Journal of Scientometric Research, 2025; 14(1):46-61.
https://www.jscires.org

Research Article

Large Language Models in Biomedicine and Health: A Holistic Evaluation of the Effectiveness, Reliability and Ethics using Altmetrics

Prema Nedungadi[1,*], Hiran Haridas Lathabai[2], Raghu Raman[3]

[1]Amrita School of Computing, Amrita Vishwa Vidyapeetham, Amritapuri, Kerala, INDIA.
[2]Amrita CREATE, Amrita Vishwa Vidyapeetham, Amritapuri, Kollam, Kerala, INDIA.
[3]Amrita School of Business, Amrita Vishwa Vidyapeetham, Amritapuri, Kollam, Kerala, INDIA.

ABSTRACT

This study investigated the application of Large Language Models (LLMs), particularly ChatGPT and Bard, in biomedical and clinical sciences and health sciences via a combination of the altmetric approach and scientometric approach. This study analyzes Altmetric scores across various journals and FoRs, focusing on 942 publications since the launch of the ChatGPT in November 2022 (up to 13 November 2023) in Biomedical and Clinical Sciences and Health Sciences. Key findings highlight the growing impact of LLMs in the biomedical and health sciences, as evidenced by high Altmetric Attention Scores. Discussions revolve around ethical issues such as data privacy, AI biases and LLMs' role at the intersection of computational linguistics, AI and healthcare. The findings underscore the potential and challenges of LLMs in healthcare, emphasizing the need for enhanced accuracy, reliability and social acceptance. This study not only presents current trends and impacts but also provides key recommendations for advancing LLM technology, fostering interdisciplinary collaboration and establishing robust validation and regulatory frameworks to successfully integrate LLMs in biomedicine and health. These insights are crucial for guiding future advancements in healthcare research and practice.

Keywords: LLM, Fields of Research, Altmetrics, Ethics, Reliability, Public health, Gender, Social media generative AI, Data privacy, AI bias.

Correspondence:
Prema Nedungadi
Amrita School of Computing,
Amrita Vishwa Vidyapeetham,
Amritapuri-690525, Kollam, Kerala, INDIA.
Email: prema@amrita.edu

Received: 08-02-2024;
Revised: 13-05-2024;
Accepted: 30-11-2024.

INTRODUCTION

The integration of Artificial Intelligence (AI), notably through Large Language Models (LLMs) such as ChatGPT and Bard, has led to a significant shift in the biomedical and healthcare landscape. These sophisticated models reshape clinical decision-making and patient interactions, demonstrating their increasing relevance across diverse applications. The range from diagnosing orthopedic diseases to enhancing disease detection from Electronic Health Records (EHRs) highlights their increasing relevance in healthcare.[1] The influence of ChatGPT spans various sectors, influencing practices in finance, medical examinations, research authoring, student engagement with AI technologies, AI-assisted medical education, gender, cybersecurity, ethics in research, human resource management, sustainable development and higher education.[2-12] The interdisciplinary nature of AI in medicine encourages progress in healthcare,[13] precision medicine and public services.[14] These models, derived from extensive data, merge diverse, intelligent tasks and medical expertise, suggesting the future of collaborative AI and medical practice.[15] This necessitates strong ethical and regulatory frameworks for integrating ChatGPT into clinical workflows.[16,17] The deployment of LLMs in biomedical contexts shows promise for enhancing clinical and research productivity despite challenges in ensuring accuracy[18,19] and addressing safety concerns.[20,21] Ethical considerations remain crucial, particularly in comprehensive AI implementation in healthcare.[22,23] This calls for a systematic analysis of LLMs for biomedical applications, not just any aspect-specific review. The body of knowledge associated with literature accrued or rapidly piled by different fields of science and technology or subfields of such fields due to varying intensity and pace of publishing and patenting activities provides a plethora of opportunities for such an analysis. Despite being in its early stage, a large volume of literature is available on LLMs such as ChatGPT, especially because of its tremendous biomedical appeal. Thus, a systematic analysis of the body of literature related to major LLM tools and their role and impact on biomedicine and health can be useful for revealing various insightful implications that may benefit various stakeholders,

DOI: 10.5530/jscires.20251166

including practitioners, researchers and policymakers. As there have been no such notable attempts thus far, we intend to address that gap in this work and the novelty of this research lies.

Scientometrics is a field of research that addresses the assessment of the progress of different fields as well as the impact assessment of various actors (or contributors) of science and technology via techniques and methods of science and technology. It offers several methods for mining and analysis of the body of scientific literature as well as for impact assessment of scholarly publications. However, there are certain perceived limitations to the scientometric approach for impact assessment. As the scientific impact of published works may take some time to manifest, traditional scientometric indicators such as citation counts are not suitable for early impact assessments of publications. Additionally, those indicators are not capable of capturing the societal impact of such works. In this era of expanding social media and online platform usage, there are means for assessing the attention garnered by published works on various social media platforms and other forums. Scientometricians are contemplating a fluid analogy between (i) the relationship between indicators such as citation to scientific impact and (ii) social media attention/visibility to societal impact, which led to the development of the field 'Altmetrics', which stands for alternative metrics and is often dubbed Scientometrics 2.0. Moreover, altmetrics, which is emerging as an innovative impact measurement tool alongside traditional citation counts, provides a detailed view of research impact, reflecting community engagement in scientific discourse.[24-27] As both approaches can complement each other, a diligent combination of the traditional scientometric approach and the altmetric approach can be useful for the effective analysis of the body of literature related to different fields, subfields, etc., especially for cases such as the one we have to address.

This served as our motivation to employ the traditional scientometric approach and Altmetric approach to analyze the role and impact of LLMs in biomedicine and health. It assesses the engagement and influence of LLM-related research in these fields, using Altmetric data for insights into their effectiveness, reliability and ethical implications. Specifically, the pursuit of the following three key Research Questions (RQs) that can be addressed via the judicious combination of traditional scientometric approaches and Altmetric analysis that we adopt in this work will be beneficial for the abovementioned stakeholders and for providing a sense of direction for further research.

RQ1: Evaluating the influence of LLM research on public health and health services.

RQ2: Undertaking a Fields of Research (FoR) map and cluster analysis focusing on effectiveness, reliability and ethical considerations.

RQ3: Investigating the role of social media in shaping public perception and academic engagement.

METHODOLOGY

As our 1st objective (RQ1) was to extract insights into the effectiveness, reliability, ethical considerations and overarching impact of LLMs in healthcare at a very early stage, Altmetric data analysis is needed. Thus, RQ1 is addressed via the identification of journals with top Altmetric scores (sum of Altmetric scores of individual articles). The identification of top notch journals (those that managed to receive substantial early attention), the identification of dominant areas in LLM healthcare research, the analysis of the top articles that gained the most attention (those with the most AAS garnering articles) and the identification of fields of research associated with those publications are the keys for identifying works that might effectively discuss the abovementioned aspects that can be major determinants of LLM usage in medicine and healthcare. The science mapping approach (scientometric approach) can address RQ2. In this approach, a FoR map associated with the body of literature related to LLMs needs to be created. This is augmented by a cluster analysis delineating the interconnections between various research fields, thereby partitioning the FoR map into discernible clusters. Subsequently, a content analysis of the leading publications within each thematic cluster was conducted to extract more specific insights that can inform various key stakeholders. In addition, it is important to analyze the dominant sources (such as geographical regions) that constitute the social media activities that shaped the early pattern of social media attention in LLM research (i.e., RQ3). This issue is addressed by analyzing the geographical patterns of Twitter (a platform that dominantly contributes to Altmetric score computation) users. The pursuit of the three RQs is achieved after the selection of a suitable database that indexes published works that form the body of literature related to major LLM usage in biomedical and healthcare.

Our major data source for collecting publication data related to ChatGPT and Bard is the Dimensions database.[28] We used the search terms (chatgpt OR "chat GPT" OR chat-GPT OR gpt3* OR gpt4*) OR ("bard")) in the title and abstract of the publications, as ChatGPT product versions and Bard are the most promising LLMs found thus far. All publication types were included except Preprints. The data were collected on November 14, 2023. Publications in 2022 and 2023 up to that day were covered.

Of the 4968 publications retrieved, we focused on 942 publications belonging to two major Fields of Research (FoRs), Biomedical and Clinical Sciences and Health Sciences, using the ANZSRC 2020 code.[28] We collected the altmetric details of the 942 publications using the DOI from Altmetric.com. The Altmetric Attention Score (AAS) is a composite score of mentions (that represent attention) in social media and other platform mentions. We also identified the most frequently mentioned source titles for these publications. The thematic clusters within FoRs are identified using cocitation maps via VOSviewer software, which enables

the visualization and analysis of scientific landscapes, revealing patterns and relationships within research fields.[29]

RESULTS AND DISCUSSION

The results obtained by applying our framework to the selected data (942 publications) related to LLM for biomedicine and health for addressing three specified RQs are discussed next.

RQ1: Influence on public health and health services

As mentioned above, to address RQ1, we determined the journals that received the highest Altmetric Attention Score (sum of the AASs of individual publications) from the dataset of 942 selected publications. The top 10 journals with the highest AAS are shown in Table 1. Upon analysis of the top AAS-receiving publications and FoR mapping, the profound linkage of LLMs to diverse fields of research in biomedical and clinical sciences and health sciences and their potential applications are revealed. The impact of LLM research on public health policies and health services is evident in the diverse range of journals featured in this research. Various areas/aspects of LLM influence in biomedical and healthcare research and practice, as evident from publications with top AASs in the top 10 journals with the highest AAS sums, are summarized below:

Validation and Transparency: The paper in JAMA Internal Medicine (i.e., the 1st paper in Table 1 and the only one in our dataset), "Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum",[30] which falls under Health Sciences (specifically Health Services and Systems), underscores the importance of transparency and the rigorous validation of AI tools in healthcare, reflecting the medical community's focus on the accuracy and ethical implications of AI interactions with patients.

Impact of AI in Different Medical Fields

Second, Radiology features the paper "Appropriateness of Breast Cancer Prevention and Screening Recommendations Provided by ChatGPT",[31] which belongs to Biomedical and Clinical Sciences. This paper focuses on the role of AI in diagnostic processes and preventive medicine, highlighting the potential impact of AI on patient care in radiology.

Collaboration across Disciplines

The 4th publication in Table 1 titled "ChatGPT and Generating a Differential Diagnosis Early in an Emergency Department Presentation"[32] in the Annals of Emergency Medicine, also in the Biomedical and Clinical Sciences field, reflects the urgency of integrating AI in emergency medicine and a strong commitment toward collaborative AI applications that can potentially enhance diagnostic accuracy in time-sensitive medical settings.

Educational Potential

In the interdisciplinary journal PLOS Digital Health, the paper "Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models"[33] (i.e., the 5th one in Table 1), which spans information and computing sciences as well as biomedical and clinical sciences, demonstrates interest in utilizing AI for educational purposes, particularly in medical training and assessment.

Performance in Specialized Tests

The 6th article in Table 1, featured in the Journal of Medical internet features "Assessing the Utility of ChatGPT Throughout the Entire Clinical Workflow: Development and Usability Study".[17] which falls under Health Sciences (Health Services and Systems), assesses LLM's ability to support clinical decision-making and highlights limitations such as possible model hallucinations.

The 9th work in Table 1 is published in the American Journal of Gastroenterology, "Chat Generative Pretrained Transformer Fails the Multiple-Choice American College of Gastroenterology Self-Assessment Test",[34] which is categorized under Biomedical and Clinical Sciences and enumerates the challenges that AI faces in specialized medical fields and the need for domain-specific AI training.

Regulatory Oversight

The 8th paper in Table 1 is from npj Digital Medicine, "The imperative for regulatory oversight of large language models (or generative AI) in healthcare",[35] which falls under Health Sciences (Health Services and Systems) and calls for attention to the need for governance in the use of AI in healthcare, resonating with the community's concern for responsible implementation.

Misinformation Risks and Public Health Communication

The 3rd paper in Table 1 is an interdisciplinary paper titled "Abstracts written by ChatGPT Fool Scientists",[36] which appeared in Nature and cuts across Biological Sciences, Language, Communication and Culture and reflects on concerns about the reliability of scientific communication in the age of AI and the importance of discerning AI-generated content.

JAMA Network Open features the 7th work in Table 1 titled "Evaluating Artificial Intelligence Responses to Public Health Questions",[37] which is about health sciences and public health, indicating a significant interest in the accuracy and reliability of AI in disseminating public health information.

Vaccine Safety Communication

Finally, the paper "Chatting with ChatGPT to learn about the safety of COVID-19 vaccines-A perspective" in human vaccines and immunotherapeutics addresses the biomedical and clinical sciences fields such as immunology, medical biotechnology

Table 1: Journals with a High Altmetric Attention Score (AAS).

Journal Name	Count of Papers	Sum of Altmetric Attention Score	Top publication based on AAS	Fields of Research of the top publication
JAMA Internal Medicine	1	6036	Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum (AAS=6036).	42 Health Sciences 4203 Health Services and Systems.
Radiology	15	1969	Appropriateness of Breast Cancer Prevention and Screening Recommendations Provided by ChatGPT. (AAS=945).	32 Biomedical and Clinical Sciences 3202 Clinical Sciences.
Nature	2	1848	Abstracts written by ChatGPT fool scientists (AAS=1848).	31 Biological Sciences 47 Language, Communication and Culture 4703 Language Studies.
Annals of Emergency Medicine	4	1656	ChatGPT and Generating a Differential Diagnosis Early in an Emergency Department Presentation (AAS=1602).	32 Biomedical and Clinical Sciences 3202 Clinical Sciences.
PLOS Digital Health	1	1542	Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models (AAS=1542).	46 Information and Computing Sciences 32 Biomedical and Clinical Sciences 3202 Clinical Sciences.
Journal of Medical Internet Research	13	1410	Assessing the Utility of ChatGPT Throughout the Entire Clinical Workflow: Development and Usability Study (AAS=902).	42 Health Sciences 4203 Health Services and Systems.
JAMA Network Open	4	1182	Evaluating Artificial Intelligence Responses to Public Health Questions (AAS=713).	42 Health Sciences 4206 Public Health.

Journal Name	Count of Papers	Sum of Altmetric Attention Score	Top publication based on AAS	Fields of Research of the top publication
npj Digital Medicine	4	585	The imperative for regulatory oversight of large language models (or generative AI) in healthcare (AAS=252).	42 Health Sciences 4203 Health Services and Systems.
American Journal of Gastroenterology	3	426	Chat Generative Pretrained Transformer Fails the Multiple-Choice American College of Gastroenterology Self-Assessment Test. (AAS=374).	32 Biomedical and Clinical Sciences 3202 Clinical Sciences.
Human vaccines Immunotherapeutics	1	370	Chatting with ChatGPT to learn about safety of COVID-19 vaccines-A perspective (AAS=370).	32 Biomedical and Clinical Sciences 3204 Immunology 3206 Medical Biotechnology 3207 Medical Microbiology.

and medical microbiology and highlights AI's potential role in public health dialogs, especially concerning vaccine safety and misinformation.[38]

Thus, the top 10 articles according to altmetric attention scores are found to shed light on the influence of LLMs on public health and health services and the imperative factors that contribute to these influences, such as effectiveness/performance, transparency, reliability, ethical and regulatory considerations and risks and safety issues. Therefore, RQ1 is successfully addressed via the use of an altmetric approach and content analysis.

RQ2: Fields of research map and cluster analysis

The Fields of Research map (Figure 1), which is a coaffiliation map of FoRs (coupled via publications in our dataset), represents the interlinkage between various FoRs associated with the body of literature analyzed. In the map, the size of the vertices (FoRs) indicates high linkage to other vertices (FoRs). Upon clustering, the map delineates two principal clusters: Biomedical and Clinical Sciences (green cluster) and Health Sciences (red cluster). Analysis of both these clusters and the top 10 publications from the list of publications that are responsible for the formation of green clusters and red clusters addresses RQ2. These issues are discussed next.

Cluster 1 Biomedical and clinical sciences (green)

As already mentioned, a heavy dominance of the FoR 'Biomedical and Clinical Sciences' can be seen in the green cluster. Other FoRs, such as 'Biological Sciences' and 'Engineering', are also found to indicate the ever-increasing interconnectedness between life sciences and engineering sciences disciplines. The major sub-FoRs that are deeply influenced by LLM research seem to include 'Clinical Sciences', 'Immunology', 'Ophthalmology and optometry' and 'Dentistry'. Some sub-FoRs of the FoR 'Health Sciences', such as 'Allied Health and Rehabilitation Science' and 'Sports Science and Exercise', are also found in the green cluster (while FoR Health Sciences dominates the red cluster). This indicates the deep connectedness between 'Biomedical and Clinical Sciences' and 'Health Sciences' and the ability of LLMs and AI to influence these major FoRs individually and simultaneously.

The top 10 publications in the biomedical and clinical sciences clustered according to the Altmetrics Attention Score, along with their further relevant details, are presented in Table 2. In Table 2, the FoR column predominantly shows works in clinical sciences, highlighting the significant focus on the practical applications and implications of LLMs in various medical fields. This interdisciplinary paper combines information and computing sciences with clinical sciences, emphasizing the growing intersection of AI technology and health care. The range

of disciplines extends to Immunology, Medical Biotechnology, Medical Microbiology, Ophthalmology and Optometry and Oncology and Carcinogenesis, showing that LLMs have varied impacts across different biomedical and clinical areas. They illustrate how LLMs are being explored for their potential in medical education, diagnostic processes and specific areas such as vaccine safety and cancer research.

According to the 1st publication in Table 2, the ChatGPT achieved a passing score or nearly moderate performance in the USMLE,[33] indicating that LLMs will soon mature enough to impact clinical medicine and that such a scenario demands open science research infrastructure in medical education.

The 2nd publication in Table 2 underscored that, in some cases, ChatGPT generated incorrect answers or explanations,[32] indicating that inconsistency in outputs underscores the unpredictability of LLMs and highlighting the need to restrict the role of LLMs as an aid rather than replace physicians' judgment.

The 3rd work suggested that although some improvement is needed, the LLM holds great potential for automating patient educational information about breast cancer prevention and screening[27] and recommended physician supervision throughout the process.

Through a comparative analysis, the 4th work revealed that GPT-4 significantly outperformed GPT-3.5 and Bard, with a lower rate of hallucinations observed for GPT-4 patients,[39] indicating the potential of LLMs in neurosurgical education and clinical decision-making.

An experiment reported in the 5th work revealed that ChatGPT provided a confident, human-like answer[38] and suggested that the use of technology, especially LLMs, can heavily impact medical writing, particularly in terms of speed and accuracy.

The 6th work[34] involved a comparative analysis of two versions of the ChatGPT on gastroenterology-related examination and found that both versions failed the test (as the pass score required was 70%), suggesting the need for substantial improvement of LLMs before implementation in healthcare.

The 7th work revealed the overall ability of ChatGPT to detect counterfeit questions related to vaccines and vaccination[40] and highlighted that the current language used is optimal (fewer technical terms without compromising scientific details), suggesting that the potential LLM is reliable medical information.

According to the 8th paper, LLMs such as the ChatGPT cannot substantially assist in training for board certification in ophthalmology,[24] as the performance of the ChatGPT was

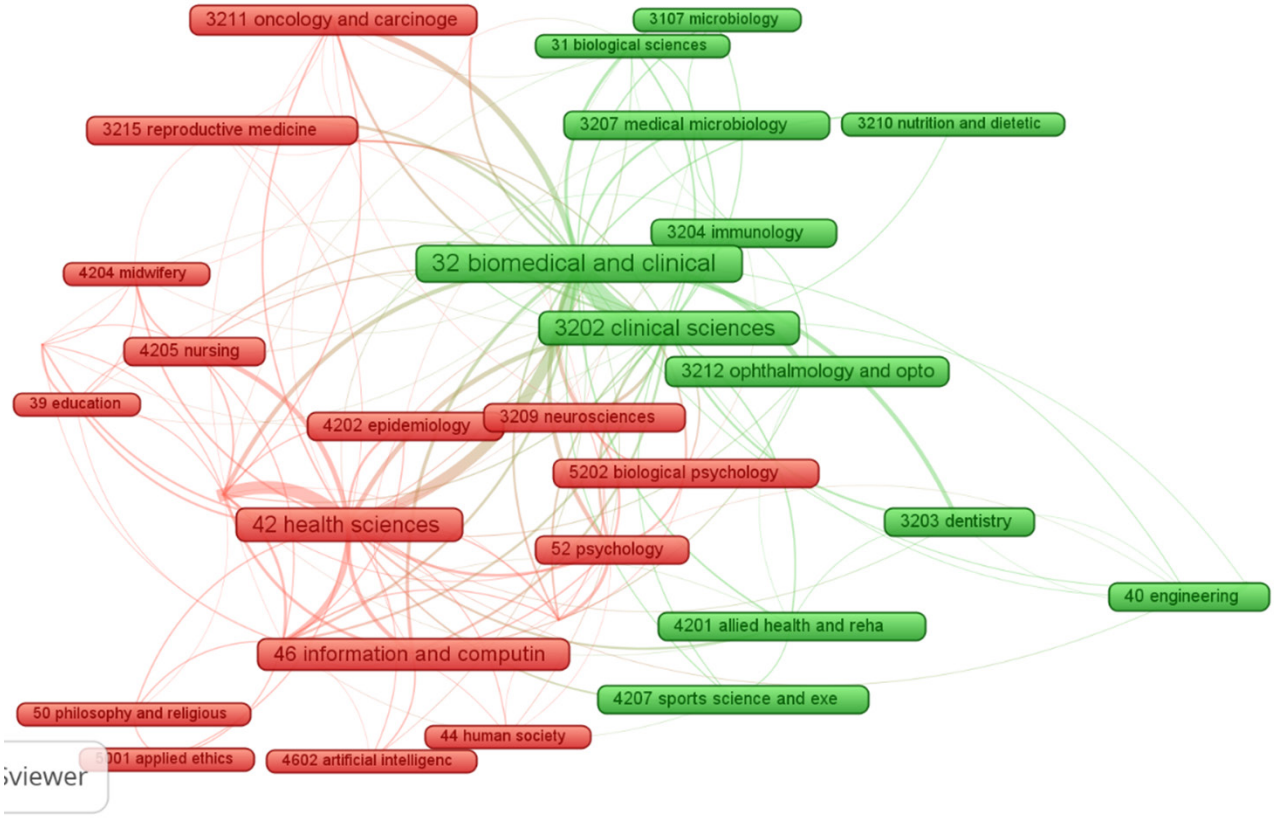

Figure 1: Fields of the research map showing two major clusters.

Table 2: Top 10 Publications in Biomedical and Clinical Sciences based on Altmetric Attention Scores (AASs).

Title of the paper	AAS	Objectives/ Research focus	Nature of Work	Specific Medical Applications Discussed (if any)	Methodology/ Investigation details	Ethical Concerns Discussed (if any)	Benchmarking (if any)/ Dataset/tools used/generated
Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models.	3957	Assessment of Performance of ChatGPT in USMLE.	Evaluation of the Potential of ChatGPT.	Medical Education.	Open-Ended (OE) prompting, Multiple choice single answer without forced justification (MC-NJ) prompting, Multiple choice single answer with forced justification (MC-J) prompting.	Nil	Performance benchmark of near passing score of AI LLM on USMLE established.
ChatGPT and Generating a Differential Diagnosis Early in an Emergency Department Presentation.	1602	Investigation of the ability of ChatGPT to generate accurate differential diagnoses in undifferentiated patients based on physician notes recorded at the initial ED presentation.	Evaluation of the Potential of ChatGPT.	Medical/Clinical diagnosis (Emergency).	Comparison by 1-way ANOVA (confidence interval determined by bootstrapping) of differential diagnosis and leading diagnosis by doctors before laboratory tests and adjusted differential and leading diagnosis after laboratory tests with output of ChatGPT (obtained by threefold entry of each case).	Emphasized considering ethical and legal requirements by LLMs for use in a medical setting.	Performance benchmark of 97% (ChatGPT v3.5) accuracy in diagnosis with laboratory data.
Appropriateness of Breast Cancer Prevention and Screening Recommendations Provided by ChatGPT.	949	Evaluation of ChatGPT for Breast Cancer Prevention and Screening.	Evaluation of the potential of ChatGPT.	Medical/Clinical diagnosis (Breast Radiology).	The latest methodology for the evaluation of cardiovascular diseases was followed. Triple entry of 25 questions to ChatGPT and rating responses as 'appropriate,' 'inappropriate,' and 'unreliable' by fellowship-trained breast radiologists.	Highlighted ChatGPT's role as a research Chatbot, not intended for medical use.	Breast Imaging-Reporting and Data System (BI-RADS) Atlas[8] Performance score of 88% appropriateness.
Performance of ChatGPT, GPT-4 and Google Bard on a Neurosurgery Oral Boards Preparation Question Bank.	565	Performance assessment of three LLMs (GPT-3.5, GPT-4 and Google Bard) on a question bank designed for neurosurgery oral board examination preparation.	Evaluation potential of LLMs (comparison).	Medical Education/Clinical Decision-making (Neurosurgery).	149-question Self-Assessment Neurosurgery Exam (SANS) Indications Exam in a single best answer MCQ format. The results were compared using the Chi-squared test, Fisher's Exact test and Univariable logistic regression test.	The potential ethical and legal implications of using LLMs should be carefully considered.	Initial benchmark (with 82.6% accuracy) in high-order and relatively open-ended clinical scenarios. Developed methods to quantify hallucinations.

Title of the paper	AAS	Objectives/ Research focus	Nature of Work	Specific Medical Applications Discussed (if any)	Methodology/ Investigation details	Ethical Concerns Discussed (if any)	Benchmarking (if any)/ Dataset/tools used/generated
ChatGPT and the Future of Medical Writing.	389	Assessment of capability of ChatGPT for medical writing.	Evaluation of the potential of ChatGPT.	Medical writing/ Education (Radiology).	Essay question about how well ChatGPT is trained in Radiology.	Raised cautions related to ethics, legal issues (including medicolegal issues), innovation, accuracy, bias and transparency for using ChatGPT for medical writing.	Nil
Chat Generative Pretrained Transformer Fails the Multiple-Choice American College of Gastroenterology Self-Assessment Test.	374	Assessment of Performance of ChatGPT on Multiple-Choice American College of Gastroenterology Self-Assessment Test.	Evaluation of the potential of ChatGPT for medical application.	Medical Education (Gastro-enterology).	Multiple Choice Questionnaire for 2021 and 2022. Comparison of ChatGPT-3 and ChatGPT-4	Recommended against using ChatGPT for medical education in gastroenterology in its current form.	ChatGPT-3 scored highest with 65.1%.
Chatting with ChatGPT to learn about the safety of COVID-19 vaccines-A perspective.	370	Assessment of ChatGPT's capacity to generate opinions on vaccine hesitancy.	Evaluation of the potential of ChatGPT for information accuracy.	Information science (Vaccine safety).	Fifty frequently asked questions about misconception, false contradictions and true contradictions. Analysis of responses by professional specialists in the field.	The present-day version of ChatGPT cannot replace expert or scientific evidence *per se*.	Accuracy of 85.5%.
Performance of an Artificial Intelligence Chatbot in Ophthalmic Knowledge Assessment.	324	Assessment of performance of ChatGPT in practice questions for board certification in Ophthalmology.	Evaluation of the potential of ChatGPT for medical application.	Medical Education (Ophthalmology).	125 text-based multiple-choice questions provided by the Ophthalmology Questions free trial for board certification examination preparation.	Nil	A score of 46%
Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma.	297	Investigation of accuracy and reproducibility of ChatGPT in questions regarding knowledge, management and emotional support for cirrhosis and HCC.	Evaluation of the potential of ChatGPT for medical application.	Medical Education/ Healthcare and Management (Cirrhosis and HCC).	There are one hundred sixty-four questions regarding knowledge, management and emotional support for Cirrhosis and HCC. Grading of responses by 2 transplant hepatologists and resolution by a 3rd reviewer. 26 quality measures of cirrhosis management were used to check ChatGPT's knowledge of cirrhosis management. An assessment of emotional support capacity was also conducted.	Nil	76.9% score for quality measures but failed to specify decision-making cutoffs and treatment durations.

Title of the paper	AAS	Objectives/ Research focus	Nature of Work	Specific Medical Applications Discussed (if any)	Methodology/ Investigation details	Ethical Concerns Discussed (if any)	Benchmarking (if any)/ Dataset/tools used/generated
Assessment of Artificial Intelligence Chatbot Responses to Top Searched Queries About Cancer.	271	Assessment of quality of information and presence of misinformation generated by 4 AI chatbots approximately 5 major types of cancer.	Evaluation of the potential of AI chatbots for medical application.	Medical Education (oncology)	Most commonly searched queries related to 5 types of cancers from the Google Trends platform in 2 years (2021 and 2002). The DISCERN instrument and Patient Education Materials Assessment Tool (PEMAT) were used to grade primary outcomes. Secondary outcomes such as misinformation were graded through a 5-item Likert Scale and readability was assessed through the Flesch-Kincaid Grade Level readability score.	Nil	66.7% for understandability (moderate), 20% for actionability (poor) and readability was at the college level.

the best in general medicine and the poorest in the retina and vitreous.

The 9th paper revealed that the ChatGPT performed well in terms of basic knowledge, lifestyle and treatment,[41] but a lack of knowledge about regional guideline variations (such as HCC screening criteria) is an area that should be properly worked upon and recommended to restrict its usage as an adjunct informational tool.

The 10th paper also identified the potential of ChatGPT for producing accurate results, but as responses were not actionable and were found to be at the college reading level,[42] it recommended its use as a supplementary tool (not as a primary source of medical information).

Cluster 2 Health sciences (red)

In the red cluster, FoRs such as Information and Computing, Psychology, Education, etc., have the highest linkages to other FoRs or sub-FoRs. This indicates the profound importance and possible impact that LLMs can have on FoRs related to medical research and practice. The presence of FoR Information and Computing Sciences highlights how the future of medical research and practice is dependent on the advancement of information, computing and learning (AI) technologies. Nursing, epidemiology, etc., are the dominant sub-FoRs within the FoR Health Sciences that are influenced by LLM advancement. In FoR psychology, sub-FoR 'biological psychology' is the subfield in which the LLM has the greatest impact at this early stage. The sub-FoR 'Neurosciences' (a subfield of Clinical Sciences) is also found in this cluster, showing its profound linkage to FoRs such as 'Health Sciences' and 'Information and Computing Sciences'.

Table 3 shows that the top publications in health sciences predominantly feature papers classified under 'Health Services and Systems. This focus reflects an academic and practical interest in understanding how AI tools can integrate into and enhance healthcare services, from clinical workflows to regulatory oversight. Notably, several studies have compared AI-generated responses with those of healthcare professionals, assessing the accuracy, reliability and utility of the ChatGPT in clinical settings. This highlights the growing curiosity about the potential role of AI in augmenting or assisting medical decision-making processes. Papers such as "Evaluating Artificial Intelligence Responses to Public Health Questions" also extend the exploration to public health, indicating a broader scope of LLM applications beyond individual patient care to community health issues.

Furthermore, studies on the role of AI in generating clinical vignettes, scientific abstracts and even the creation of discharge summaries point to the diverse and innovative ways LLMs are utilized. These studies suggest that AI has transformative potential for reshaping various facets of healthcare service delivery and medical communications, opening doors to new methods of patient engagement, medical education and healthcare management. This trend underscores the importance of continued research and regulatory considerations as AI becomes increasingly embedded in health services systems.

The 1st paper in Table 3 shows that the LLM is competent in generating quality, empathetic and more detailed (than physicians) responses to queries in online forums,[30] indicating

Table 3: Top 10 Publications in Health Sciences based on Altmetric Attention Scores (AASs).

Title of the paper	AAS	Objectives/ Research focus	Nature of Work	Specific Medical Applications Discussed (if any)	Methodology/ Investigation details	Ethical Concerns Discussed (if any)	Benchmarking (if any)/ Dataset/ tools used/ generated
Chatbot Responses to Patient Questions Posted to a Public Social Media Forum.	6094	Assessment of the potential of ChatGPT to assist in answering patient questions with quality and empathy comparable to that of physicians.	Evaluation of the potential of AI chatbots for medical application	Medical/ Clinical assistance	Comparison of responses by Chatbot and Physicians on 195 randomly drawn questions from patients on social media forums. Triplicate evaluation by a team of licensed healthcare professionals.	Nil	78.6% score for preference to Chatbot response over physician response. The score for quality by Chatbot was 78.5% (against 22.1% by physicians) The score for empathy by Chatbot was 45.1% (against 4.1% by physicians).
Assessing the Utility of ChatGPT Throughout the Entire Clinical Workflow: Development and Usability Study.	912	Assessment of ChatGPT's capacity for ongoing clinical decision support.	Evaluation of ChatGPT for medical applications.	Medical/ Clinical decision support.	36 Clinical vignettes published in Merck Sharpe and Dohme (MSD) Clinical Manual and compared accuracy concerning various clinical aspects based on age, gender and sensitivity. Linear regression was also conducted to assess contributing factors toward ChatGPT's clinical task performance.	Cautioned against model hallucination and unclear composition of ChatGPT's training dataset	Overall score of 71.7%. Accuracy score of 76.9% for final diagnosis. The lowest accuracy was for the aspect differential diagnosis, which was 60.3%.
Evaluating Artificial Intelligence Responses to Public Health Questions.	713	Assessment of the potential of ChatGPT's ability to respond to health inquiries by the lay public.	Evaluation of ChatGPT for healthcare applications.	Health care support	Twenty-three questions (in 4 categories) were used to evaluate ChatGPT. The methodology was followed by Miner *et al.* (2016).[26] Nobles *et al.* (2020)[28] and STROBE reporting guidelines. Questions used a common help-seeking structure. An automated readability index was used. Analyses were done in the R tool.	AI companies should be encouraged to use government-recommended resources. It is recommended that the government limit the liability of AI companies using government-recommended resources.	Accuracy of 91%.
Identifying depression and its determinants upon initiating treatment: ChatGPT versus primary care physicians.	369	Comparative evaluation of the ability of ChatGPT versions 3.5 and 4 for depressive episodes and suggested protocols with that of primary care physicians.	Evaluation of ChatGPT for primary health care.	Medical/ Clinical applications (Primary care).	Vignettes containing symptoms of depression of hypothetical patients during initial treatment were used. Consistency and reliability were tested by repeating the experiment 10 times.	Recommended further research related to potential risks and ethical issues for the usage of AI in primary care.	Nil

Title of the paper	AAS	Objectives/ Research focus	Nature of Work	Specific Medical Applications Discussed (if any)	Methodology/ Investigation details	Ethical Concerns Discussed (if any)	Benchmarking (if any)/ Dataset/ tools used/ generated
ChatGPT for Clinical Vignette Generation, Revision and Evaluation.	273	Assessment of the ability of ChatGPT to generate, rewrite and evaluate clinical vignettes.	Evaluation of ChatGPT for medical writing.	Medical Writing and Editing	Natural language prompting for the generation of 10 sets of 10 vignettes For common childhood illness By providing symptoms. Additionally, a prompt for rewriting 15 existing pediatric vignettes was done. Fifteen vignettes generated from the parent's perspective are rewritten at the physician's level, 8th-grade reading level, etc. and again from the parent's perspective.	Cautioned against caveats including unpredictability, tendency to generate references to nonexistential scholarly article.	75.6% 1st passage diagnostic accuracy. 57.8% triage accuracy.
The imperative for regulatory oversight of large language models (or generative AI) in healthcare.	260	Review of regulations for usage of LLMs in healthcare and suggestion of recommendations.	Perspective on LLM usage in healthcare	Medical/ Clinical applications	Systematically underscored the difference between other AI technologies and LLMs. Reviewed existing (Pre-LLM) AI regulations. Listed out possible applications/use cases of LLM for medical practitioners and patients. Finally, enumerated regulatory challenges for LLMs.	Highlighted the necessity of using a proactive approach to regulation to address involved risks and ethical challenges.	Nil
Accuracy and Reliability of Chatbot Responses to Physician Questions.	244	Assessment of accuracy and comprehensiveness of chatbot-generated responses to physician-developed medical queries and to determine the reliability and limitations of–generated medical information.	Evaluation of ChatGPT's ability for medical/clinical assistance.	Medical/ clinical assistance	Two hundred eighty-four questions generated by 33 physicians across 17 specialties were used. Physicians graded responses on a 6-point Likert scale for accuracy and on a 3-point Likert scale for comprehensiveness. Descriptive statistics, the Mann-Whitney U test and the Kruskal-Wallis test were used for analysis.	Further research and models were developed recommended for overcoming concerns related to ethical, transparency, data security and privacy and medicolegal aspects.	More than 50% (nearly all correct or completely correct)
Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers.	193	Assessment of capability of ChatGPT for scientific writing (medical).	Evaluation of ChatGPT for medical/ scientific writing.	Medical/ scientific writing	Selected five abstracts of articles published in high-impact medical journals. Provided titles and journal names of same to ChatGPT for generating abstracts. Generated and original abstracts were reviewed by a GPT-2 output detector and human experts to detect fake ones.	Recommended usage of AI output decoders as useful editorial tools.	Nil.

Title of the paper	AAS	Objectives/ Research focus	Nature of Work	Specific Medical Applications Discussed (if any)	Methodology/ Investigation details	Ethical Concerns Discussed (if any)	Benchmarking (if any)/ Dataset/ tools used/ generated
Artificial Intelligence Can Generate Fraudulent but Authentic-Looking Scientific Medical Articles: Pandora's Box Has Been Opened.	190	Assessment of capabilities of ChatGPT 3 in generating high-quality fraudulent medical articles.	Evaluation of ChatGPT for medical/ scientific writing.	Medical/ Scientific writing (Neurosurgery, Psychiatry and Statistics).	I posted questions to ChatGPT and refined responses iteratively to generate fabricated medical articles. Experts in Neurosurgery, Psychiatry and Statistics reviewed accuracy and coherence.	Highlighted the need to be vigilant about the misuse of AI in scientific research. Recommended development and usage of better detection methods, ethical guidelines and best practices to combat the potential misuse of AI.	Nil
ChatGPT: the future of discharge summaries?	143	Discussion of ChatGPT's ability to generate discharge summaries.	Commentary of ChatGPT for medical/health care assistance.	Medical/health care assistance	Nil	Nil	Deep consideration should be given for every aspect, including privacy and security issues related to the medical data of patients, technology failure issues, etc., for adopting AI tools in healthcare.

its ability to significantly reduce the burnout of physicians and suggesting further exploration of the use of LLMs in clinical settings.

The 2nd study reported that the ChatGPT performed well in terms of general medical knowledge,[17] but its performance is inferior for differential diagnosis and clinical management and highlights the ability to improve accuracy for more clinical information as the major strength of LLMs.

An investigation in the 3rd work revealed that the ChatGPT consistently provided evidence-based answers[43] and recommended encouragement of public-private partnerships for the use of AI and LLMs in healthcare wherein public health agencies should develop and maintain database resources for AI companies to use.

A comparative analysis of the ChatGPT with physicians carried out *in the* 4th work revealed that the ChatGPT showed no gender or socioeconomic biases in response to the recommendations[44] and confirmed the congruence of the ChatGPTs 3.5 and 4 with the accepted guidelines for managing mild and severe depression.

The 5th work reported the results of an analysis of LLM's performance in medical writing[45] and as LLMs demonstrated versatility in generating results customizable to the literacy levels of patients, their ability to perform medical writing and editing is indicated, though not without supervision, due to caveats.

The 6th work recommended the development of a specific and distinct regulatory framework for LLMs (rather than using a common framework for LLMs and non-LLM AI technologies) and such a framework should target the regulation of companies developing LLMs rather than the regulation of every iteration of LLMs.[46]

The 7th work revealed the ability of LLMs to generate accurate and comprehensive information[47] and enumerated limitations such as hallucination, inability to seek clarifications for ambiguous queries and tendency to make inaccurate citations (that questions transparency).

The 8th work reported the results of an examination of LLM-generated abstracts and original abstracts by human reviewers[48] and confirmed the ability of the ChatGPT to write seemingly believable scientific abstracts and highlighted the need for careful discussions to determine ethical considerations.

According to the 9th work, ChatGPT produced fraudulent papers that may convince researchers, even experienced ones,[21] though

there were inaccurate and nonexistent references, duplicate citations, etc. and emphasized the importance of effective and responsible harnessing of LLMs.

The 10th work recommended that ChatGPT be upgraded to autoscrap medical information from patients' digital records by sufficiently addressing data privacy concerns[49] after (i) sufficient pilot testing, (ii) gathering stakeholders' perspectives and (iii) careful considerations of technology failure strategies.

Thus, FoR map analysis revealed important FoRs and interlinkages between FoRs belonging to two major clusters. Additionally, the presence of sub-FoRs of FoRs belonging to the red cluster in the green cluster and vice versa indicates the interdependence of the two major clusters and the possible ability of LLMs to impact both clusters together. In-depth analysis of both clusters selected for 10npublications (according to the AAS) also reinforced this fact and such an exercise helped to garner useful insights related to almost all the imperative aspects related to LLMs, such as accuracy and performance, diverse applicability, reliability, transparency, other limitations and ethical considerations. Thus, as envisaged, a combination of a scientometric approach and an altmetric approach helped to successfully address RQ2. The insights gained from the in-depth analysis and possible recommendations based on such insights are discussed in the 'Future directions' section.

RQ3 Impact of social media

Figure 2 shows the notable presence of LLM research discussions on Twitter, underlining the importance of such platforms in spreading and deliberating scientific findings. The tweet analysis points to a conversation dominated by the Global North, particularly the United States, underscoring its leading technological role. This dominance is somewhat less evident but still noticeable in countries such as the United Kingdom, Japan, Canada and various European nations. Conversely, the Global South, represented by nations such as Brazil, India and South Africa, appears less engaged in these AI dialogs, suggesting a disparity in digital and developmental engagement. This situation underscores the imperative for the inclusive development of technologies with far-reaching effects.

Thus, the analysis of the geographical distribution of tweets (major contributors to AAS) provided insights into the existence of global disparities (between the Global North and South) related to the evolution of LLMs, emphasizing the dominance of certain countries or regions in shaping public opinion and perceptions related to technologies, especially those with game-changing potential. This also underlines the importance of altmetric sources and platforms in the progress of LLM research and applications in the biomedical and healthcare fields and points to the need for some nations/regions to consider proactive social media engagement to popularize their research outcomes and to remain updated about advancements. Thus, RQ3 is also effectively addressed by almetric analysis, as expected.

CONCLUSION AND FUTURE DIRECTIONS

This framework, which is based on a combination of scientometric and altmetric data-driven analyses of ChatGPT and Bard's applications in biomedicine and health, offers a comprehensive view of their potential applications, challenges and interdisciplinary impact.

Identified potential benefits of LLMs and opportunities

The study highlights active engagement with LLM research in biomedicine and health, particularly in publications such as JAMA Internal Medicine and Radiology. This study emphasizes the accuracy and performance of LLMs and AI in patient care and diagnostic processes. Science mapping and cluster analysis revealed key FoRs and sub-FoRs that are mostly influenced by LLM research and advancement. Analysis of the top publications

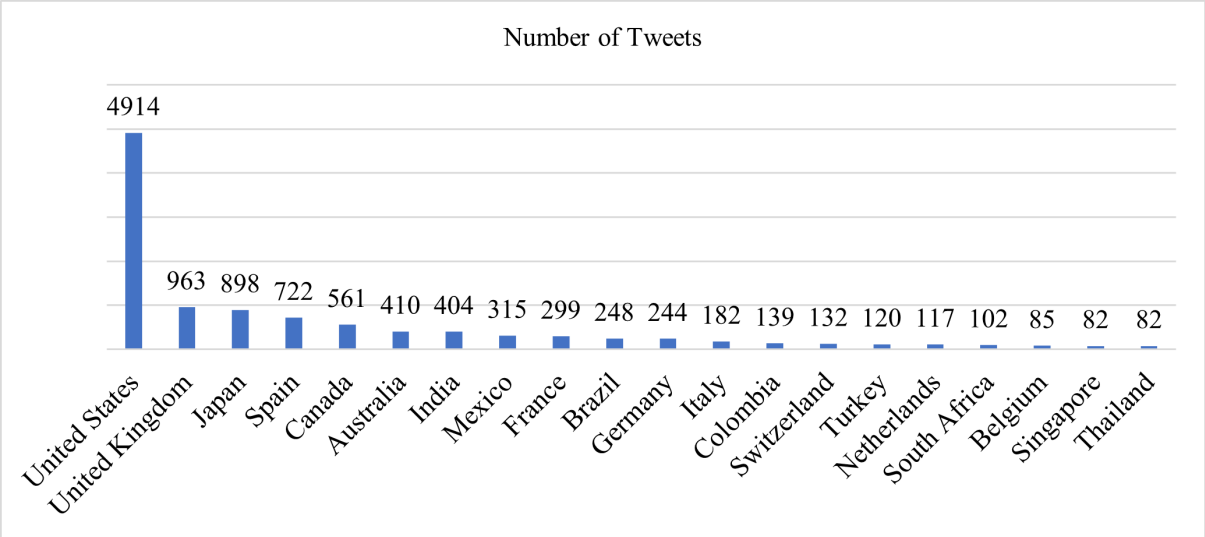

Figure 2: Leading countries based on Tweets.

within two major clusters encompassing ChatGPT's performance on the USMLE[33] and its applications in emergency medicine[32] hints at the potential of LLMs across different functional tiers in these sectors. Scholarly works[50,51] have shown improvements in the accuracy and thoroughness of LLMs, indicating a progression in their adaptation to medical contexts.

Thus, our study revealed the growing influence of LLMs, from emergency medicine to education and regulation; LLMs and AI transform medical training, diagnostics, public health and scientific communication, indicating their ability to reshape healthcare service delivery, education and management. LLMs in clinical sciences and health services and systems are noted for their practical applications and potential to transform healthcare, with research spanning diverse fields, from immunology to oncology.

An updated altmetric analysis revealed significant academic and public interest in LLMs, driven by high attention scores and social media engagement. This widespread interest revolves around LLM efficacy and reliability and their role in bridging the gap between language, AI and healthcare. This integration shapes current research trends, highlighting the potential of LLMs to advance healthcare research and inform future policies.

CHALLENGES AND ETHICAL CONSIDERATIONS

This study also highlights ethical implications identified from various important works identified through altmetric analysis of the whole literature and literature related to two important clusters. Some works have stressed the importance of transparency and the need for the validation of LLMs for healthcare. Some acknowledged its limitations in specialized tests and vulnerability to misinformation. The reliability of the LLMs was also challenged in some works because they are prone to hallucinations and because of their tendency to add inaccurate and/or nonexistent references. Additionally, as LLMs and AI integrate more deeply into health systems, robust research and regulations are crucial. The need to separate LLMs from non-LLM AI applications in medicine and healthcare is also identified as a crucial challenge, but if it is addressed properly, it can be converted into a great plethora of opportunities.

From a critical perspective, these data showcase the promise and the perils of integrating generative AI tools such as ChatGPT and Bard into healthcare. The high attention scores reflect the public and professional community's vested interest in these developments. However, they also call for cautious and considered integration of these technologies, emphasizing validation, transparency and ethical usage to ensure that they truly benefit patient care and outcomes. Given the rapid evolution of AI technologies, continuous monitoring, evaluation and adaptation of these systems are required to ensure that they align with healthcare goals and ethical standards. This integration shapes current healthcare research trends, highlighting the contributions of LLMs to advancing healthcare research and informing future policy-making. The ethical challenges of specific domains, such as dermatology, emphasize the need for varied datasets and conscientious AI deployment in healthcare.[11]

FUTURE DIRECTIONS

Our findings suggest several future research possibilities for applying Large Language Models (LLMs) in diverse healthcare contexts. These prospects are informed by the latest data, observed trends and identified limitations of LLM tools, as revealed in widely referenced studies and various exploratory endeavors. However, as the key limitations contribute immensely to further exploration and development, these limitations are summarized again as follows for quick perusal.

ACCURACY AND RELIABILITY

Ensuring that LLMs consistently and accurately respond to medical inquiries is paramount.

Unpredictability and hallucination: Reducing the generation of incorrect or irrelevant information is crucial for the safe use of LLMs in medical education and clinical practice.

Regional guideline awareness: LLMs must be trained on regional medical guidelines and regulations.

AI output detectors: AI output detectors are critical for identifying and eliminating fictitious, AI-generated scientific content.

Social acceptance: Addressing data privacy and security concerns is essential for building public trust in LLMs in healthcare.

Building upon the findings of our study on the impact of Large Language Models (LLMs), such as ChatGPT and Bard, on biomedicine and health, we propose the following recommendations. These findings are closely aligned with our research insights and aim to enhance the efficacy, ethical application and sustainable integration of LLMs in healthcare environments:

Encouraging public-private partnerships and investment in research infrastructure: Given the study's revelations on the interdisciplinary and significant influence of LLMs and the need for sophisticated research infrastructure, we advocate nurturing Public-Private Partnerships (PPPs) and augmenting research infrastructure. Such PPPs should be structured and operationalized within a robust and comprehensive legal framework that, in turn, is set up on top of a standard ethical framework. Special emphasis should be placed on the sharing and misuse of clinical data, other sensitive information and the structure and nature of the engagement of various entities within PPPs to ensure the safety, security and dignity of people.

Technological enhancement of AI LLMs: The analysis of LLMs' performance underscores the necessity for improvements in accuracy and reliability, which would facilitate their evolution as a source of medical knowledge, an informational tool and potentially as an assistant in diagnosis and decision-making for physicians and health researchers. Special care should be taken to avoid the known, forewarned and unknown caveats of technological overreliance and overuse in biomedical and health research and practice, as these FoRs are directly associated with the life and health of patients. This can be addressed via constant explorations and discussions on aspects wherein technology should be used and not used and fostering technological development and advancement in that direction so that room can be always reserved for human supervision and intervention wherever and whenever necessary. Human-centric technological development is thus desirable for AI and LLMs.

Interdisciplinary collaboration and strategic social media utilization

The findings on substantial engagement with LLM research on social media platforms underscore the importance of strategically using these channels. The varied applications of LLMs highlight the need for cooperative approaches across distinct disciplines. Such collaboration, knowledge exchange and pooling can improve overall LLM performance by reducing limitations such as hallucination tendencies. Such interdisciplinary as well as multinational collaborations might be useful for determining global regulations and standards and at the same time determining locally acceptable exceptions in regional guidelines.

Robust validation processes for LLMs

In line with our findings on high Altmetric scores and engagement levels, we recommend the implementation of stringent validation protocols for LLMs to ensure their dependability and effectiveness, particularly in specialized medical fields. This aligns with the observed interest and engagement in LLMs, emphasizing the necessity for precision and reliability. As proposed in some of the works with high Altmetric Attention Scores, effective validation frameworks can serve as benchmarks and be further refined. The role of interdisciplinary and multinational collaboration can also be vital in this respect and validation exercises based on multiple relevant criteria (as applicable) should be considered.

Policy development and regulatory frameworks

Table 1, which shows journals with high Altmetric Attention Scores, highlights the need for diverse policies to govern LLM integration in healthcare, reflecting the interdisciplinary applications of these policies. Policies should be adaptable; evolving in tandem with LLM tools and include systematic determinations for the equitable use of AI in medical and academic publishing. Regulatory frameworks in strict congruence with ethical and legal frameworks should be established to monitor and enforce preventive and corrective action wherever applicable in the case of materialized or potential ethical breaches as well as legal violations.

ACKNOWLEDGEMENT

We would like to express our immense gratitude to our beloved Chancellor, Mata Amritanandamayi Devi (AMMA), for providing the motivation and inspiration for this research.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

1. Kuroiwa T, Sarcon A, Ibara T, Yamada E, Yamamoto A, Tsukamoto K, *et al.* The potential of ChatGPT as a self-diagnostic tool in common orthopedic diseases: exploratory study. J Med Internet Res. 2023;25:e47621. doi: 10.2196/47621, PMID 37713254.
2. Dowling M, Lucey B. ChatGPT for (finance) research: the Bananarama conjecture. Fin Res Lett. 2023;53:103662. doi: 10.1016/j.frl.2023.103662.
3. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, *et al.* Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLoS Digit Health. 2023;2(2):e0000198. doi: 10.1371/journal.pdig.0000198, PMID 36812645.
4. Lee P, Bubeck S, Petro J. Benefits, limits and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med. 2023;388(13):1233-9. doi: 10.1056/NEJMsr2214184, PMID 36988602.
5. Nedungadi P, Ramesh M, Govindaraju V, Rao B, Berbeglia P, Raman R. Emerging leaders or persistent gaps? Generative AI research may foster women in STEM. Int J Inf Manag. 2024;77:102785. doi: 10.1016/j.ijinfomgt.2024.102785.
6. Raman R. Transparency in research: an analysis of ChatGPT usage acknowledgment by authors across disciplines and geographies. Acc Res. 2023:1-22. doi: 10.1080/08989621.2023.2273377, PMID 37877216.
7. Raman R, Lathabhai H, Mandal S, Kumar C, Nedungadi P. Contribution of business research to sustainable development goals: bibliometrics and science mapping analysis. Sustainability. 2023;15(17):12982. doi: 10.3390/su151712982.
8. Raman R, Venugopalan M, Kamal A. Evaluating human resources management literacy: A performance analysis of ChatGPT and bard. Heliyon. 2024;10(5):e27026. doi: 10.1016/j.heliyon.2024.e27026, PMID 38486738.
9. Raman R, Lathabai HH, Mandal S, Das P, Kaur T, Nedungadi P. ChatGPT: literate or intelligent about UN sustainable development goals? PLOS One. 2024;19(4):e0297521. doi: 10.1371/journal.pone.0297521, PMID 38656952.
10. Raman R, Calyam P, Achuthan K. ChatGPT or Bard: who is a better Certified Ethical Hacker? Comput Sec. 2024;140:103804. doi: 10.1016/j.cose.2024.103804.
11. Raman R, Mandal S, Das P, Kaur T, Sanjanasri JP, Nedungadi P. Exploring University Students' Adoption of ChatGPT Using the Diffusion of Innovation Theory and Sentiment analysis with Gender Dimension. Hum Behav Emerg Technol. 2024 (in press);2024(1). doi: 10.1155/2024/3085910.
12. Stokel-Walker C. ChatGPT listed as author on research papers: many scientists disapprove. Nature. 2023;613(7945):620-1. doi: 10.1038/d41586-023-00107-z, PMID 36653617.
13. Meskó B. The impact of multimodal large language models on health care's future. J Med Internet Res. 2023;25:e52865. doi: 10.2196/52865, PMID 37917126.
14. Kusters R, Misevic D, Berry H, Cully A, Le Cunff Y, Dandoy L, *et al.* Interdisciplinary research in artificial intelligence: challenges and opportunities. Front Big Data. 2020;3:577974. doi: 10.3389/fdata.2020.577974, PMID 33693418.
15. Qiu J, Li L, Sun J, Peng J, Shi P, Zhang R, *et al.* Large ai models in health informatics: applications, challenges and the future. IEEE J Biomed Health Inform. 2023 Sep 22;27(12):6074-87. doi: 10.1109/JBHI.2023.3316750, PMID 37738186.
16. Martinengo L, Lin X, Jabir AI, Kowatsch T, Atun R, Car J, *et al.* Conversational agents in health care: expert interviews to inform the definition, classification and conceptual framework. J Med Internet Res. 2023;25:e50767. doi: 10.2196/50767, PMID 37910153.
17. Rao A, Pang M, Kim J, Kamineni M, Lie W, Prasad AK, *et al.* Assessing the utility of ChatGPT throughout the entire clinical workflow: development and usability study. J Med Internet Res. 2023;25:e48659. doi: 10.2196/48659, PMID 37606976.
18. Pal S, Bhattacharya M, Lee SS, Chakraborty C. A domain-specific next-generation large language model (LLM) or ChatGPT is required for biomedical engineering and research. Ann Biomed Eng. 2024;52(3):451-4. doi: 10.1007/s10439-023-03306-x, PMID 37428337.
19. Thirunavukarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. Nat Med. 2023;29(8):1930-40. doi: 10.1038/s41591-023-02448-8, PMID 37460753.

20. Levkovich I, Elyoseph Z. Identifying depression and its determinants upon initiating treatment: ChatGPT versus primary care physicians. Fam Med Community Health. 2023;11(4):e002391. doi: 10.1136/fmch-2023-002391, PMID 37844967.
21. Májovský M, Černý M, Kasal M, Komarc M, Netuka D. Artificial intelligence can generate fraudulent but authentic-looking scientific medical articles: Pandora's box has been opened. J Med Internet Res. 2023;25:e46924. doi: 10.2196/46924, PMID 37256685.
22. Ferreira AL, Lipoff JB. The complex ethics of applying ChatGPT and language model artificial intelligence in dermatology. J Am Acad Dermatol. 2023;89(4):e157-8. doi: 10 .1016/j.jaad.2023.05.054, PMID 37263382.
23. Kluge EW. Artificial intelligence in healthcare: ethical considerations. Healthc Manag Forum: Los Angeles. 2020;33(1):47-9. Sage CA. doi: 10.1177/0840470419850438, PMID 31340674.
24. Banshal SK, Singh VK, Muhuri PK. Can Altmetric mentions predict later citations? A test of validity on data from ResearchGate and three social media platforms. Online Inf Rev. 2021;45(3):517-36. doi: 10.1108/OIR-11-2019-0364.
25. Banshal SK, Gupta S, Lathabai HH, Singh VK. Power Laws in Altmetric s: an empirical analysis. J Inf. 2022;16(3):101309. doi: 10.1016/j.joi.2022.101309.
26. Priem J, Hemminger BH. Scientometrics 2.0: new metrics of scholarly impact on the social Web. First Monday. 2010 Jul 2. doi: 10.5210/fm.v15i7.2874.
27. Priem J, Parra C, Piwowar HA, Groth P, Waagmeester A. Uncovering impacts: a case study in using Altmetric s tools. Ins EngPublica@ ESWC. 2012 May 28:(40-4).
28. Raman R, Mandal S, Das P, Kaur T, Sanjanasri JP, Nedungadi P. University students as early adopters of ChatGPT: innovation Diffusion Study. doi: 10.21203/rs.3.rs-2734142 /v1.
29. Van Eck NJ, Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics. 2010;84(2):523-38. doi: 10.1007/s11192-009 -0146-3, PMID 20585380.
30. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, *et al.* Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. 2023;183(6):589-96. doi: 10.1001/jamainternmed.2 023.1838, PMID 37115527.
31. Haver HL, Ambinder EB, Bahl M, Oluyemi ET, Jeudy J, Yi PH. Appropriateness of breast cancer prevention and screening recommendations provided by ChatGPT. Radiology. 2023;307(4):e230424. doi: 10.1148/radiol.230424, PMID 37014239.
32. Ten Berg HT, van Bakel B, van de Wouw L, Jie KE, Schipper A, Jansen H, *et al.* ChatGPT and generating a differential diagnosis early in an emergency department presentation. Ann Emerg Med. 2024;83(1):83-6. doi: 10.1016/j.annemergmed.2023.0 8.003, PMID 37690022.
33. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLoS Digit Health. 2023;2(2):e0000198. doi: 10.1371/journal.pdig. 0000198, PMID 36812645.
34. Suchman K, Garg S, Trindade AJ. Chat generative pretrained transformer fails the multiple-choice American College of Gastroenterology self-assessment test. Off J Am Coll Gastroenterol ACG. 2023;118(12):2280-2. doi: 10.14309/ajg.000000000000 2320, PMID 37212584.
35. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. npj Digit Med. 2023;6(1):120. doi: 10.1038/ s41746-023-00873-0, PMID 37414860.
36. Else H. Abstracts written by ChatGPT fool scientists. Nature. 2023;613(7944):423. doi: 10.1038/d41586-023-00056-7, PMID 36635510.
37. Nobles AL, Leas EC, Caputi TL, Zhu SH, Strathdee SA, Ayers JW. Responses to addiction help-seeking from Alexa, Siri, Google Assistant, Cortana and Bixby intelligent virtual assistants. npj Digit Med. 2020;3(1):11. doi: 10.1038/s41746-019-0215-9, PMID 32025572.
38. Salas A, Rivero-Calle I, Martinón-Torres F. Chatting with ChatGPT to learn about safety of COVID-19 vaccines–A perspective. Hum Vaccin Immunother. 2023;19(2):2235200. doi: 10.1080/21645515.2023.2235200, PMID 37660470.
39. Ali R, Tang OY, Connolly ID, Fridley JS, Shin JH, Sullivan PL, *et al.* Performance of ChatGPT, GPT-4 and Google Bard on a neurosurgery oral boards preparation question bank. Neurosurgery. 2022:10-227.
40. Mihalache A, Popovic MM, Muni RH. Performance of an artificial intelligence chatbot in ophthalmic knowledge assessment. JAMA Ophthalmol. 2023;141(6):589-97. doi: 1 0.1001/jamaophthalmol.2023.1144, PMID 37103928.
41. Yeo YH, Samaan JS, Ng WH, Ting PS, Trivedi H, Vipani A, *et al.* Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol. 2023;29(3):721-32. doi: 10.3350/cmh.2023.0089, PMID 36946005.
42. Pan A, Musheyev D, Bockelman D, Loeb S, Kabarriti AE. Assessment of artificial intelligence chatbot responses to top searched queries about cancer. JAMA Oncol. 2023;9(10):1437-40. doi: 10.1001/jamaoncol.2023.2947, PMID 37615960.
43. Ayers JW, Zhu Z, Poliak A, Leas EC, Dredze M, Hogarth M, *et al.* Evaluating artificial intelligence responses to public health questions. JAMA Netw Open. 2023;6(6):e2317517. doi: 10.1001/jamanetworkopen.2023.17517, PMID 37285160.
44. Levkovich I, Elyoseph Z. Identifying depression and its determinants upon initiating treatment: ChatGPT versus primary care physicians. Fam Med Community Health. 2023;11(4):e002391. doi: 10.1136/fmch-2023-002391, PMID 37844967.
45. Benoit JR. ChatGPT for clinical vignette generation, revision and evaluation. MedRxiv. 2023 Feb 8; 02. doi: 10.1101/2023.02.04.23285478.
46. Goodman RS, Patrinely JR, Stone CA, Zimmerman E, Donald RR, Chang SS, *et al.* Accuracy and reliability of chatbot responses to physician questions. JAMA Netw Open. 2023;6(10):e2336483. doi: 10.1001/jamanetworkopen.2023.36483, PMID 37782499.
47. Gao CA, Howard FM, Markov NS, Dyer EC, Ramesh S, Luo Y, *et al.* Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. npj Digit Med. 2023;6(1):75. doi: 10.1038/s41746-023-00819-6, PMID 37100871.
48. Patel SB, Lam K. ChatGPT: the future of discharge summaries? Lancet Digit Health. 2023;5(3):e107-8. doi: 10.1016/S2589-7500(23)00021-3, PMID 36754724.
49. D'Orsi CJ, Sickles EA, Mendelson EB, Morrie EA. ACR BI-RADS® atlas, Breast Imaging Reporting and Data System. Reston, VA: American College of Radiology; 2013.
50. Pugliese N, Wai-Sun Wong V, Schattenberg JM, Romero-Gomez M, Sebastiani G, NAFLD Expert Chatbot Working Group, *et al.* Accuracy, reliability and comprehensibility of ChatGPT-generated medical responses for patients with nonalcoholic fatty liver disease. Clin Gastroenterol Hepatol. 2024;22(4):886-889.e5. doi: 10.1016/j.cgh.2023.08.033, PMID 37716618.
51. Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical considerations of using ChatGPT in health care. J Med Internet Res. 2023;25:e48009. doi: 10.2196/48009, PMID 37566454.

Cite this article: Nedungadi P, Lathabai HH, Raman R. Large Language Models in Biomedicine and Health: A Holistic Evaluation of the Effectiveness, Reliability and Ethics using Altmetrics. J Scientometric Res. 2025;14(1):46-61.