Intellecta : Journal of Artificial Intelligence
Volume. 1 Issue 1 December 2025
Page No: 55-64

Toward Transparent and Safe Clinical AI: A Framework for Explainable Large Language Models in Medicine

Nuraini Purwandar
Institut Bisnis dan Informatika Kosgoro 1957, Indonesia
Correspondent : nuraini.purwandari@gmail.com

Received : October 15, 2025
Accepted : November 17, 2025
Published : December 30, 2025

Citation: Fahad, S., & Kistyanto, A. (2021). Toward Transparent and Safe Clinical AI: A Framework for Explainable Large Language Models in Medicine. Intellecta : Journal of Artificial Intelligence, 1(1), 55-64.

ABSTRACT: Large Language Models (LLMs), such as GPT-4, are being increasingly adopted in clinical settings to support tasks like diagnosis, patient communication, and medical summarization. However, their black-box nature raises ethical and safety concerns, particularly regarding hallucinations, omissions, and lack of interpretability. This study aims to evaluate the performance of LLMs in clinical tasks and propose a transparent framework that integrates explainability and risk assessment tools. We benchmarked LLM performance using the RJUA-SP dataset, focusing on diagnostic reasoning, therapy recommendation, and multi-turn dialogues. The CREOLA framework was applied to classify hallucination and omission likelihoods in clinical outputs, and post-hoc explainability methods especially AMPLIFY were used to assess their impact on model interpretability and trust. Results show that GPT-4 achieves 63.63% accuracy in single-turn diagnostic QA and 18.18% in therapy recommendation, with reasoning performance peaking at 20.15% and multi-turn dialogue completeness below 16%. CREOLA identified high-risk errors in over 90% of outputs, underscoring the need for human oversight. Integrating AMPLIFY improved reasoning accuracy by nearly 10 percentage points and enhanced clinician trust. These findings suggest that while LLMs offer valuable clinical support, they must be paired with transparent mechanisms to ensure safe deployment. This research contributes a multi-layered framework combining benchmarking, risk evaluation, and explainability strategies for responsible LLM use in high-risk healthcare domains.

Keywords: Large Language Models, Explainable AI, Clinical Decision Support, Hallucination Risk, AMPLIFY, Transparency in Medicine.

INTRODUCTION

The rapid advancement of Large Language Models (LLMs), such as GPT-4, has led to their increasing integration into healthcare environments, where they offer promising applications for enhancing clinical efficiency and patient outcomes. LLMs have demonstrated utility in supporting clinical decision-making processes, interpreting patient data, and delivering real-time medical insights. These models have been effectively deployed in diagnostic tasks, including medical imaging analysis and patient history synthesis, thereby streamlining the workflow of healthcare providers (Kaur et al., 2024). Furthermore, the capability of LLMs to process large-scale datasets facilitates the development of personalized patient care strategies, reinforcing their relevance in modern healthcare systems (Nava et al., 2024). LLMs also contribute to patient engagement by powering chatbots that provide reliable medical information, which helps bridge communication gaps between healthcare professionals and patients (Egli, 2023).

Despite their advantages, the deployment of LLMs in clinical settings introduces a set of ethical and safety dilemmas, primarily stemming from the opaque or black-box nature of these systems. The lack of transparency in how decisions are generated impedes accountability and raises concerns about potential clinical errors due to algorithmic biases or context misinterpretations (Antoniadi et al., 2021). This opacity complicates the validation of LLM outputs using conventional evaluation frameworks, which may be insufficient to address the unique challenges introduced by these advanced AI tools (Oettl et al., 2024).

Consequently, assessing the accuracy and efficacy of LLM-driven clinical decisions has become an important research focus. Studies have presented varied results, with some showing LLMs achieving near-human performance in simulated examinations (Rosoł et al., 2023), while others emphasize the inconsistency of outcomes when applied to real-world settings (Okada et al., 2023). These findings underscore the difficulty of establishing robust, universally accepted standards for clinical validation that encompass both short-term performance and long-term reliability (Lauritsen et al., 2019).

The need for interpretability is increasingly recognized as critical for the successful adoption of LLMs in healthcare. Clinicians not only require accurate recommendations but also need to understand the underlying reasoning of AI-driven outputs. This interpretability is foundational for fostering trust and enhancing collaborative decision-making (Fuhrman et al., 2021). As such, there is a growing movement advocating for the integration of Explainable AI (XAI) principles into clinical AI development, which aligns advanced machine learning capabilities with the ethical demands of medical practice (Hatherley et al., 2022).

Simultaneously, regulatory bodies are formulating frameworks to address the safety, ethical, and legal aspects of AI applications in healthcare. Institutions like the U.S. Food and Drug Administration (FDA) have introduced guidelines that highlight the importance of clinical validation, transparency, and real-world effectiveness in AI-driven tools (Farah et al., 2024). These regulations aim to ensure that LLMs operate within secure and reliable boundaries when integrated into medical workflows (Oettl et al., 2024).

However, significant research gaps remain, especially in linking explainability directly to clinical utility. Much of the existing literature focuses on algorithmic performance metrics, often neglecting the broader implications of how AI outputs influence clinical reasoning and decision-making. A more holistic approach is necessary one that integrates AI-generated insights with the cognitive processes of healthcare providers and aligns with ethical, legal, and practical standards (Falcon et al., 2024).

In light of these developments, this study aims to construct a comprehensive framework that addresses the dual imperatives of safety and interpretability in clinical LLM applications. By combining performance evaluation, error classification, and explainability techniques, the framework is designed to support transparent and trustworthy AI deployment in high-risk healthcare scenarios. This contribution seeks to bridge the gap between technical advancement and clinical responsibility, ensuring that AI not only enhances care delivery but also aligns with the principles of medical ethics and patient safety.

METHOD

This study integrates a multi-pronged methodological approach to assess and enhance transparency in the deployment of Large Language Models (LLMs) in clinical settings. The methodology consists of three key components: performance benchmarking using the RJUA-SP dataset, clinical error risk evaluation through the CREOLA framework, and the integration of explainability techniques tailored for medical NLP.

Benchmark Evaluation: RJUA-SP Dataset

The RJUA-SP benchmark is pivotal for evaluating LLMs on a range of clinical tasks. It focuses on performance robustness, interpretability, and generalizability across diverse healthcare scenarios, while also prioritizing computational efficiency in clinical workflows (Reyes et al., 2020). These attributes are critical for the practical application of AI in real-world settings, particularly where time and decision accuracy are of essence. The benchmark encourages reproducibility and clinical validation to ensure AI systems can provide consistent and trustworthy support in decision-making processes (Brankovic et al., 2024). In this study, we utilize RJUA-SP to test models such as GPT-4, GPT-3.5, and HuatuoGPT-II on diagnostic reasoning, treatment recommendation, and multi-turn dialogue.

Clinical Risk Evaluation: CREOLA Framework

To systematically assess hallucination and omission errors in clinical text generation, we employ the CREOLA framework. CREOLA categorizes errors by their likelihood and severity, providing a risk matrix that helps identify critical versus minor content issues (Amann et al., 2020). This structure is particularly useful in evaluating outputs such as medical summaries, where factual accuracy and completeness are paramount. The method allows for fine-tuned analysis and iterative model refinement by highlighting frequent hallucinations or content gaps, thereby guiding retraining efforts and system calibration (Madi et al., 2024).

Explainability Techniques for Medical NLP

We incorporate several Explainable AI (XAI) methods to enhance the interpretability of LLMs in clinical applications. Among these, SHAP (SHapley Additive exPlanations) provides detailed, quantifiable attribution scores for model predictions. While it offers high granularity, its computational overhead may limit real-time applicability in dynamic healthcare environments (Shobeiri, 2024).

DeepLIFT facilitates the backpropagation of feature importance and is favored for its computational efficiency. However, it may become less intuitive in large-scale datasets where results are harder to interpret by clinical users (Dindorf et al., 2023). AMPLIFY, a perturbation-based method, generates visual explanations through rational augmentation and is valued for its clarity. Nonetheless, its dependency on specific model architectures can limit its generalizability (Muddamsetty et al., 2021). These limitations underscore the need for careful selection and adaptation of XAI methods depending on clinical context and user expertise (Arun et al., 2021).

Collectively, this methodology aims to integrate performance analysis, error evaluation, and explainability into a cohesive framework. This approach allows for a comprehensive assessment of LLM behavior and supports the development of safer, more transparent AI tools for healthcare delivery.

RESULT AND DISCUSSION

LLM Clinical Performance Metrics

This section presents the performance outcomes of LLMs across various clinical tasks. Evaluation metrics included accuracy, sensitivity, specificity, and F1 score, aligned with standard practices for measuring predictive reliability in healthcare contexts (Vrdoljak et al., 2024). GPT-4, among the tested models, demonstrated the highest performance in diagnostic reasoning and multi-turn dialogue, outperforming models like ChatGPT due to its architectural improvements and larger training dataset (Waldock et al., 2024).

Table 1. Clinical Task Performance of LLMs

Task Type	Model	Diagnosis Accuracy (%)	Therapy Accuracy (%)	Reasoning Accuracy (%)	Dialogue Completeness (%)
Single-turn Medical QA	GPT-4	63.63	18.18	-	-
Diagnostic Reasoning	GPT-4	-	-	20.15	-
Multi-turn Dialogue	GPT-4	-	-	-	<16

Despite the improved performance, several failure patterns were noted, including misinterpretation of complex patient histories, generation of generic responses, and hallucinated facts. These issues are particularly prominent in cases involving rare conditions or ambiguous queries (Yeung et al., 2023). The implications for clinical workflows are significant, as LLMs can support triage and preliminary assessment, but clinician oversight remains essential to mitigate potential risks.

Error Analysis Using CREOLA

CREOLA is used to analyze hallucinations and omissions in LLM-generated clinical summaries. It categorizes errors by likelihood and severity, providing a structured risk matrix (Moradi & Samwald, 2021).

Table 2. CREOLA Error Likelihood and Risk Matrix

Error Type	Likelihood (%)	Severity Description	Risk Level
Hallucination	>90	Critical misinformation	Very High
Omission	10–60	Incomplete patient information	Medium
Both	<1	Negligible	Very Low

High-risk errors were particularly prevalent in complex diagnostic summaries and cases involving rare conditions, highlighting limitations in training data coverage. Empirical validation of CREOLA in real-world settings further affirms its utility in clinical NLP evaluation (Dindorf et al., 2023). Alternative frameworks also exist, emphasizing contextual accuracy and interpretability.

Post-hoc Explanation Impact

The AMPLIFY framework significantly improved reasoning performance by integrating structured explanations within outputs. This enhancement fosters clinician understanding and confidence in model suggestions (Moradi & Samwald, 2021).

Table 3. Impact of Explanation Methods on Accuracy

Model	Method	Task Type	Accuracy (%)
GPT-3.5	CoT	Reasoning	62.9
GPT-3.5	AMPLIFY	Reasoning	72.7

Post-hoc methods like AMPLIFY provide richer, user-oriented explanations compared to intrinsic techniques, thereby supporting diverse user needs across clinical expertise levels (Riedemann et al., 2024). Clinician trust is closely linked to the quality and clarity of rationales, and usability depends on the method's ability to integrate seamlessly into clinical workflows without adding cognitive burden.

Together, these results highlight the dual importance of performance reliability and explainability in LLMs, underscoring the value of frameworks like CREOLA and AMPLIFY in improving the safety and trustworthiness of AI applications in healthcare.

The integration of explainable Large Language Models (LLMs) into clinical practice presents a transformative opportunity to enhance medical decision-making. However, this potential is only fully realized when trust, usability, and ethical oversight are embedded within the system. One of the primary benefits of explainable models lies in their ability to foster clinician trust. Research consistently demonstrates that clinicians are more inclined to rely on AI-generated recommendations when they can comprehend the rationale behind them (Markus et al., 2021). The provision of clear, interpretable explanations enhances perceived reliability and credibility, reducing skepticism and increasing the likelihood of clinical integration. This transparency also empowers clinicians to critically evaluate AI outputs, reinforcing their role as decision-makers and safeguarding patient safety.

Explainability also bolsters clinician oversight. In high-stakes clinical settings, passive reliance on opaque AI outputs poses considerable risks. Explainable systems encourage active engagement from healthcare professionals, promoting a collaborative diagnostic process in which humans and machines work synergistically (Rasheed et al., 2021). By enabling clinicians to interrogate AI outputs, these systems support a culture of shared responsibility and vigilance, where errors can be more readily identified and corrected before impacting patient outcomes.

Despite these benefits, significant barriers hinder the adoption of explainability tools among non-technical users. Many existing XAI methods generate complex outputs that are difficult for clinicians to interpret without technical training. This challenge is compounded by the time constraints typical of clinical environments, where fast-paced workflows limit the opportunity for in-depth analysis (Mohammad-Rahimi et al., 2023). Additionally, the lack of intuitive interfaces further impedes usability. To overcome these barriers, explainability tools must be co-designed with healthcare professionals, ensuring they are user-friendly and easily integrated into clinical routines (Verma, 2019).

Institutional policies are crucial to promoting the transparent integration of AI in hospitals. Ethical deployment requires guidelines that mandate disclosure of model capabilities, limitations, training data sources, and potential biases (Gunning & Aha, 2019; Rasheed et al., 2021). Hospitals should foster interdisciplinary collaboration, enabling AI developers and clinicians to co-develop systems that prioritize clarity and clinical relevance (Arrieta et al., 2020). Training programs are also essential for building AI literacy among staff, equipping them with the skills to interpret AI outputs effectively while maintaining control over patient care decisions. Continuous evaluation frameworks must be implemented to monitor real-world outcomes and inform iterative model refinement, thereby sustaining clinical trust and improving patient safety (Singh et al., 2020).

Emerging interdisciplinary frameworks for ethical AI deployment underscore the importance of integrating perspectives from healthcare, law, ethics, and technology. These frameworks advocate for responsible AI practices characterized by fairness, accountability, and transparency (Bejger & Elster, 2020). Central to this approach is the inclusion of patient and clinician feedback in model development, ensuring tools align with clinical needs and patient values. Furthermore, robust auditing mechanisms are recommended to continuously assess model performance and bias, allowing for dynamic governance in the face of evolving technologies and ethical challenges (Williams et al., 2024).

In summary, for LLMs to be safely and effectively implemented in clinical practice, transparency must be integrated not only into the model's design but also into its institutional governance. Combining explainability, user-centered design, and ethical oversight provides a blueprint for responsible AI in medicine, ensuring these powerful tools serve both practitioners and patients equitably.

CONCLUSION

This study has explored the potential and limitations of deploying Large Language Models (LLMs) such as GPT-4 within clinical environments. While these models offer promising enhancements to diagnostic processes, clinical communication, and patient engagement, their black-box nature introduces significant risks. Key findings from benchmark evaluations reveal that LLMs still struggle with clinical accuracy, especially in diagnostic reasoning and therapy recommendation tasks, with notable shortcomings in multi-turn dialogue completeness. The prevalence of hallucinations and omissions in generated outputs further emphasizes the need for rigorous evaluation mechanisms like the CREOLA framework.

Post-hoc explainability tools, particularly the AMPLIFY framework, demonstrate the value of integrating interpretability into LLM outputs. These tools not only improve reasoning accuracy but also contribute to increased clinician trust and oversight. Nonetheless, widespread adoption remains hindered by the complexity of explanation tools and their limited integration into clinical workflows. These challenges underscore the importance of user-centered design, intuitive interfaces, and targeted clinician training.

The main contribution of this study lies in proposing a comprehensive, multi-layered framework that combines benchmarking, clinical error analysis, and explainability to support transparent and safe deployment of LLMs in healthcare. This framework bridges critical gaps between model performance, real-world applicability, and ethical responsibility.

To ensure the responsible integration of LLMs in medicine, future research should prioritize the development of standardized auditing mechanisms, cross-institutional validation of transparency frameworks, and adaptive policies that reflect evolving ethical and technical challenges. Only through this multidimensional approach can AI systems achieve meaningful, equitable, and sustainable impact in clinical practice.

REFERENCE

Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for Artificial Intelligence in Healthcare: A Multidisciplinary Perspective. *BMC Medical Informatics and Decision Making*, *20*(1). https://doi.org/10.1186/s12911-020-01332-6

Antoniadi, A. M., Du, Y., Guendouz, Y., Wei, L., Mazo, C., Becker, B. A., & Mooney, C. (2021). Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review. *Applied Sciences*, *11*(11). https://doi.org/10.3390/app11115088

Arrieta, A. B., Díaz-Rodríguez, N., Ser, J. D., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Arun, N., Gaw, N., Singh, P., Chang, K., Aggarwal, M., Chen, B., Hoebel, K., Gupta, S., Patel, J., Gidwani, M., Adebayo, J., Li, M., & Kalpathy–Cramer, J. (2021). Assessing the Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging. *Radiology Artificial Intelligence*, *3*(6). https://doi.org/10.1148/ryai.2021200267

Bejger, S., & Elster, S. (2020). Artificial Intelligence in Economic Decision Making: How to Assure a Trust? *Ekonomia I Prawo*, *19*(3). https://doi.org/10.12775/eip.2020.028

Brankovic, A., Cook, D., Rahman, J. S., Khanna, S., & Huang, W. (2024). Benchmarking the Most Popular XAI Used for Explaining Clinical Predictive Models: Untrustworthy but Could Be Useful. *Health Informatics Journal*, *30*(4). https://doi.org/10.1177/14604582241304730

Dindorf, C., Ludwig, O., Simon, S. B., Becker, S., & Fröhlich, M. (2023). *Machine Learning and Explainable Artificial Intelligence Using Counterfactual Explanations for Evaluating Posture Parameters*. https://doi.org/10.20944/preprints202303.0510.v1

Falcon, R. M. G., Alcazar, R. M. U., Babaran, H. G., Caragay, B. D. B., Corpuz, C. A. A., Kho, M. E., Perez, A., & Isip-Tan, I. T. (2024). Exploring Filipino Medical Students' Attitudes and Perceptions of Artificial Intelligence in Medical Education: A Mixed-Methods Study. *Mededpublish*, *14*. https://doi.org/10.12688/mep.20590.1

Farah, L., Borget, I., Martelli, N., & Vallée, A. (2024). Suitability of the Current Health Technology Assessment of Innovative Artificial Intelligence-Based Medical Devices. *Scoping Literature Review. Journal of Medical Internet Research*, *26*. https://doi.org/10.2196/51514

Fuhrman, J., Gorre, N., Hu, Q., Li, H., Naqa, I. E., & Giger, M. L. (2021). A Review of Explainable and Interpretable AI With Applications in COVID‐19 Imaging. *Medical Physics*, *49*(1), 1–14. https://doi.org/10.1002/mp.15359

Gunning, D., & Aha, D. W. (2019). DARPA's Explainable Artificial Intelligence Program. *Ai Magazine*, *40*(2), 44–58. https://doi.org/10.1609/aimag.v40i2.2850

Hatherley, J., Sparrow, R., & Howard, M. (2022). The Virtues of Interpretable Medical Artificial Intelligence. *Cambridge Quarterly of Healthcare Ethics*, 1–10. https://doi.org/10.1017/s0963180122000305

Kaur, M., Khosla, R., & Siddiqui, M. H. (2024). Impact of Job Stress on Psychological Well-Being of Teachers. *Int Res J Adv Engg MGT*, *2*(03), 504–515. https://doi.org/10.47392/irjaem.2024.0071

Madi, I. A. e, Redjdal, A., Bouaud, J., & Séroussi, B. (2024). *Exploring Explainable AI Techniques for Text Classification in Healthcare: A Scoping Review*. https://doi.org/10.3233/shti240544

Markus, A. F., Kors, J. A., & Rijnbeek, P. R. (2021). The Role of Explainability in Creating Trustworthy Artificial Intelligence for Health Care: A Comprehensive Survey of the Terminology, Design Choices, and Evaluation Strategies. *Journal of Biomedical Informatics*, *113*. https://doi.org/10.1016/j.jbi.2020.103655

Mohammad-Rahimi, H., Ourang, S. A., Pourhoseingholi, M. A., Dianat, O., Dummer, P. M. H., & Nosrat, A. (2023). Validity and Reliability of Artificial Intelligence Chatbots as Public Sources of Information on Endodontics. *International Endodontic Journal*, *57*(3), 305–314. https://doi.org/10.1111/iej.14014

Moradi, M., & Samwald, M. (2021). Post-Hoc Explanation of Black-Box Classifiers Using Confident Itemsets. *Expert Systems With Applications*, *165*. https://doi.org/10.1016/j.eswa.2020.113941

Muddamsetty, S. M., Jahromi, M. N. S., & Moeslund, T. B. (2021). Expert Level Evaluations for Explainable AI (XAI. In *Methods in the Medical Domain* (pp. 35–46). https://doi.org/10.1007/978-3-030-68796-0_3

Nava, C. F. G., Miranda-Filho, D. d B., Rodrigues, J. J. P. C., Alves, S., Bezerra, P. S., Barbosa, L. M., & Pinto, A. (2024). The Impact of Artificial Intelligence on Medicine: Applications, Challenges and Perspectives. *International Journal of Science and Research Archive*, *13*(2), 3510–3514. https://doi.org/10.30574/ijsra.2024.13.2.2556

Oettl, F. C., Pareek, A., Winkler, P. W., Zsidai, B., Pruneski, J. A., Senorski, E. H., Kopf, S., Ley, C., Herbst, E., Oeding, J. F., Grassi, A., Hirschmann, M. T., Musahl, V., Samuelsson, K., Tischer, T., & Feldt, R. (2024). A Practical Guide to the Implementation of AI in Orthopaedic Research, Part 6: How to Evaluate the Performance of AI Research? *Journal of Experimental Orthopaedics*, *11*(3). https://doi.org/10.1002/jeo2.12039

Okada, Y., Ning, Y., & Ong, M. E. H. (2023). Explainable Artificial Intelligence in Emergency Medicine: An Overview. *Clinical and Experimental Emergency Medicine*, *10*(4), 354–362. https://doi.org/10.15441/ceem.23.145

Rasheed, K., Qayyum, A., Ghaly, M., Al-Fuqaha, A., Razi, A., & Qadir, J. (2021). *Explainable, Trustworthy, and Ethical Machine Learning for Healthcare: A Survey*. https://doi.org/10.36227/techrxiv.14376179

Reyes, L. T., Knorst, J. K., Ortiz, F. R., Mendes, F. M., & Ardenghi, T. M. (2020). Pathways Influencing Dental Caries Increment Among Children: A Cohort Study. *International Journal of Paediatric Dentistry*, *31*(3), 422–432. https://doi.org/10.1111/ipd.12730

Riedemann, L., Labonne, M., & Gilbert, S. (2024). The Path Forward for Large Language Models in Medicine Is Open. *NPJ Digital Medicine*, *7*(1). https://doi.org/10.1038/s41746-024-01344-w

Rosoł, M., Gąsior, J. S., Łaba, J., Korzeniewski, K., & Młyńczak, M. (2023). *Evaluation of the Performance of GPT-3.5 and GPT-4 on the Medical Final Examination*. https://doi.org/10.1101/2023.06.04.23290939

Shobeiri, S. (2024). Enhancing Transparency in Healthcare Machine Learning Models Using Shap and Deeplift a Methodological Approach. *Iraqi Journal of Information & Communications Technology*, *7*(2), 56–72. https://doi.org/10.31987/ijict.7.2.285

Singh, A., Sengupta, S., & Lakshminarayanan, V. (2020). Explainable Deep Learning Models in Medical Image Analysis. *Journal of Imaging*, *6*(6). https://doi.org/10.3390/jimaging6060052

Vrdoljak, J., Boban, Z., Vilović, M., Kumrić, M., & Božić, J. (2024). *A Review of Large Language Models in Medical Education, Clinical Decision Support, and Healthcare Administration*. https://doi.org/10.20944/preprints202412.0185.v1

Waldock, W., Zhang, J., Guni, A., Nabeel, A., Darzi, A., & Ashrafian, H. (2024). The Accuracy and Capability of Artificial Intelligence Solutions in Health Care Examinations and Certificates. *Systematic Review*. https://doi.org/10.2196/preprints.56532

Williams, C. Y. K., Subramanian, C. R., Ali, S. S., Apolinario, M., Askin, E., Barish, P., Cheng, M., Deardorff, W. J., Donthi, N., Ganeshan, S., Huang, O., Kantor, M. A., Lai, A., Manchanda, A., Moore, K., Muniyappa, A., Nair, G., Patel, P. P., Santhosh, L., & Rosner, B. (2024). *Physician- And Large Language Model-Generated Hospital Discharge Summaries: A Blinded, Comparative Quality and Safety Study*. https://doi.org/10.1101/2024.09.29.24314562

Yeung, J. A., Kraljević, Ž., Luintel, A., Balston, A., Idowu, E., Dobson, R., & Teo, J. (2023). AI Chatbots Not Yet Ready for Clinical Use. *Frontiers in Digital Health*, *5*. https://doi.org/10.3389/fdgth.2023.1161098