Mesopotamian Journal of Artificial Intelligence in Healthcare
Vol.2024, **pp**. 69–75
DOI: https://doi.org/10.58496/MJAIH/2024/010 ; ISSN: 3005-365X
https://mesopotamian.press/journals/index.php/MJAIH

Brief Communication

Benchmarking Generative AI: A Call for Establishing a Comprehensive Framework and a Generative AIQ Test

Malik Sallam[1,2,*], Roaa Khalil[1], Mohammed Sallam[3]

[1] *Department of Pathology, Microbiology and Forensic Medicine, School of Medicine, The University of Jordan, Amman, Jordan.*

[2] *Department of Clinical Laboratories and Forensic Medicine, Jordan University Hospital, Amman, Jordan.*

[3] *Department of Pharmacy, Mediclinic Parkview Hospital, Mediclinic Middle East, Dubai, United Arab Emirates.*

ARTICLE INFO

Article History

Received 15 Apr 2024
Accepted 08 Jun 2024
Published 02 Jul 2024

Keywords

Artificial intelligence
Clinical decision-making
Benchmarking
GPT-4
ChatGPT

ABSTRACT

The introduction and rapid evolution of generative artificial intelligence (genAI) models necessitates a refined understanding for the concept of "intelligence". The genAI tools are known for its capability to produce complex, creative, and contextually relevant output. Nevertheless, the deployment of genAI models in healthcare should be accompanied appropriate and rigorous performance evaluation tools. In this rapid communication, we emphasizes the urgent need to develop a "Generative AIQ Test" as a novel tailored tool for comprehensive benchmarking of genAI models against multiple human-like intelligence attributes. A preliminary framework is proposed in this communication. This framework incorporates miscellaneous performance metrics including accuracy, diversity, novelty, and consistency. These metrics were considered critical in the evaluation of genAI models that might be utilized to generate diagnostic recommendations, treatment plans, and patient interaction suggestions. This communication also highlights the importance of orchestrated collaboration to construct robust and well-annotated benchmarking datasets to capture the complexity of diverse medical scenarios and patient demographics. This communication suggests an approach aiming to ensure that genAI models are effective, equitable, and transparent. To maximize the potential of genAI models in healthcare, it is important to establish rigorous, dynamic standards for its benchmarking. Consequently, this approach can help to improve clinical decision-making with enhancement in patient care, which will enhance the reliability of genAI applications in healthcare.

1. INTRODUCTION

Integrating the generative artificial intelligence (genAI) tools into healthcare practice marks a paradigm shift, with enhancement in the technological role to improve the quality of patient care [1, 2]. These genAI tools offer possible improvements across various aspects of healthcare including the diagnosis, management, and prevention of diseases [1, 3, 4]. GenAI tools are characterized by the ability to create diverse range of content (e.g., text, images). Additionally, these tools are capable of data predictions. Hence, genAI tools can massively impact several core aspects of healthcare [5-7]. For example, genAI tools in the field of radiology can help in analyzing complex radiologic images with precision and speed beyond the reach of human radiologists [8-10]. A practical illustration of AI capabilities in radiology has been shown through the use of AI algorithms to detect early-stage cancers and to interpret complex scans in neurology [11-13]. Of note, genAI tools' ability gains momentum over time through learning from thousands of previous cases to become more reliable in detecting subtle abnormalities that might be overlooked by the human eye [14].

GenAI models such as ChatGPT, Bing, and Gemini can revolutionize diagnostics, personalized medicine, and patient interactions [15]. However, the recognized limitations of genAI models should be highlighted and addressed particularly in the context of healthcare as follows [1, 4]. First, the AI-generated content depends on the training data; thus, the quality, diversity, and volume of training data would significantly impact genAI models' performance and reliability [16]. Consequently, the use of incomplete or biased datasets to train genAI models can lead to biased predictions and decisions [17, 18]. This is of particular concern in healthcare settings [1]. For example, if an AI-based diagnostic tool is trained on data retrieved predominantly from certain demographic groups, the AI performance would be less effective for the underrepresented groups [19].

*Corresponding author. Email: malik.sallam@ju.edu.jo

Second, the genAI models are described as "black boxes", which denotes that the content generated by AI algorithms lacks transparency [1, 20]. This lack of transparency would represent a significant barrier in healthcare settings, where the rationale behind diagnostic or treatment recommendations should be understood for both trust and accountability [1, 21]. Thus, physicians and patients may become hesitant to rely on AI recommendations without a clear understanding of how the conclusions were derived, especially in critical or life-threatening scenarios [22-24].

Third, the integration of genAI models in the routine healthcare mandates complex regulations and rigorous standards in order to protect patients' safety and privacy [5, 25]. The compliance of genAI models with the strict healthcare regulations (e.g., Health Insurance Portability and Accountability Act on patient data privacy, Food and Drug Administration guidelines for medical devices), can be a challenge for these models [26]. Additionally, the rapid evolution of genAI capabilities could exceed the pace of development of corresponding regulation frameworks. Subsequently, this would lead to potential gaps in genAI oversight in healthcare practice [27, 28].

Finally, genAI models must consistently show high reliability in performance in order to be trusted with clinical decisions. Variability in genAI models' performance, which could be related to variable training data or different operational conditions, would result in inconsistent clinical outcomes. Subsequently, this would undermine trust in genAI models and further complicate the utility of AI in healthcare [1, 29].

Based on the aforementioned points suggesting the complex features of genAI models, it is necessary to establish a comprehensive and meticulous benchmarking framework to ensure the AI safety, efficacy, and ethical compliance [30]. Safety is of particularly important in healthcare, where the risks associated with incorrect or unsafe AI-generated recommendations can have ominous consequences [31]. Safety of genAI in healthcare entails validation of AI performance across diverse patient demographics and variable disease presentation. The failure to perform rigorous testing for safety of genAI in healthcare could lead to patient harm, loss of trust in medical AI applications, and legal consequences for healthcare providers [1, 32].

Most of all, the reliability of genAI models' performance in healthcare must be demonstrated in real-world [1, 33]. This requires conducting longitudinal research investigations to compare AI-assisted interventions against evidence-based best practices in healthcare. Another important aspect that should be taken into consideration in AI integration in healthcare is the ethical compliance. The ethical compliance involves genAI operation without bias as well as the performance respecting patients' autonomy and privacy. To ensure the ethical compliance of genAI in healthcare, benchmarking is required to detect and prevent biases which might arise from biased training datasets. For example, a genAI model initially developed for genetic disorders' screening requires evaluation to make sure it does not prioritize or disadvantage patients' groups based on age, ethnicity, or socioeconomic status. Additionally, ethical genAI benchmarking must guarantee the privacy standards, which involves handling sensitive in a secure manner. Furthermore, genAI benchmarking involves providing transparency to patients and healthcare providers regarding how AI recommendations are generated. A notable example for AI benchmarking frameworks is illustrated by MedPerf, which was introduced by Alexandros Karargyris *et al*., which benchmarks medical AI models via federated evaluation across different facilities [34]. This approach would help to ensure the privacy and efficiency via direct assessment of AI model performance using real-world data under human supervision [34].

Despite the absence of a consensus definition, "intelligence" can be defined based on the ability to comprehend complex concepts, adapt to varying conditions, learn from experience, employ reasoning, and to handle challenges using cognitive processes [35]. Thus, intelligence among humans is recognized as a dynamic, multi-dimensional construct which can be viewed as a broad reflection of human cognitive abilities [36].

Due to limitations in genAI models' adaptability compared to the human brain, the measurement and benchmarking of the currently available genAI models represent a challenging task. Specifically, the human brain is characterized by the ability to process a wide range of sensory data, adapt to new environments, and learn from minimum information [37]. On the other hand, genAI models require large training datasets to learn and can struggle to make rational generalizations beyond the models' training conditions [38]. This aspect in particular emphasizes the need for sophisticated benchmarking tools to effectively measure the cognitive capabilities of genAI models, and to ensure that these AI models are both robust and versatile.

This brief communication aimed to highlight the urgent need to develop a Generative AI Quotient Test (genAIQ), in order to enable the robust evaluation of the cognitive genAI capabilities. The genAIQ aims to provide a standardized approach to evaluate how the genAI tools perform tasks typically associated with human intelligence, such as learning, reasoning, and problem-solving, under varied and unexpected conditions. For example, a genAI model trained in diagnostic radiology should excel in recognizing the conditions it has been explicitly trained on, and also demonstrate adaptability to handle rare, ambiguous, or complex cases that differ significantly from those presented in training data.

The call for establishing a preliminary framework to rigorously evaluate genAI, as discussed in this communication, involves the incorporation of diverse concepts of human intelligence. This includes the basic cognitive processes as well as the aspects

of emotional and social intelligence, which are can challenging for genAI tools to emulate. In particular, the ability to understand and interpret human emotions can be crucial for AI applications in healthcare settings, where empathetic communication and sensitivity to the patients' emotional states are crucial.

Establishing a comprehensive framework for benchmarking would ensure that genAI tools are technically competent, safe, and effective in real-world settings. This involves continuous revision and updates of the benchmarking tools to keep pace with advances in AI technology. Ultimately, the development of a genAIQ test and its associated framework represents a critical step towards the responsible integration of AI technologies into sectors where human-like cognitive abilities are essential, which would ensure that these AI tools augment human capabilities without causing collateral damage.

2. Suggested Metrics for a Benchmarking Framework for GenAI Models

Given the rapid advances in genAI models, there is an urgent need to develop a framework to benchmark its performance, especially in healthcare settings [1, 4, 39]. This initiative involves the establishment of specific performance metrics to evaluate the effectiveness of genAI applications. These metrics are proposed as foundational elements for future comprehensive frameworks aiming for rigorous assessment of genAI capabilities, and to ensure that these models are effective, equitable, and successfully integrated into clinical practice (**Figure 1**).

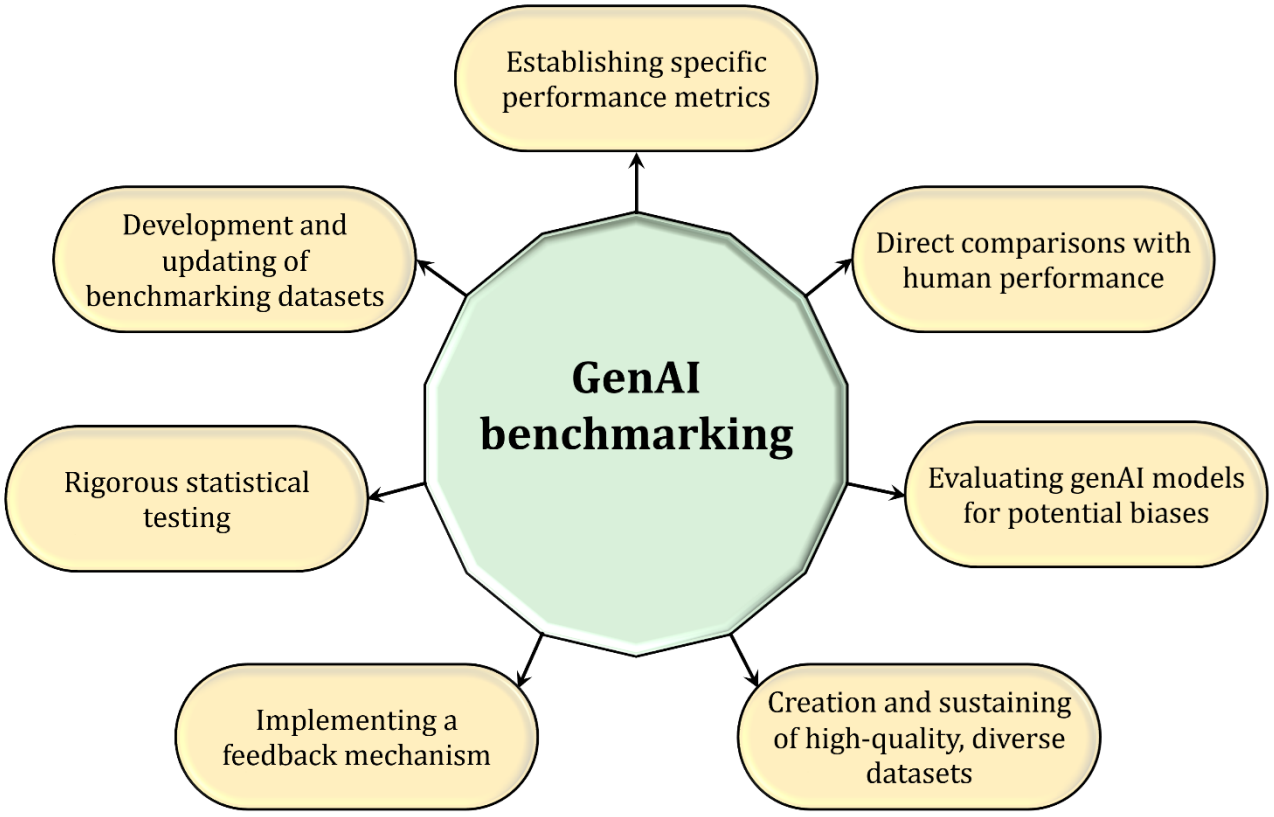

Fig. 1. A proposed framework for benchmarking generative artificial intelligence (genAI).

The first performance metric is the accuracy metric which would help to assess how AI-generated output aligns with the strict standards approved by healthcare experts. This should cover varying aspects such as the diagnostic accuracy and reliability of treatment recommendations by genAI models. The second performance metric involves diversity which is another crucial metric to be considered, focusing on the genAI's ability to handle a broad spectrum of clinical scenarios. This would reflect the complexity and variety of patient cases in clinical practice, which would help to ensure that genAI models are flexible and robust enough to meet the standards in healthcare.

Third, novelty should also be considered as a performance metric, which involves measurement of genAI ability to produce new insights and novel solutions beyond the training data. This metric aims to support the development of genAI models that introduce innovative solutions in patient care; thus, enhancing the existing medical practices. Consistency is also a critical performance metric, evaluating the reliability of genAI models across various cases and conditions. This metric aims

to ensure that genAI outputs are consistent and reliable, which would aid in genAI integration into routine clinical practice. Finally, the speed and efficiency of genAI models must be assessed among the performance metrics. This metric is particularly important to assess genAI ability to effectively operate in real-time clinical settings.

The second aspect necessary for genAI benchmarking involves the development of comprehensive, annotated datasets. These datasets should accurately represent real-world patients' demographics and various medical conditions. Additionally, these datasets must be updated regularly to incorporate the emerging health conditions and disease trends.

Third, genAI models must undergo rigorous benchmarking against the current best practices in healthcare and benchmarking against prior AI models. Additionally, the benchmarking should include direct comparisons with human performance. This process is important to assess whether genAI can meet or even surpass the established gold standards of patient care.

Forth, the employment of rigorous statistical methods is crucial to determine the significance of performance differences among genAI models. This would help to ensure that any observed improvements or failures are not merely due to random variations in the data.

Fifth, evaluating genAI models for potential biases is essential to ensure safe outcomes. It is equally important to maintain transparency in genAI decision-making processes and ensure that these AI models can be understood and reviewed with ease by clinicians.

Sixth, implementing a feedback mechanism that utilizes real-world application data to continuously train and benchmark genAI models is critical. This ensures that AI remains adaptive to new information and consistently aligns with evolving healthcare standards.

Finally, a significant challenge in benchmarking genAI is the creation and sustaining of high-quality, diverse datasets free from biases. The subjective nature of certain medical judgments further complicates the standardization of benchmarks. As AI technologies advance, the frameworks used for their evaluation must also evolve, necessitating continuous adaptation of benchmarking standards.

3. Suggestion for the Development of a Generative AIQ Test

The integration of genAI into cognitive assessment represents both a remarkable opportunity and a complex challenge. A suggestion is the exploration of a genAIQ test that would measure genAI models' cognitive abilities, similar to assessments utilized to benchmark human intelligence. This concept would involve evaluating whether genAI can produce correct solutions and demonstrate creativity, adaptability, and a deep understanding of varying contexts. The development of such a test would require a comprehensive approach that mirrors the complex nature of human intelligence and acknowledges the unique capabilities and limitations of genAI.

Reflecting on the established theories of human intelligence, the development of a genAIQ test requires a comprehensive evaluation of AI across various cognitive domains. This includes assessment of problem-solving and reasoning abilities, particularly how genAI handle complex, unfamiliar problems and to propose effective solutions. Additionally, the test should measure the genAI model learning and adaptability in response to new information, its creativity and innovation in generating novel ideas absent from training data, and its capability to comprehend and contextualize instructions within a specific framework.

Methodologically, the creation of a genAIQ test involves important considerations. The design of test items should challenge genAI models across a broad range of cognitive abilities, including logic, abstract reasoning, pattern recognition, and creative thinking. Unlike static assessments, a genAIQ test should incorporate real-time data to adapt based on the genAI responses, to enable a deeper exploration of the AI cognitive processes. Establishing relevant benchmarks and a fair scoring system that accounts for the diversity of AI capabilities is also crucial.

However, the development of such a test is fraught with technical and ethical challenges. Technically, the relevance of the test across various AI forms with potentially different architectures must be ensured, adding layers of complexity. Ethically, it is important to avoid embedding human biases within the AI, especially biases related to cultural or contextual interpretations.

4. CONCLUSIONS

As genAI continues to advance with expanding influence, the establishment of a robust and comprehensive benchmarking framework becomes imperative. This framework is essential for the evaluation of genAI models across multiple dimensions, to mirror the complex and adaptive qualities of human intelligence that these models aim to replicate. Rigorous testing and continuous monitoring of these genAI models is required in order to enable healthcare professionals to integrate genAI into their practices with confidence, thereby enhancing patient outcomes while upholding the highest standards of care and ethical responsibility.

This communication has briefly outlined the critical components of a genAI benchmarking framework and advocates for its swift development and implementation. The creation of a genAIQ test stands at the forefront of AI and cognitive psychology, offering a method to measure the intellectual capabilities of genAI models thoroughly and accurately. This process is essential for improving the understanding, utilization, and governance of genAI, to ensure that these models operate reliably, safely, and ethically.

Furthermore, the implementation of a genAIQ test could revolutionize the development, deployment, and trust in genAI models across various sectors, particularly in healthcare. It would offer a standardized approach for comparing different genAI models, and enhance our comprehension of genAI evolving capabilities and ensuring they meet ethical standards and practical demands.

Funding

The author's paper clearly indicates that the research was conducted without any funding from external sources.

Conflicts Of Interest

The author's disclosure statement confirms the absence of any conflicts of interest.

Acknowledgment

AI Tools are Used to Edit and Proofread This Paper.

References

[1] M. Sallam, "ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns," (in eng), *Healthcare (Basel)*, vol. 11, no. 6, p. 887, Mar 19 2023, doi: 10.3390/healthcare11060887.

[2] M. Mijwil, R. Doshi, K. Hiran, I. Bala, and A. L. I. Guma, "The Effect of Human-Computer Interaction on New Applications by Exploring the Use Case of ChatGPT in Healthcare Services," in *Modern Technology in Healthcare and Medical Education: Blockchain, IoT, AR, and VR,*, K. Hiran, R. Doshi, and M. Patel Eds. Hershey, PA: IGI Global, 2024, pp. 74-87.

[3] P. P. Ray, "ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope," *Internet of Things and Cyber-Physical Systems*, vol. 3, pp. 121-154, 2023/01/01/ 2023, doi: 10.1016/j.iotcps.2023.04.003.

[4] M. Sallam, A. Al-Farajat, and J. Egger, "Envisioning the Future of ChatGPT in Healthcare: Insights and Recommendations from a Systematic Identification of Influential Research and a Call for Papers," *Jordan Medical Journal*, vol. 58, no. 1, pp. 95-108, 02/19 2024, doi: 10.35516/jmj.v58i1.2285.

[5] S. Reddy, "Generative AI in healthcare: an implementation science informed translational path on application, integration and governance," (in eng), *Implement Sci*, vol. 19, no. 1, p. 27, Mar 15 2024, doi: 10.1186/s13012-024-01357-9.

[6] Z. Lv, "Generative artificial intelligence in the metaverse era," *Cognitive Robotics*, vol. 3, pp. 208-217, 2023/01/01/ 2023, doi: 10.1016/j.cogr.2023.06.001.

[7] M. M. Raza, K. P. Venkatesh, and J. C. Kvedar, "Generative AI and large language models in health care: pathways to implementation," *npj Digital Medicine*, vol. 7, no. 1, p. 62, 2024/03/07 2024, doi: 10.1038/s41746-023-00988-4.

[8] S. Srivastav *et al.*, "ChatGPT in Radiology: The Advantages and Limitations of Artificial Intelligence for Medical Imaging Diagnosis," (in eng), *Cureus*, vol. 15, no. 7, p. e41435, Jul 2023, doi: 10.7759/cureus.41435.

[9] M. F. Russe *et al.*, "Performance of ChatGPT, human radiologists, and context-aware ChatGPT in identifying AO codes from radiology reports," *Scientific Reports*, vol. 13, no. 1, p. 14215, 2023/08/30 2023, doi: 10.1038/s41598-023-41512-8.

[10] P. Keshavarz *et al.*, "ChatGPT in radiology: A systematic review of performance, pitfalls, and future perspectives," *Diagnostic and Interventional Imaging*, vol. In Press, 2024/04/27/ 2024, doi: 10.1016/j.diii.2024.04.003.

[11] B. Hunter, S. Hindocha, and R. W. Lee, "The Role of Artificial Intelligence in Early Cancer Diagnosis," (in eng), *Cancers (Basel)*, vol. 14, no. 6, p. 1524, Mar 16 2022, doi: 10.3390/cancers14061524.

[12] H. Silva *et al.*, "The use of artificial intelligence tools in cancer detection compared to the traditional diagnostic imaging methods: An overview of the systematic reviews," (in eng), *PLoS One*, vol. 18, no. 10, p. e0292063, 2023, doi: 10.1371/journal.pone.0292063.

[13] A. M. J. Zubair Rahman *et al.*, "Advanced AI-driven approach for enhanced brain tumor detection from MRI images utilizing EfficientNetB2 with equalization and homomorphic filtering," *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, p. 113, 2024/04/30 2024, doi: 10.1186/s12911-024-02519-x.

[14] Y. Kumar, A. Koul, R. Singla, and M. F. Ijaz, "Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda," (in eng), *J Ambient Intell Humaniz Comput*, vol. 14, no. 7, pp. 8459-8486, 2023, doi: 10.1007/s12652-021-03612-z.
[15] P. Lee, S. Bubeck, and J. Petro, "Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine," (in eng), *N Engl J Med*, vol. 388, no. 13, pp. 1233-1239, Mar 30 2023, doi: 10.1056/NEJMsr2214184.
[16] N. Nagarajan, "A comprehensive review of AI's dependence on data," *International Journal of Artificial Intelligence and Data Science*, vol. 1, no. 1, pp. 1-11, 2024, doi: 10.13140/RG.2.2.27033.63840.
[17] L. Belenguer, "AI bias: exploring discriminatory algorithmic decision-making models and the application of possible machine-centric solutions adapted from the pharmaceutical industry," (in eng), *AI Ethics*, vol. 2, no. 4, pp. 771-787, 2022, doi: 10.1007/s43681-022-00138-8.
[18] N. Norori, Q. Hu, F. M. Aellen, F. D. Faraci, and A. Tzovara, "Addressing bias in big data and AI for health care: A call for open science," (in eng), *Patterns (N Y)*, vol. 2, no. 10, p. 100347, Oct 8 2021, doi: 10.1016/j.patter.2021.100347.
[19] D. Ueda *et al.*, "Fairness of artificial intelligence in healthcare: review and recommendations," *Japanese Journal of Radiology*, vol. 42, no. 1, pp. 3-15, 2024/01/01 2024, doi: 10.1007/s11604-023-01474-3.
[20] E. A. M. van Dis, J. Bollen, W. Zuidema, R. van Rooij, and C. L. Bockting, "ChatGPT: five priorities for research," (in eng), *Nature*, vol. 614, no. 7947, pp. 224-226, Feb 2023, doi: 10.1038/d41586-023-00288-7.
[21] L. L. d. Souza *et al.*, "ChatGPT and medicine: a potential threat to science or a step towards the future?," *Journal of Medical Artificial Intelligence*, vol. 6, 2023, doi: 10.21037/jmai-23-70.
[22] L. Petersson *et al.*, "Challenges to implementing artificial intelligence in healthcare: a qualitative interview study with healthcare leaders in Sweden," *BMC Health Services Research*, vol. 22, no. 1, p. 850, 2022/07/01 2022, doi: 10.1186/s12913-022-08215-8.
[23] R. Yokoi, Y. Eguchi, T. Fujita, and K. Nakayachi, "Artificial Intelligence Is Trusted Less than a Doctor in Medical Treatment Decisions: Influence of Perceived Care and Value Similarity," *International Journal of Human–Computer Interaction*, vol. 37, no. 10, pp. 981-990, 2021/06/15 2021, doi: 10.1080/10447318.2020.1861763.
[24] A. Martinho, M. Kroesen, and C. Chorus, "A healthy debate: Exploring the views of medical doctors on the ethics of artificial intelligence," *Artificial Intelligence in Medicine*, vol. 121, p. 102190, 2021/11/01/ 2021, doi: 10.1016/j.artmed.2021.102190.
[25] M. Mohammad Amini, M. Jesus, D. Fanaei Sheikholeslami, P. Alves, A. Hassanzadeh Benam, and F. Hariri, "Artificial Intelligence Ethics and Challenges in Healthcare Applications: A Comprehensive Review in the Context of the European GDPR Mandate," *Machine Learning and Knowledge Extraction*, vol. 5, no. 3, pp. 1023-1035, 2023, doi: 10.3390/make5030053.
[26] C. Wang, S. Liu, H. Yang, J. Guo, Y. Wu, and J. Liu, "Ethical Considerations of Using ChatGPT in Health Care," (in eng), *J Med Internet Res*, vol. 25, p. e48009, Aug 11 2023, doi: 10.2196/48009.
[27] D. Yim, J. Khuntia, V. Parameswaran, and A. Meyers, "Preliminary Evidence of the Use of Generative AI in Health Care Clinical Services: Systematic Narrative Review," (in eng), *JMIR Med Inform*, vol. 12, p. e52073, Mar 20 2024, doi: 10.2196/52073.
[28] D. Leslie and A. Perini, "Future Shock: Generative AI and the International AI Policy and Governance Crisis," *Harvard Data Science Review*, vol. Special Issue 5, 05/31 2024, doi: 10.1162/99608f92.88b4cc98.
[29] H. A. Younis *et al.*, "A Systematic Review and Meta-Analysis of Artificial Intelligence Tools in Medicine and Healthcare: Applications, Considerations, Limitations, Motivation and Challenges," (in eng), *Diagnostics (Basel)*, vol. 14, no. 1, p. 109, Jan 4 2024, doi: 10.3390/diagnostics14010109.
[30] P. P. Ray, "Benchmarking, ethical alignment, and evaluation framework for conversational AI: Advancing responsible development of ChatGPT," *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, vol. 3, no. 3, p. 100136, 2023/09/01/ 2023, doi: 10.1016/j.tbench.2023.100136.
[31] A. Choudhury and O. Asan, "Role of Artificial Intelligence in Patient Safety Outcomes: Systematic Literature Review," (in eng), *JMIR Med Inform*, vol. 8, no. 7, p. e18599, Jul 24 2020, doi: 10.2196/18599.
[32] Y. Mu and D. He, "The Potential Applications and Challenges of ChatGPT in the Medical Field," (in eng), *Int J Gen Med*, vol. 17, pp. 817-826, 2024, doi: 10.2147/ijgm.S456659.
[33] T. Dave, S. A. Athaluri, and S. Singh, "ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations," (in eng), *Front Artif Intell*, vol. 6, p. 1169595, 2023, doi: 10.3389/frai.2023.1169595.
[34] A. Karargyris *et al.*, "Federated benchmarking of medical artificial intelligence with MedPerf," (in eng), *Nat Mach Intell*, vol. 5, no. 7, pp. 799-810, Jul 2023, doi: 10.1038/s42256-023-00652-2.
[35] S. Legg and M. Hutter, "A Collection of Definitions of Intelligence," presented at the Proceedings of the 2007 conference on Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms: Proceedings of the AGI Workshop 2006, 2007.
[36] L. Rinaldi and A. Karmiloff-Smith, "Intelligence as a Developing Function: A Neuroconstructivist Approach," (in eng), *J Intell*, vol. 5, no. 2, p. 18, Apr 29 2017, doi: 10.3390/jintelligence5020018.

[37] L. Zhao *et al.*, "When brain-inspired AI meets AGI," *Meta-Radiology*, vol. 1, no. 1, p. 100005, 2023/06/01/ 2023, doi: 10.1016/j.metrad.2023.100005.
[38] H. Naveed *et al.*, "A comprehensive overview of large language models," *arXiv preprint arXiv:2307.06435*, 2023, doi: 10.48550/arXiv.2307.06435.
[39] M. Sallam, "Bibliometric top ten healthcare-related ChatGPT publications in the first ChatGPT anniversary," *Narra J*, vol. 4, no. 2, p. e917, 2024, doi: 10.52225/narra.v4i2.917.