Essay:
Trustworthy Large Language Models for Clinical Decision Support

Clinical Decision Support

Master of Science in Artificial Intelligence in Medicine
University of Bern

Carla Paloma Malo Nivon

Matriculation number: 24-110-066

Choekyel Nyungmartsang

Matriculation number: 21-876-693

Monday 30th June, 2025

UNIVERSITÄT
BERN

Master of Science
Artificial Intelligence in Medicine

Eigenständigkeitserklärung

Hiermit versichern wir, dass wir die Arbeit mit dem Titel

---

eigenständig und nur mit den angegebenen Hilfsmitteln verfasst haben. Alle Texte, derer wir uns bedient haben, inklusive Quellen aus dem Internet, haben wir im Literaturverzeichnis aufgeführt. Wörtliche Übernahmen von Textpassagen Anderer haben wir als wörtliche Zitate gekennzeichnet. Wir haben auch die paraphrasierende Wiedergabe oder freie Übernahme fremder Gedanken durch Verweis auf einen Originaltext belegt. Falls wir Werkzeuge und Dienste der künstlichen Intelligenz eingesetzt haben oder die Arbeit von anderen Menschen haben gegenlesen lassen, haben wir dies in der Arbeit ausgewiesen. Wir versichern, dass die vorliegende Arbeit trotz Einsatz der etwaig aufgeführten Hilfsmittel und Unterstützungen im Wesentlichen unsere eigene Leistung ist.

Uns ist bewusst, dass ein Verstoss gegen diese Vorgaben den Regeln der guten wissenschaftlichen Praxis widerspricht und ein Plagiat darstellt, das nach Richtlinien der Universität Bern geahndet werden kann.

Damit diese und andere Arbeiten auf Plagiate überprüft werden können, erklären wir uns damit einverstanden, dass die Universität Bern diese Arbeit mit einer Plagiatssoftware überprüft und in einer Datenbank speichert, mit deren Hilfe zukünftige Arbeiten auf Plagiate überprüft werden können.

Bern, Monday 30th June, 2025 _____

Bern, Monday 30th June, 2025 _____

In 2015 a team taught pigeons to spot breast-cancer patterns on pathology slides. After only two weeks the birds reached 85 % accuracy, and when their answers were combined they matched expert pathologists [1]. If pigeons can learn patterns from pictures, machines that read huge amounts of text should do even better.

Large Language Models (LLMs) are deep neural networks that read billions of sentences and learn to guess the next word. They can draft notes, answer questions, and explain lab results in plain language. Because most clinical work is text-based, researchers now test LLMs as helpers in medical decisions. Early systems such as Med-PaLM, GatorTron, and ChatGPT have already passed medical exams and can summarise electronic health records in minutes.

Interest is high, but medicine is demanding. Wrong or biased advice can harm patients, leak private data, and damage public trust. This essay asks a simple question: what must change before LLMs can be trusted for clinical decision support? The next sections describe the current technology, examine five key barriers and then suggest practical ways to move ahead.

Overview of current state of technology

LLMs have advanced quickly since the transformer design made large-scale text learning practical. Today these models range from a few million to more than five hundred billion parameters, and many receive extra training on medical text to build clinical skill [2].

The first wave of general models was led by ChatGPT, trained on broad web data. It drew attention in medicine when it scored at or near the pass mark on all three steps of the United States Medical Licensing Exam (USMLE) even without special medical tuning [3]. Although the test used sample questions, the result showed that a text model can recall much of the knowledge doctors learn in school.

Researchers soon built health-specific models, tailored to different types of medical data. For instance, Med-PaLM and its successor Med-PaLM 2, trained on structured question-answer datasets, achieved 67 % and later on 86 % accuracy on USMLE-style medical exams, surpassing the human pass threshold [4]. In contrast, GatorTron, trained on a massive corpus of 90 billion words from unstructured EHRs, scaled up to 8.9 billion parameters. It outperformed previous biomedical and clinical transformers on all five clinical NLP tasks such as concept extraction and inference [5].

To compare these systems fairly, several benchmarks have been proposed, like MultiMedQA, that blends 6 medical licensing exams and a consumer dataset with new web health queries. Human reviewers then score answers for factual correctness, reasoning, and possible harm, exposing weaknesses that simple accuracy numbers can miss [6].

Ethical barriers to trustworthy LLMs

Research shows that language models can pass medical exams, but real-world care still brings serious risks. The five issues that follow hallucinations, bias, privacy leaks, lack of explanation, and loss of public trust, separate current early systems from safe daily use. Each subsection sets out the problem, offers evidence, and explains why it matters for patients and clinicians.

A. Hallucinations and factual errors

LLMs can speak with full confidence while inventing facts. This flaw is called a hallucination. A recent study in NEJM AI asked leading models to assign ICD-10 codes to real patient notes and found that even GPT-4 chose the exact correct code in fewer than half of the cases, while many outputs were imprecise or made-up [7, 8].

In routine care such slips are not minor quirks. A non-existent drug or wrong dose can delay treatment or cause an overdose, and incorrect codes can skew billing records and research data. The burden falls on clinicians, who must read every AI line but stay responsible if a hidden mistake appears in the record. Ethically this breaks the rule of "do no harm" and reduces informed consent because neither doctor nor patient can tell when the system is guessing. Until hallucinations drop to a level close to human error and models can flag uncertainty in plain language, their output cannot be trusted as a standalone source for clinical decisions.

B. Cognitive bias and healthcare disparities

LLMs learn from the text they read. If that text carries old race-based ideas, the model can copy them. A 2023 Nature study asked four popular models questions about kidney function, lung capacity, and pain. All four at times told users to adjust care by race or claimed that Black patients feel pain differently, even though these views have been disproved [9].

When a model repeats bias, the harm can spread quickly. Skewed dose or triage advice may steer doctors toward the wrong test or delay care for people who already struggle to get treatment. Biased cost or prognosis estimates can also push health plans and policy makers to unfair choices. These mistakes violate the duty of justice, raise the risk of unequal treatment, and weaken trust among patients who have seen bias before.

C. Privacy and data security

Language models keep traces of the text they saw during training. If that text includes real patient notes, skilled users can pull private facts back out. A 2024 Nature study showed that even after names and dates were removed, an attacker could ask short probe questions and learn whether a patients record had been in the training set [10].

Leaks break laws like the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA) and can harm patients far beyond the clinic. A revealed HIV status or rare-disease note could hurt a persons chance to get insurance or a job. Fear of such leaks may also stop hospitals from sharing data that is needed to train fairer models. Technical fixes like local hospital deployment, differential-privacy noise, and strict logging are under study, but none fully block the risk yet [11]. Until clear guardrails show that no sensitive text can escape, privacy worries will limit how far LLMs can move into everyday clinical work.

D. Lack of explainability

Most language models can give an answer but cannot show how they reached it. Apple researchers tested new "reasoning" models and saw them solve easy puzzles, then fail on harder ones while still speaking with full confidence. They called this gap the "illusion of thinking" and showed that longer chains of thought did not lead to better logic [12].

In medicine this lack of clarity is a problem. Doctors must explain to patients and colleagues why they order a test or change a drug dose. If the model offers only a smooth answer without clear evidence, the clinician has little ground to defend the choice. Hidden logic also blocks safety checks because reviewers cannot trace where an error began. This weakens accountability and leaves both patients and providers unsure whether to follow the advice. Until models can show reliable, easy-to-read reasons for each clinical step, their guidance will stay outside the main decision process.

E. Erosion of trust in clinical AI overall

Bad news spreads quickly. A 2024 Nature Medicine survey asked people how much they trust doctors who use AI support. Many said they prefer a doctor who works without any AI, and headlines about failed tools were the main reason [13]. When a well-known system fails, people start to doubt every AI tool. Each public failure makes it harder for newer, safer models to gain acceptance. Good systems may never reach patients if the public sees all clinical AI as risky. Fixing this will need strict testing, open sharing of errors, and clear proof that new language models avoid the mistakes of the old ones.

Implications for key stakeholders

Clinicians are the first to handle LLM output, and patients depend on the advice that follows. Responsibility for care still rests with the human professional, so every AI suggestion needs a quick but careful review. The extra time spent checking can extend visits, and patients notice the delay. A 2024 Nature Medicine survey found that many people trust medical advice less once they learn an AI helped, even when the answer is correct [13]. When hallucinations, bias, or data leaks slip through, trust can collapse and workloads can rise instead of fall.

Regulators are acting in parallel. The EU AI Act puts clinical decision support in the high-risk category, which means external audits, full trace records, and clear proof of safety are required before launch [14]. For developers this means privacy by design, clear uncertainty flags, and public test reports are now expected. Without these steps, public concern can keep even the best tools out of daily use.

Conclusion

LLMs will not earn a place in clinical care until five gaps close: fewer hallucinations, regular bias checks, strong privacy protection, clear explanations, and open error reports. Each fix is within reach. Benchmarks such as MultiMedQA now measure reasoning and potential harm, giving teams a clear target. The EU AI Act already lists decision support as high risk, so builders must show safety data before release. Local deployment, retrieval-augmented design, and simple uncertainty flags are reducing mistakes in trial units.

Progress is steady, but trust will depend on proof from daily practice. Strong audits, shared lessons, and plain-language reports can show patients and doctors that these tools are safe. And while pigeons will never roam hospital halls, their success at spotting cancer reminds us how powerful pattern learning can be. With the right safeguards, language models can become reliable partners for clinicians and their patients.

Bibliography

[1] Richard M. Levenson et al. “Pigeons (Columba Livia) as Trainable Observers of Pathology and Radiology Breast Cancer Images”. In: *PLOS ONE* 10.11 (Nov. 18, 2015). Ed. by Jonathan A Coles, e0141357. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0141357. URL: https://dx.plos.org/10.1371/journal.pone.0141357 (visited on 06/26/2025).

[2] Hongjian Zhou et al. “A Survey of Large Language Models in Medicine: Progress, Application, and Challenge”. In: ().

[3] Tiffany H. Kung et al. “Performance of ChatGPT on USMLE: Potential for AI-assisted Medical Education Using Large Language Models”. In: *PLOS Digital Health* 2.2 (Feb. 9, 2023). Ed. by Alon Dagan, e0000198. ISSN: 2767-3170. DOI: 10.1371/journal.pdig.0000198. URL: https://dx.plos.org/10.1371/journal.pdig.0000198 (visited on 06/28/2025).

[4] *Med-PaLM: A Medical Large Language Model - Google Research*. Med-PaLM: A Medical Large Language Model - Google Research. URL: https://sites.research.google/med-palm/ (visited on 06/28/2025).

[5] Xi Yang et al. “A Large Language Model for Electronic Health Records”. In: *npj Digital Medicine* 5.1 (Dec. 26, 2022), p. 194. ISSN: 2398-6352. DOI: 10.1038/s41746-022-00742-2. URL: https://www.nature.com/articles/s41746-022-00742-2 (visited on 06/28/2025).

[6] Karan Singhal et al. “Large Language Models Encode Clinical Knowledge”. In: *Nature* 620.7972 (Aug. 3, 2023), pp. 172–180. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-023-06291-2. URL: https://www.nature.com/articles/s41586-023-06291-2 (visited on 06/28/2025).

[7] *AI Not Ready to Perform Basic Medical Coding | Norwood*. Apr. 26, 2024. URL: https://www.norwood.com/nejm-study-ai-not-ready-to-perform-basic-medical-coding-let-alone-replace-people/ (visited on 06/28/2025).

[8] Ali Soroush et al. “Large Language Models Are Poor Medical Coders — Benchmarking of Medical Code Querying”. In: *NEJM AI* 1.5 (Apr. 25, 2024). ISSN: 2836-9386. DOI: 10.1056/AIdbp2300040. URL: https://ai.nejm.org/doi/10.1056/AIdbp2300040 (visited on 06/28/2025).

[9] Jesutofunmi A. Omiye et al. “Large Language Models Propagate Race-Based Medicine”. In: *npj Digital Medicine* 6.1 (Oct. 20, 2023), p. 195. ISSN: 2398-6352. DOI: 10.1038/s41746-023-00939-z. URL: https://www.nature.com/articles/s41746-023-00939-z (visited on 06/28/2025).

[10] Atiquer Rahman Sarkar et al. “De-Identification Is Not Enough: A Comparison between de-Identified and Synthetic Clinical Notes”. In: *Scientific Reports* 14.1 (Nov. 29, 2024), p. 29669. ISSN: 2045-2322. DOI: 10.1038/s41598-024-81170-y. URL: https://www.nature.com/articles/s41598-024-81170-y (visited on 06/28/2025).

[11] Jitendra Jonnagaddala and Zoie Shui-Yee Wong. "Privacy Preserving Strategies for Electronic Health Records in the Era of Large Language Models". In: *npj Digital Medicine* 8.1 (Jan. 16, 2025), p. 34. ISSN: 2398-6352. DOI: 10.1038/s41746-025-01429-0. URL: https://www.nature.com/articles/s41746-025-01429-0 (visited on 06/28/2025).

[12] Parshin Shojaee et al. "The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity". In: ().

[13] Moritz Reis, Florian Reis, and Wilfried Kunde. "Influence of Believed AI Involvement on the Perception of Digital Medical Advice". In: *Nature Medicine* 30.11 (Nov. 2024), pp. 3098–3100. ISSN: 1078-8956, 1546-170X. DOI: 10.1038/s41591-024-03180-7. URL: https://www.nature.com/articles/s41591-024-03180-7 (visited on 06/28/2025).

[14] Hannah Van Kolfschooten and Janneke Van Oirschot. "The EU Artificial Intelligence Act (2024): Implications for Healthcare". In: *Health Policy* 149 (Nov. 2024), p. 105152. ISSN: 01688510. DOI: 10.1016/j.healthpol.2024.105152. URL: https://linkinghub.elsevier.com/retrieve/pii/S0168851024001623 (visited on 06/28/2025).