balloch, J., Yeung, J. A., & Pimenta, D. (2025). A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *npj Digital Medicine*, *8*(1), Article 274. https://doi.org/10.1038/s41746-025-01670-7

Brodeur, P. G., Buckley, T. A., Kanjee, Z., Goh, E., Ling, E. B., Jain, P., Cabral, S., Abdulnour, R. E., Haimovich, A. D., Freed, J. A., Olson, A., Morgan, D. J., Hom, J., Gallo, R., McCoy, L. G., Mombini, H., Lucas, C., Fotoohi, M., Gwiazdon, M., . . . Rodman, A. (2026). Performance of a large language model on the reasoning tasks of a physician. *Science, 392*(6797), 524–527. https://doi.org/10.1126/science.adz4433

Cao, W., Zhang, Q., Liu, J., & Liu, S. (2026). From agents to governance: Essential AI skills for clinicians in the large language model era. *Journal of Medical Internet Research, 28*, Article e86550. https://doi.org/10.2196/86550

Chang, Y., Hsieh, M. H., Ju, P. C., Liu, Y. C., & Chang, C. C. (2026). Clinical plausibility in large language model robustness testing for medicine: A scoping review.

© IIARJournal. All rights reserved.

Journal of Medical Systems, *50*(1), Article 77. https://doi.org/10.1007/s10916-026-02405-1

CHART Collaborative, Huo, B., Collins, G. S., Chartash, D., Thirunavukarasu, A. J., Flanagin, A., Iorio, A., Cacciamani, G., Chen, X., Liu, N., Mathur, P., Chan, A. W., Laine, C., Pacella, D., Berkwits, M., Antoniou, S. A., Camaradou, J. C., Canfield, C., Mittelman, M., . . . Guyatt, G. H. (2025). Reporting guideline for chatbot health advice studies: The CHART statement. *JAMA Network Open*, *8*(8), Article e2530220. https://doi.org/10.1001/jamanetworkopen.2025.30220

Chen, H., Zeng, D., Qin, Y., Fan, Z., Ng Yu Ci, F., Klonoff, D. C., Ji, J. S., Zhang, S., Amissah-Arthur, K. N., Jiménez de Tavárez, M. M., Masood, S., Van Le, P., Keane, P. A., Sheng, B., Wong, T. Y., & Tham, Y. C. (2025). Large language models and global health equity: A roadmap for equitable adoption in LMICs. *The Lancet Regional Health – Western Pacific*, *63*, Article 101707. https://doi.org/10.1016/j.lanwpc.2025.101707

Chuma, K. G., & Sibiya, P. T. (2026). Integrating ChatGPT to support children and adolescents with diabetes mellitus in South African healthcare: Perspectives of healthcare professionals and academics. *African Journal of Primary Health Care & Family Medicine*, *18*(1), Article a5433. https://doi.org/10.4102/phcfm.v18i1.5433

Cornick, R. V., Picken, S. C., Awotiwon, A., Rabe, M., Wattrus, C., Fredericks, T., Timmerman, V., & Fairall, L. R. (2025). Addressing health information inequities: Making evidence-based clinical content more accessible in low- and middle-income primary care. *BMJ Global Health*, *9*(Suppl 3), Article e013814. https://doi.org/10.1136/bmjgh-2023-013814

Ferber, D., El Nahhas, O. S. M., Wölflein, G., Wiest, I. C., Clusmann, J., Leßmann, M.-E., Foersch, S., Lammert, J., Tschochohei, M., Jäger, D., Salto-Tellez, M., Schultz, N., Truhn, D., & Kather, J. N. (2025). Development and validation of an autonomous artificial intelligence agent for clinical decision-making in oncology. *Nature Cancer*, *6*(8), 1337–1349. https://doi.org/10.1038/s43018-025-00991-6

Ferreira, J. C., & Rosa, I. (2026). Large language models in healthcare: Applications, evaluation frameworks, and governance pathways: A scoping review and multidimensional framework. *Frontiers in Digital Health*, *8*, Article 1865568. https://doi.org/10.3389/fdgth.2026.1865568

Gaber, F., Shaik, M., Allega, F., Bilecz, A. J., Busch, F., Goon, K., Franke, V., & Akalin, A. (2025). Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis. *npj Digital Medicine*, *8*(1), Article 263. https://doi.org/10.1038/s41746-025-01684-1

Goh, E., Gallo, R., Hom, J., Strong, E., Weng, Y., Kerman, H., Cool, J. A., Kanjee, Z., Parsons, A. S., Ahuja, N., Horvitz, E., Yang, D., Milstein, A., Olson, A. P. J., Rodman, A., & Chen, J. H. (2024). Large language model influence on diagnostic reasoning: A randomized clinical trial. *JAMA Network Open*, *7*(10), Article e2440969. https://doi.org/10.1001/jamanetworkopen.2024.40969

Guan, Z., Li, H., Hernández-Jiménez, S., Amissah-Arthur, K. N., Schmidt, M. I., Masood, S., Zeng, D., Jia, W., Lim, L. L., & Sheng, B. (2026). Can artificial intelligence bridge the gaps for primary diabetes care in low-income and middle-income countries? *The Lancet Diabetes & Endocrinology*, *14*(5), 412–421. https://doi.org/10.1016/S2213-8587(26)00010-0

Gupta, S., Gal, Z. T., Touray, J., Luiselli, G. A., Ceesay, A., Manneh, E. K., Cham, M., Rolston, J. D., Chrenek, R., Bah, M. G., Golby, A. J., Esene, I. N., Arnaout, O., Sanchez, C., Janneh, L., Smith, T. R., & Jabang, J. N. (2026). Large language model for postoperative clinical decision support in a neurosurgery ward in the Gambia: A prospective pilot feasibility study. *Neurosurgery*, *99*(5), 1127–1135. https://doi.org/10.1227/neu.0000000000003963

Hager, P., Jungmann, F., Holland, R., Bhagat, K., Hubrecht, I., Knauer, M., Vielhauer, J., Makowski, M., Braren, R., Kaissis, G., & Rueckert, D. (2024). Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine*, *30*(9), 2613–2622. https://doi.org/10.1038/s41591-024-03097-1

Hasan, M. M., Hossain, M. A., Sayem, F. H., Paul, B. K., Rahman, Z., Uddin, M. S., & Mostafiz, R. (2026). CLIN-LLM: A safety-constrained hybrid framework for clinical diagnosis and treatment generation. *PLOS ONE*, *21*(8), Article e0348611. https://doi.org/10.1371/journal.pone.0348611

Huo, B., Boyle, A., Marfo, N., Tangamornsuksan, W., Steen, J. P., McKechnie, T., Lee, Y., Mayol, J., Antoniou, S. A., Thirunavukarasu, A. J., Sanger, S., Ramji, K., & Guyatt, G. (2025). Large language models for chatbot health advice studies: A systematic review. *JAMA Network Open*, *8*(2), Article e2457879. https://doi.org/10.1001/jamanetworkopen.2024.57879

Hussein, A. M., & Yusuf, F. Y. (2026). Artificial intelligence and mental health services in Somalia: Opportunities and health-system considerations. *Journal of Multidisciplinary Healthcare*, *19*, 1–4. https://doi.org/10.2147/JMDH.S622655

Kimera, R., Kuyeso, R., Maleka, E., Tukamushaba, F., William, W., Mwavu, R., Bamutura, D. S., Ngonzi, J., Ainomugisha, B., Namuli, A., Namiiro, P., Nyakato, C., Celi, L. A., & Kaggwa, F. (2026). Dialogues of delivery: A multilingual question-answer dataset for maternal healthcare in East African languages. *BMC Research Notes*, *19*(1), Article 317. https://doi.org/10.1186/s13104-026-07899-z

Li, J., Guan, Z., Wang, J., Cheung, C. Y., Zheng, Y., Lim, L. L., Lim, C. C., Ruamviboonsuk, P., Raman, R., Corsino, L., Echouffo-Tcheugui, J. B., Luk, A. O. Y., Chen, L. J., Sun, X., Hamzah, H., Wu, Q., Wang, X., Liu, R., Wang, Y. X., . . . Wong, T. Y. (2024). Integrated image-based

© IIARJournal. All rights reserved.

deep learning and language models for primary diabetes care. *Nature Medicine, 30*(10), 2886–2896. https://doi.org/10.1038/s41591-024-03139-8

Lima, H. A., Trocoli-Couto, P. H. F. S., Moazzam, Z., Rocha, L. C. D., Pagano, A., Martins, F. F., Brabo, L. T., Reis, Z. S. N., Keder, L., Begum, A., Mamede, M. H., Pawlik, T. M., & Resende, V. (2025). Quality assessment of large language models' output in maternal health. *Scientific Reports, 15*(1), Article 22474. https://doi.org/10.1038/s41598-025-03501-x

Liu, S., Peterson, J. F., & Wright, A. (2026). How generative AI should transform clinical decision support. *JAMA, 336*(7), 543–546. https://doi.org/10.1001/jama.2026.13187

Low, Y. S., Jackson, M. L., Hyde, R. J., Brown, R. E., Sanghavi, N. M., Baldwin, J. D., Pike, C. W., Muralidharan, J., Hui, G., Alexander, N., Hassan, H., Nene, R. V., Pike, M., Pokrzywa, C. J., Vedak, S., Yan, A. P., Yao, D. H., Zipursky, A. R., Dinh, C., . . . Gombar, S. (2025). Answering real-world clinical questions using large language model, retrieval-augmented generation, and agentic systems. *Digital Health, 11*, Article 20552076251348850. https://doi.org/10.1177/20552076251348850

Mahajan, A., Obermeyer, Z., Daneshjou, R., Lester, J., & Powell, D. (2025). Cognitive bias in clinical large language models. *npj Digital Medicine, 8*(1), Article 428. https://doi.org/10.1038/s41746-025-01790-0

Malpani, R., & Moodley, K. (2025). Large multimodal models: Boon or burden for low- and middle-income countries. *International Health, 17*(5), 853–855. https://doi.org/10.1093/inthealth/ihae073

Menon, V., Shimelash, N., Rutunda, S., Nshimiyimana, C., Archer, L., Emmanuel-Fabula, M., Berhe, D. F., Gill, J., Hezagira, E., Remera, E., Riley, R., Wong, R., Denniston, A. K., Mateen, B. A., & Liu, X. (2025). Assessing the potential utility of large language models for assisting community health workers: Protocol for a prospective, observational study in Rwanda. *BMJ Open, 15*(10), Article e110927. https://doi.org/10.1136/bmjopen-2025-110927

Meskó, B., & Topol, E. J. (2023). The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *npj Digital Medicine, 6*(1), Article 120. https://doi.org/10.1038/s41746-023-00873-0

Mourão, C. F., & Juliasse, L. E. (2026). Accountability for large language models in health care. *Bulletin of the World Health Organization, 104*(10), 698–704. https://doi.org/10.2471/BLT.26.296186

Muliokela, R. K., Banda, K., Hussen, A. M., Malumo, S. B., Kashoka, A., Mwiche, A., Chiboma, I., Barreix, M., Nyirenda, M., Sithole, Z., Ratanaprayul, N., Endehabtu, B. F., Telake, H. A., Weldeab, A., Probert, W. J. M., Tunçalp, Ö., Maya, E., Woldetsadik, M., Tilahun, B., . . . Tamrat, T. (2025). Implementation of WHO SMART guidelines-digital adaptation kits in pathfinder countries in Africa: Processes and early lessons learned. *JMIR Medical Informatics, 13*, Article e58858. https://doi.org/10.2196/58858

Okoroafor, S. C., & Dela Christmals, C. (2024). Health professions education strategies for enhancing capacity for task-shifting and task-sharing implementation in Africa: A scoping review. *Journal of Continuing Education in the Health Professions, 44*(3), 194–202. https://doi.org/10.1097/CEH.0000000000000517

Ong, J. C. L., Jin, L., Elangovan, K., Lim, G. Y. S., Lim, D. Y. Z., Sng, G. G. R., Ke, Y. H., Tung, J. Y. M., Zhong, R. J., Koh, C. M. Y., Lee, K. Z. H., Chen, X., Ch'ng, J. K., Than, A., Goh, K. J., Lim, C. P., Ng, T. M., Liu, N., & Ting, D. S. W. (2025). Large language model as clinical decision support system augments medication safety in 16 clinical specialties. *Cell Reports Medicine, 6*(10), Article 102323. https://doi.org/10.1016/j.xcrm.2025.102323

Osonuga, A., Osonuga, A. A., Fidelis, S. C., Osonuga, G. C., Juckes, J., & Olawade, D. B. (2025). Bridging the digital divide: Artificial intelligence as a catalyst for health equity in primary care settings. *International Journal of Medical Informatics, 204*, Article 106051. https://doi.org/10.1016/j.ijmedinf.2025.106051

Peiffer-Smadja, N., Descousse, S., Courrèges, E., Nganbou, A., Jeanmougin, P., Birgand, G., Lénaud, S., Beaumont, A.-L., Durand, C., Delory, T., Le Bel, J., Bouvet, E., Lariven, S., D'Ortenzio, E., Konaté, I., Bouyou-Akotet, M. K., Ouedraogo, A.-S., Kouakou, G. A., Poda, A., . . . Tanon, A. (2024). Implementation of a clinical decision support system for antimicrobial prescribing in sub-Saharan Africa: Multisectoral qualitative study. *Journal of Medical Internet Research, 26*, Article e45122. https://doi.org/10.2196/45122

Pieterse, P. (2025). Health workers in sub-Saharan Africa: Concurrent skilled health worker shortages and under-employment. *The International Journal of Health Planning and Management, 40*(6), 1335–1340. https://doi.org/10.1002/hpm.70001

Rath, S., Barzani, S., Okesanya, O. J., Tawfik, A. M., Shehada, W., Aziz, M. M., Urbon, S., & Alsabri, M. (2025). Optimizing pediatric emergency triage in low-resource settings: Evidence-based strategies, task-shifting, and technological innovations. *International Journal of Emergency Medicine, 18*(1), Article 216. https://doi.org/10.1186/s12245-025-00998-x

Sandmann, S., Hegselmann, S., Fujarski, M., Bickmann, L., Wild, B., Eils, R., & Varghese, J. (2025). Benchmark evaluation of DeepSeek large language models in clinical decision-making. *Nature Medicine, 31*(8), 2546–2549. https://doi.org/10.1038/s41591-025-03727-2

Scipion, C. E. A., Manchester, M. A., Federman, A., Wang, Y., & Arias, J. J. (2025). Barriers to and facilitators of clinician acceptance and use of artificial intelligence in healthcare settings: A scoping review. *BMJ Open, 15*(4), Article e092624. https://doi.org/10.1136/bmjopen-2024-092624

© IIARJournal. All rights reserved.

Sheffel, A., Andrews, K. G., Conner, R., Di Giorgio, L., Evans, D. K., Gatti, R., Lindelow, M., Sharma, J., Svensson, J., Wane, W., & Welander Tärneberg, A. (2024). Human resource challenges in health systems: Evidence from 10 African countries. *Health Policy and Planning, 39*(7), 693–709. https://doi.org/10.1093/heapol/czae034

Shidende, N., & Mwogosi, A. (2025). Exploring the impact of generative AI tools on healthcare delivery in Tanzania. *Journal of Health Organization and Management, 39*(8), 1580–1599. https://doi.org/10.1108/JHOM-01-2025-0007

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., . . . Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature, 620*(7972), 172–180. https://doi.org/10.1038/s41586-023-06291-2

Su, H., Sun, Y., Li, R., Zhang, A., Yang, Y., Xiao, F., Duan, Z., Chen, J., Hu, Q., Yang, T., Xu, B., Zhang, Q., Zhao, J., Li, Y., & Li, H. (2025). Large language models in medical diagnostics: Scoping review with bibliometric analysis. *Journal of Medical Internet Research, 27*, Article e72062. https://doi.org/10.2196/72062

Tegenaw, G. S., Sori, D. A., Teklemariam, G. K., Verbeke, F., Cornelis, J., & Jansen, B. (2024). Evaluation of a computer-aided clinical decision support system for point-of-care use in low-resource primary care settings: Acceptability evaluation study. *JMIR Human Factors, 11*, Article e47631. https://doi.org/10.2196/47631

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nature Medicine, 29*(8), 1930–1940. https://doi.org/10.1038/s41591-023-02448-8

Weissman, G. E., Mankowitz, T., & Kanter, G. P. (2025). Unregulated large language models produce medical device-like output. *npj Digital Medicine, 8*(1), Article 148. https://doi.org/10.1038/s41746-025-01544-y

Yu, L., Darmstadt, G. L., Ward, V., Wong, R. J., Stevenson, D. K., & Maric, I. (2026). Large language models for maternal and neonatal health care in low- and middle-income countries. *The Journal of Pediatrics, 293*, Article 115037. https://doi.org/10.1016/j.jpeds.2026.115037

Zhao, W., Wu, C., Fan, Y., Qiu, P., Zhang, X., Sun, Y., Zhou, X., Zhang, S., Peng, Y., Wang, Y., Sun, X., Zhang, Y., Yu, Y., Sun, K., & Xie, W. (2026). An agentic system for rare disease diagnosis with traceable reasoning. *Nature, 651*(8106), 775–784. https://doi.org/10.1038/s41586-025-10097-9

© IIARJournal. All rights reserved.