Oryx Centre for Strategic Futures

Research and policy initiative — AI, geopolitics, emerging technologies, strategic foresight and the Global South

Riya Nalwa, Proprietor — Udyam-registered micro enterprise, UDYAM-UP-56-0189301
BR 82, RWA Sector 45, Noida, UP 201303 · https://oryxcentre.dpdns.org · amartya@oryxcentre.dpdns.org

Audit Certificate



Seller stamp design — design only, not a physical stamp. Udyam number to be proofed character by character before any stamp maker use.

Report: What does published evidence say about large language models in healthcare, 2023–2026?

Job ID: 20261031-011705-c98c4d | **Model:** glm-4.7-flash | **Status at draft:** needs_review

Reviewer: Riya Nalwa | **Approval date:** 31 October 2026

Reviewer verdict (31 Oct 2026, 19:09 IST, verbatim): “content as reviewed, with no corrections — so it can proceed to final export, create all documents too, like yu will with my seller kit”

Operator approval: Amartya Kumar, 31 Oct 2026 ~19:12 IST, following the reviewer verdict above. Operator approval is recorded in the accompanying README.

Metrics (mechanical, unchanged from the draft): 10 candidates screened, 7 included, 3 excluded; 7 sources, 7 cited; evidence-table rows 7, verified 7/7; claim sentences 35, cited 21, sourced-claim rate 60%; citation markers 25,

invalid markers none; retrieval note: Semantic Scholar rate-limited (shared-IP budget).

Important: The adversarial-review flags in this report are preserved verbatim. The reviewer approved the content as reviewed, with no corrections — the flags were NOT resolved, removed, or hidden by this export. Citation precision is judged by the human reviewer, not computed mechanically. This certificate attests to review and approval of this draft as written; it is not a claim that every cited source supports every claim, and it is not a client delivery.

App-approval note: The public beta (oryxai.pythonanywhere.com) was down at export time, so the beta app's own Approve button could not be clicked. This package was therefore produced from the verified draft and the written approval chain above — no app approval is faked or claimed.

Oryx AI — Auditable Research Report

Question: What does published evidence say about large language models in healthcare, 2023-2026?

Status: FINAL — Human-reviewed (see Audit Certificate below). NOT sent to any client; delivered to Amartya Kumar and Riya Nalwa for a final look only.

Retrieval note: Semantic Scholar: rate-limited (shared-IP budget).

Screening: 10 candidates screened, 7 included, 3 excluded — see 03-screening-log.md.

Framing

Sub-questions

1. **Clinical Utility and Safety:** What is the published evidence regarding the diagnostic accuracy, treatment recommendation accuracy, and safety outcomes (e.g., hallucinations, adverse events) of LLMs applied to clinical decision-making tasks in 2023-2026?

2. **Operational Efficiency and Workflow Integration:** How have LLMs been utilized to automate clinical documentation, streamline administrative workflows, and impact clinician burnout in healthcare settings during this period?
3. **Regulatory and Ethical Frameworks:** What are the prevailing regulatory stances, ethical guidelines, and liability frameworks established for LLM deployment in healthcare between 2023 and 2026?
4. **Data Privacy and Security:** What evidence exists regarding the vulnerabilities of LLMs to data breaches, prompt injection attacks, and the preservation of patient confidentiality in healthcare contexts?

Scope

- **Timeframe:** January 1, 2023, to December 31, 2026.
- **Subject:** Large Language Models (LLMs) specifically, including proprietary models (e.g., GPT-4, Claude 3) and open-source models (e.g., Llama 2/3, Mistral), as applied to healthcare.
- **Context:** Healthcare settings, including hospitals, clinics, and telemedicine platforms.
- **Output:** Peer-reviewed literature (journals, conferences), clinical guidelines, and regulatory white papers.

Exclusions

- **Non-clinical applications:** Research focused solely on LLMs in finance, law, or general customer service, even if the underlying technology is the same.
- **Pre-2023 research:** Historical reviews or retrospective studies that do not include data or findings from the specified 2023-2026 window.
- **Non-peer-reviewed sources:** Op-eds, blog posts, or industry press releases, unless they are cited within peer-reviewed literature.
- **Non-LLM AI:** Research focusing on traditional machine learning (e.g., random forests, SVMs) or computer vision applied to medical imaging, unless the LLM is used for the reporting or administrative layer.

False-premise check

- **Premise:** The premise that a cohesive, definitive body of evidence exists for LLMs in healthcare in 2023-2026 is **potentially flawed**.

- **Reasoning:** The period 2023-2026 represents the "early adoption" and "rapid scaling" phase of generative AI. While there is a massive volume of literature, it is likely characterized by a mix of pilot studies, case reports, and early-phase trials, rather than large-scale, definitive randomized controlled trials (RCTs) or long-term outcome studies. The evidence is likely to be fragmented and evolving rapidly.
- **Reframed Question:** What are the current limitations and early findings of published evidence regarding the clinical utility and safety of large language models in healthcare, 2023-2026?

Findings

Executive answer

Evidence from 2023-2026 indicates that LLMs in healthcare are in a rapid scaling phase characterized by a significant "evaluation lag" where studies often lag behind the release of new models by up to 6 quarters [S1]. While promising for clinical decision-making and operational efficiency, the literature reveals substantial safety risks, with nearly 20% of outputs containing safety risks or biases that require rigorous adversarial testing [S5]. Regulatory frameworks are struggling to adapt to the flexibility of LLMs, creating a gap where applications can be repurposed or misused, challenging traditional liability and device definitions [S4][S7].

1. Clinical Utility and Safety

Confidence: High The evidence suggests that while LLMs demonstrate high capability in benchmark tasks, they suffer from significant safety vulnerabilities and a lack of standardized clinical reasoning evaluation [S3][S5][S6].

- **Diagnostic and Treatment Accuracy:** LLMs have shown high performance on standardized benchmarks (e.g., Med-PaLM) but these "exam-style" metrics do not reliably predict real-world clinical utility [S6]. A structured review of clinical reasoning dimensions indicates that current benchmarks often fail to assess critical skills like temporal synthesis and differential diagnosis effectively [S6].
- **Safety and Adverse Events:** There is a documented prevalence of safety risks and biases in LLM outputs, with recent data indicating nearly 20% of outputs contain such risks [S5]. Adversarial testing approaches are necessary to identify

vulnerabilities, as models can exhibit unpredictable behaviors when challenged by misleading inputs [S5].

- **Evaluation Lag:** A meta-analysis of 11,628 records found a widening "evaluation gap," where the mean lag between the release of a new model family and the publication of studies evaluating it has increased from 1.33 to 6.0 quarters [S1]. This lag suggests that the clinical literature is struggling to keep pace with the rapid evolution of the technology.

Evidence does NOT show: * Definitive proof of improved patient outcomes from LLMs in real-world clinical settings due to the lack of large-scale, long-term randomized controlled trials (RCTs) [S1]. * A unified standard for evaluating clinical reasoning that goes beyond multiple-choice style accuracy [S6].

2. Operational Efficiency and Workflow Integration

Confidence: Medium Evidence points to the potential of LLMs to automate documentation and streamline workflows, though specific data on clinician burnout reduction is currently limited to theoretical or early-stage applications [S2].

- **Clinical Documentation:** LLMs are increasingly utilized to automate clinical documentation, shifting the paradigm from retrospective record-keeping to prospective care provision [S2].
- **Workflow Integration:** There is a focus on multi-modal LLMs (MLLMs) that unify textual, imaging, and genomic data into single diagnostic and care pipelines, which could streamline complex workflows [S2].
- **Burnout:** While the structural inflection point of LLMs is expected to reduce clinician burnout by automating administrative tasks, specific published evidence quantifying this impact during 2023-2026 is not detailed in the provided sources [S2].

Evidence does NOT show: * Quantitative data on the reduction of clinician burnout rates in specific healthcare settings during this period.

3. Regulatory and Ethical Frameworks

Confidence: High The prevailing regulatory stance is one of adaptation and uncertainty, with a specific focus on the challenge of classifying LLMs as medical devices under frameworks like the MDR and AI Act [S4][S7].

- **Regulatory Stance:** Regulatory bodies are grappling with the challenge of regulating LLM-enabled medical AI, particularly where it extends beyond traditional medical device boundaries [S4].
- **Liability and Device Definition:** There is a regulatory gap identified regarding LLM applications that can be repurposed, malfunction, or be inadvertently misused for medical purposes, complicating liability frameworks [S7].
- **Governance:** Expert perspectives emphasize the need for robust governance to manage the flexibility of generative AI, which introduces new challenges for patient safety [S4].

Evidence does NOT show: * A finalized, universally adopted regulatory framework that fully addresses the unique risks of LLMs in healthcare.

4. Data Privacy and Security

Confidence: Medium The evidence highlights the critical need for rigorous security evaluation but does not provide specific data on the prevalence of breaches or successful attacks on healthcare LLMs during 2023-2026.

- **Vulnerabilities:** The flexibility of LLMs introduces new challenges regarding data privacy and security, as applications might be used in unintended ways that compromise patient confidentiality [S7].
- **Adversarial Attacks:** The literature emphasizes the need for adversarial testing and "red teaming" to identify system vulnerabilities, acknowledging that LLMs can be susceptible to prompt injection attacks and other security threats [S5].
- **Confidentiality:** While the sources discuss the risk to patient confidentiality, they do not provide specific statistics on the frequency of data breaches or successful attacks in healthcare contexts.

Evidence does NOT show: * Specific statistics on the frequency of data breaches or successful prompt injection attacks on healthcare LLMs during the 2023-2026 period.

Evidence table (verified — UNVERIFIED rows are marked, never dropped)

Source	Key finding	Evidence pointer	Limitation / gap	Relevance	Verification
[S1] PREreview of "The widening evaluation gap in medical large language model research 2023 to 2026"	The study maps 11,628 PubMed-indexed records on generative language models in healthcare from January 2023 to June 2026 and finds that the mean evaluation lag between model release and publication widens from 1.33 to 6.0 quarters.	[S1] This is an evidence map of 11,628 PubMed-indexed records on generative language models in healthcare, January 2023 to June 2026... Mean lag widens from 1.33 to 6.0	The source is a PREreview, which is a non-peer-reviewed synthesis of existing literature, and does not provide the detailed statistical analysis or full dataset of the underlying research.	Provides a quantitative overview of the volume of literature and the temporal gap between model development and evaluation in the specified timeframe.	**VERIFIED**
[S2] Generative AI and Large Language Models in Healthcare	LLMs represent a structural inflection point for care delivery, shifting focus from retrospective record-keeping to prospective, language-dependent clinical decision-making, and integrating multi-modal data (text, imaging,	[S2] ...are a structural inflection point for the delivery of care. A move away from retrospective record-keeping toward prospectives and language–dependent clinical decision-making.	The article is a review and does not present original empirical data or specific clinical outcomes from LLM deployment.	Describes the transformative potential and architectural evolution of LLMs in healthcare.	**VERIFIED**

Source	Key finding	Evidence pointer	Limitation / gap	Relevance	Verification
	genomics) into diagnostic pipelines.				
[S3] A primer on evaluation methods for large language models in healthcare	The review outlines four key areas for LLM evaluation in medicine: study design principles, statistical methods, capability evaluation (benchmarks), and clinical application evaluation.	[S3] This review covers four key areas of LLM evaluation: principles of study design, statistical methods, capability evaluation and clinical application evaluation.	The source is a review and does not provide new empirical evidence or specific performance metrics for LLMs.	Offers a framework for understanding how LLMs are currently being assessed in the medical literature.	**VERIFIED**
[S4] Expert perspectives on US regulatory approaches to large language models in healthcare	The article explores the regulatory challenges of LLM-enabled medical AI, particularly where it extends beyond traditional medical device boundaries, based on interviews with FDA advisors.	[S4] ...we explore the challenge of regulating large language model (LLM)-enabled medical AI, particularly where it extends beyond the usual boundaries of medical devices and therefore challenges traditional regulatory frameworks.	The findings are based on expert interviews and do not constitute a comprehensive legal analysis or regulatory policy proposal.	Provides insight into the regulatory landscape and the specific challenges of governing LLMs in the US healthcare system.	**VERIFIED**
[S5] Evaluating	Nearly 20% of LLM outputs	[S5] Recent data indicate	The source is a review and	Highlights the critical	**VERIFIED**

Source	Key finding	Evidence pointer	Limitation / gap	Relevance	Verification
the safety of large language models in healthcare and dentistry: adversarial testing approaches	contain safety risks or biases, necessitating rigorous adversarial testing (red teaming) to identify vulnerabilities before clinical use.	that nearly 20% of LLM outputs contain safety risks or biases, necessitating rigorous evaluation prior to clinical use.	does not provide the specific dataset or methodology used to derive the "nearly 20%" statistic.	importance of safety and bias mitigation in LLM deployment.	
[S6] A rubric landscape for evaluating clinical reasoning in large language models: what exists, what is missing, and what needs to be combined	The review maps existing assessment instruments and benchmarks to define clinical reasoning as integrating evidence over time, proposing a framework based on six dimensions (problem representation, temporal synthesis, etc.).	[S6] We define clinical reasoning as integrating and updating evidence across time and sources to form, revise and justify a patient's problem representation and a defensible plan.	The source is a narrative review and does not provide empirical validation of the proposed rubric or benchmarks.	Addresses the specific challenge of evaluating the reasoning capabilities of LLMs in clinical contexts.	**VERIFIED**
[S7] When are large language models medical devices? Navigating the MDR and AI Act for healthcare applications	The authors identify a regulatory gap where flexible LLM applications can be repurposed or misused, challenging the intended purpose of manufacturers	[S7] We have identified a regulatory gap whereby LLM applications that could be repurposed, malfunction or be inadvertently misused for	The source is a legal analysis and does not provide empirical data on the frequency of such regulatory gaps or misuses.	Discusses the regulatory and liability implications of LLM flexibility in healthcare.	**VERIFIED**

Source	Key finding	Evidence pointer	Limitation / gap	Relevance	Verification
	and raising liability concerns.	medical purposes			

Adversarial review flags (NOT fixes — reviewer must resolve each)

(a) claims whose cited source does NOT support them * "nearly 20% of outputs containing safety risks or biases that require rigorous adversarial testing" * "mean lag between the release of a new model family and the publication of studies evaluating it has increased from 1.33 to 6.0 quarters" * "LLMs are increasingly utilized to automate clinical documentation, shifting the paradigm from retrospective record-keeping to prospective care provision" * "focus on multi-modal LLMs (MLLMs) that unify textual, imaging, and genomic data into single diagnostic and care pipelines" * "structural inflection point of LLMs is expected to reduce clinician burnout by automating administrative tasks" * "regulatory gap identified regarding LLM applications that can be repurposed, malfunction, or be inadvertently misused for medical purposes, complicating liability frameworks" * "flexibility of LLMs introduces new challenges regarding data privacy and security, as applications might be used in unintended ways that compromise patient confidentiality" * "LLMs can be susceptible to prompt injection attacks and other security threats"

(b) material claims with NO citation * "LLMs in healthcare are in a rapid scaling phase" * "evidence suggests that while LLMs demonstrate high capability in benchmark tasks" * "LLMs have shown high performance on standardized benchmarks (e.g., Med-PaLM)" * "Definitive proof of improved patient outcomes from LLMs in real-world clinical settings" * "A unified standard for evaluating clinical reasoning that goes beyond multiple-choice style accuracy" * "LLMs are increasingly utilized to automate clinical documentation" * "focus on multi-modal LLMs (MLLMs) that unify textual, imaging, and genomic data" * "structural inflection point of LLMs is expected to reduce clinician burnout by automating administrative tasks" * "Regulatory bodies are grappling with the challenge of regulating LLM-enabled medical AI" * "regulatory gap identified regarding LLM applications that can be repurposed, malfunction, or be inadvertently misused for medical purposes, complicating liability frameworks" * "Expert perspectives emphasize the need for robust governance to manage the flexibility of generative AI" * "A finalized, universally adopted regulatory framework that fully addresses the unique risks of LLMs in healthcare" * "The flexibility of LLMs introduces

new challenges regarding data privacy and security" * "The literature emphasizes the need for adversarial testing and 'red teaming' to identify system vulnerabilities" * "LLMs can be susceptible to prompt injection attacks and other security threats"

(c) overstatements, hedged claims presented as facts, or numbers/dates/quotations not traceable to a cited source * "Evidence from 2023-2026 indicates that LLMs in healthcare are in a rapid scaling phase" * "nearly 20% of outputs containing safety risks or biases that require rigorous adversarial testing" * "mean lag between the release of a new model family and the publication of studies evaluating it has increased from 1.33 to 6.0 quarters" * "LLMs are increasingly utilized to automate clinical documentation, shifting the paradigm from retrospective record-keeping to prospective care provision" * "focus on multi-modal LLMs (MLLMs) that unify textual, imaging, and genomic data into single diagnostic and care pipelines" * "structural inflection point of LLMs is expected to reduce clinician burnout by automating administrative tasks" * "regulatory gap identified regarding LLM applications that can be repurposed, malfunction, or be inadvertently misused for medical purposes, complicating liability frameworks" * "flexibility of LLMs introduces new challenges regarding data privacy and security, as applications might be used in unintended ways that compromise patient confidentiality" * "LLMs can be susceptible to prompt injection attacks and other security threats"

UNSOURCED claims (quarantined — never findings)

- Confidence: High
- This lag suggests that the clinical literature is struggling to keep pace with the rapid evolution of the technology.
- Evidence does NOT show:
- Confidence: Medium
- Evidence does NOT show:
- Quantitative data on the reduction of clinician burnout rates in specific healthcare settings during this period.
- Confidence: High
- Evidence does NOT show:
- A finalized, universally adopted regulatory framework that fully addresses the unique risks of LLMs in healthcare.

- Confidence: Medium
- The evidence highlights the critical need for rigorous security evaluation but does not provide specific data on the prevalence of breaches or successful attacks on healthcare LLMs during 2023-2026.
- Confidentiality:** While the sources discuss the risk to patient confidentiality, they do not provide specific statistics on the frequency of data breaches or successful attacks in healthcare contexts.
- Evidence does NOT show:
- Specific statistics on the frequency of data breaches or successful prompt injection attacks on healthcare LLMs during the 2023-2026 period.

Sources (included — cited as [Sn])

1. PREreview of "The widening evaluation gap in medical large language model research 2023 to 2026" (2026). Evgeny V. Arsentyev. https://doi.org/10.5281/zenodo.22734669
2. Generative AI and Large Language Models in Healthcare (2026). G Muthupandi, Vishnu B, Dishanth R, Karuppiah S, Harish R K. https://doi.org/10.70593/978-81-69589-96-3_2
3. A primer on evaluation methods for large language models in healthcare (2026). Suzannah E. McKinney, Phuc Vu, Samuel A. Justice, Christopher Humphries, Alyssa Pradhan, Timothy Keyes, Kameron Collin Black, John L. Havlik, David JH Wu, Jack William O’Sullivan, Ethan Goh, Sarah Mercaldo, James Michael Hillis. https://doi.org/10.48550/arxiv.2609.14819
4. Expert perspectives on US regulatory approaches to large language models in healthcare (2026). Stephen Gilbert, Tinglong Dai. https://doi.org/10.1038/s41746-026-02787-z
5. Evaluating the safety of large language models in healthcare and dentistry: adversarial testing approaches (2026). Fahad Umer, Shaikh Mr, Arshalooz Jamila Rahman. https://doi.org/10.1038/s41405-026-00462-9
6. A rubric landscape for evaluating clinical reasoning in large language models: what exists, what is missing, and what needs to be combined (2026). Zhangshu Joshua Jiang, Zina M. Ibrahim, James Teo. https://doi.org/10.48550/arxiv.2610.01938

7. When are large language models medical devices? Navigating the MDR and AI Act for healthcare applications (2026). Jessica D. Workum, Sade Krijgsman, Ildikó Vajda, Robin van Stokkum, Jan van den Brand, Diederik A M P J Gommers, Ari Ercole, Michel Egide van Genderen. https://doi.org/10.1136/bmjdh-2026-000003

Excluded candidates are listed with reasons in 03-screening-log.md; they are never cited.

Metrics box (mechanical only — nothing estimated)

- Candidates screened: 10
- Sources included: 7
- Sources cited: 7
- Citation markers found: 25
- Invalid citation markers: none
- Evidence-table rows: 7
- Evidence-table rows verified: 7/7
- Claim sentences: 35
- Cited claim sentences: 21
- Sourced-claim rate: 60% (mechanical sentence rule: a sentence counts as cited if it contains at least one valid [Sn] marker; headings excluded)
- Citation precision: NOT mechanically computable — see adversarial review flags above; human reviewer judges whether each cited source supports its claim.
- Overall confidence: not computed mechanically — see per-section confidence in Findings; reviewer confirms.
- Human review: DONE — reviewer Riya Nalwa, verdict 31 Oct 2026 (see Audit Certificate).

Appendix: method note

Agent v0.4 staged run: framing, evidence acquisition, screening, synthesis, structured evidence table, mechanical table verification, mechanical claim-level verification, adversarial review, human gate. Sources: Semantic Scholar: rate-limited (shared-IP budget). Model: glm-4.7-flash. What would change the answer: additional primary sources, or resolution of the adversarial flags above.
