Abu-Jeyyab M, Alrosan S, alkhawaldeh IM. Harnessing Large Language Models in Medical Research and Scientific Writing: A Closer Look to The Future. *High Yield Medical Reviews*. Published online December 1, 2023. doi:10.59707/hymrFBYA5348

Online first

Harnessing Large Language Models in Medical Research and Scientific Writing: A Closer Look to The Future

Mohammad Abu-Jeyyab[1,2], Sallam Alrosan[3], Ibraheem M alkhawaldeh[2]

[1] Red Crescent Hospital, Amman, Jordan, [2] School of Medicine, Mutah University, Al-Karak, Jordan, [3] Internal medicine, Saint Luke's Health System

Keywords: Artificial intelligence, Medical Research, Scientific Writing, Large Language Models

https://doi.org/10.59707/hymrFBYA5348

High Yield Medical Reviews

Large Language Models (LLMs), a form of artificial intelligence that can generate natural language responses based on user input. They have been widely used in various applications such as entertainment, education, and customer service, but their use in the medical field is still nascent and underexplored This review aims to provide a comprehensive overview of the current state-of-the-art and future directions of LLMs for medical research and scientific writing. It covers the following aspects: (1) the data sources and challenges for training and testing LLMs in the medical domain, such as electronic health records, clinical trials, and biomedical literature; (2) the ethical implications and risks of using LLMs for medical purposes, such as privacy, consent, bias, and accountability; (3) the methods and criteria for evaluating the performance and quality of LLMs, such as accuracy, coherence, relevance, and user satisfaction; and (4) the potential applications and benefits of LLMs for various medical tasks and scenarios, such as diagnosis, treatment, patient education, clinical decision support, and scientific writing. The review concludes that LLMs have great potential to enhance the efficiency and quality of medical research and scientific writing, but also emphasize the need for rigorous design, validation, and regulation to ensure their safety and reliability.

INTRODUCTION

Artificial intelligence (AI) is the ability of machines to perform tasks that normally require human intelligence, such as reasoning, learning, and decision-making. AI has been advancing rapidly in recent years, thanks to the availability of large amounts of data, powerful computing resources, and novel algorithms. One of the most prominent forms of AI is natural language processing (NLP), which is the ability of machines to understand and generate natural language. NLP has many applications, such as machine translation[1] sentiment analysis,[2] and text summarization.[3]

One of the most impressive achievements of NLP is the development of Large Language Models (LLMs) which are models that can generate natural language texts based on user input. LLMs can engage in natural and coherent conversations with humans on various topics, as well as generate creative content.[4-6] LLMs are based on deep neural networks, which are complex mathematical models that can learn from data and produce outputs. Some examples of LLMs are GPT-3,[7] BlenderBot,[8] and DialoGPT.[9] LLMs have been widely used for various applications in entertainment, education, and customer service, but their potential in the medical field has not been fully explored. The medical field is a domain that requires high-quality and reliable information and communication, both for research and clinical purposes. Medical research involves conducting experiments, analyzing data, and writing scientific papers. Clinical practice involves diagnosing patients, recommending treatments, and educating patients. Both research and clinical practice require the use of natural language to communicate complex and technical concepts.

Medical chatbots are conversational agents that can interact with users via natural language and provide them with health-related information, advice, diagnosis, or treatment. Medical chatbots have gained popularity in recent years due to their potential benefits for patients, health professionals, and health systems. However, despite the growing interest and development of medical chatbots, there is a lack of systematic reviews that synthesize and evaluate the current trends and challenges of this emerging field. Therefore, in this paper, we aim to fill this gap by conducting a comprehensive and critical review of the existing literature on medical chatbots.

However, applying LLMs in the medical field poses several challenges and opportunities. On one hand, LLMs need to be trained and tested on data sources that reflect the medical domain knowledge and terminology, such as electronic health records,[10] clinical trials,[11] and biomedical literature.[12] These data sources are diverse in terms of format, content, and quality, and require careful preprocessing and filtering to ensure their validity and relevance. On the other hand, LLMs need to adhere to ethical standards and regulations that ensure the privacy, consent, bias, and accountability of the users and patients, such as HIPAA,[13] GDPR,[14] or IRB approval.[15] Moreover, LLMs need to be evaluated using rigorous and relevant methods that measure their accuracy, coherence, relevance, and user satis-

faction. Finally, LLMs need to demonstrate their usefulness and effectiveness for various medical tasks and scenarios that can enhance the quality and efficiency of medical research and scientific writing.

Therefore, this article aims to provide a comprehensive overview of the current state-of-the-art and future directions of LLMs for medical research and scientific writing. It covers the following aspects: (1) the data sources and challenges for training and testing LLMs in the medical domain; (2) the ethical implications and risks of using LLMs for medical purposes; (3) the methods and criteria for evaluating the performance and quality of LLMs; (4) the potential applications and benefits of LLMs for various medical tasks and scenarios. The article is organized as follows: Section 2 examines the different types of data that can be used to train and test LLMs for medical purposes; Section 3 analyzes the ethical implications of using LLMs in the medical field; Section 4 evaluates the different methods that can be used to measure the performance and quality of LLMs for medical research and scientific writing; Section 5 explores the different ways that LLMs can be applied in the medical field, such as diagnosis,[16] treatment,[17] patient education,[18] clinical decision support,[19] and scientific writing[20]; Section 6 concludes the article and provides some directions for future work.

DIFFERENT TYPES OF DATA THAT CAN BE USED TO TRAIN AND TEST LLMS

Large Language Models (LLMs) are based on large neural networks that learn from massive amounts of text data, such as books, websites, and social media posts.[4] Chatbot AI tools can be used for various purposes, such as entertainment, education, customer service, and healthcare.

One of the potential applications of LLMs is in the medical domain, where they can assist physicians and patients with diagnosis,[16] treatment,[17] and information.[18] However, to ensure the quality and reliability of the chat models, they need to be trained and tested on appropriate data sources that reflect medical knowledge and context.

There are different types of data that can be used to train and test LLMs for medical purposes. Some examples are:

- Medical textbooks and journals: These are authoritative sources of medical information that cover various topics, such as anatomy, physiology, pathology, and pharmacology. They can provide chat models with factual knowledge and terminology that are relevant to the medical domain.[5] However, these sources may not capture the latest advances or controversies in the field, and they may not reflect the real-world scenarios or challenges that physicians and patients face.[21]
- Medical question-answering datasets: These are collections of questions and answers that test the medical knowledge and reasoning skills of humans or machines. They can be used to evaluate the chat models' ability to answer complex and specific medical queries. Some examples are MedQA (USMLE exam questions),[4] BioASQ (biomedical literature questions),[5] and NEJM Knowledge+ (board review questions).[6] However, these datasets may not cover all the possible types or formats of questions that users may ask, and they may not provide sufficient feedback or explanations for the answers.[22]
- Medical dialog datasets: These are transcripts or simulations of conversations between doctors and patients or between doctors themselves. They can be used to train and test the chat models' ability to engage in natural and coherent dialogs that involve medical topics, such as symptoms, diagnosis, treatment, follow-up, etc. Some examples are MIMIC-III (critical care dialogs),[7] MedDialog (primary care dialogs),[8] and CoCo (counseling dialogs).[9] However, these datasets may not represent the diversity or variability of the dialog participants, such as their age, gender, language, culture, personality, etc., and they may not capture the emotional or social aspects of the dialogs.[23]
- Medical images: These are visual representations of medical conditions or procedures, such as X-rays, CT scans, MRI scans, and ultrasound images. They can be used to train and test the chat models' ability to process multimodal inputs (text and image) and generate relevant outputs (text). For example, a chat model could be given an image of a chest X-ray and asked to describe what it shows or diagnose a condition.[24] However, these images may not be easily available or accessible due to privacy or ethical issues, and they may require specialized knowledge or skills to interpret or analyze.[25]

These types of data can help LLMs learn from diverse and rich sources of medical information and improve their performance and accuracy in the medical domain. However, there are also some challenges and limitations that need to be addressed when using these data sources, such as:

Domain-specific knowledge: The medical domain requires a high level of expertise and understanding of complex and technical concepts and terminology. However, LLMs may not have sufficient or accurate domain knowledge to generate appropriate and relevant responses. For example, they may not know the meaning or usage of medical abbreviations, acronyms, or symbols, or they may not recognize the difference between similar or synonymous terms. Therefore, LLMs need to be trained and tested on domain-specific data sources that can provide them with adequate and correct domain knowledge. Additionally, the model performance mismatch problem is one example of how low-quality datasets might impair model performance. When the model performs well on the training dataset but badly on the test dataset, this issue develops. Overfitting, which occurs when the model learns noise or patterns in the training dataset that do not transfer well to new data, might be the cause.[26]

Clinical variability: The medical domain involves a high degree of variability and uncertainty in clinical situations and outcomes. However, Chatbot AI tools may not be able to handle or account for this variability and uncertainty in their responses. For example, they may not consider the in-

dividual differences or preferences of patients, such as their age, gender, ethnicity, medical history, or comorbidities, or they may not acknowledge the limitations or risks of their recommendations, such as side effects, contraindications, or interactions. Therefore, LLMs need to be trained and tested on diverse and realistic data sources that can capture the variability and uncertainty of the medical domain.

Interpretability and explainability: The medical domain requires a high level of transparency and accountability in the generation and communication of information and decisions. However, Chatbot AI tools may not be able to provide clear and understandable explanations or justifications for their responses. For example, they may not reveal the sources or evidence that support their answers, or they may not provide the rationale or logic behind their suggestions.[27] Therefore, LLMs need to be trained and tested on data sources that can enable them to generate interpretable and explainable responses that can increase the trust and confidence of the users and patients.

- Validation and verification: The medical domain requires a high level of accuracy and reliability in the generation and communication of information and decisions. However, LLMs may not be able to validate or verify their responses against other sources or standards. For example, they may not check the validity or currency of their references, or they may not compare their outputs with other models or methods. Therefore, LLMs need to be trained and tested on data sources that can allow them to validate and verify their responses and ensure their quality and consistency.
- Real-time decision-making: The medical domain requires a high level of speed and efficiency in the generation and communication of information and decisions. However, may not be able to generate or communicate their responses in a timely or effective manner. For example, they may take too long to process or respond to the user's input, or they may use too much or too little information or detail in their output. Therefore, Chatbot AI tools need to be trained and tested on data sources that can enable them to generate and communicate their responses in a real-time or near-real-time fashion and meet the user's expectations and needs.
- Lack of data standardization: The medical domain involves a lack of standardization or uniformity in the format or structure of the data sources. However, LLMs may not be able to process or understand different or inconsistent data formats or structures. For example, they may not recognize the difference between American and British spelling or punctuation, or they may not handle different types or units of measurement. Therefore, LLMs need to be trained and tested on data sources that follow a common or standardized format or structure that can be easily processed and understood by the models.
- Data bias and generalization: The medical domain involves a risk of bias or generalization in the data sources that may affect the outputs of the LLMs. However, LLMs may not be able to detect or correct these biases or generalizations in their responses. For example, they may reflect or reinforce existing stereotypes or prejudices in the data, such as gender, race, age, or culture, or they may overgeneralize or oversimplify their answers based on limited or narrow data. Therefore, LLMs need to be trained and tested on data sources that are fair and inclusive for all users and patients and that can enable them to generate accurate and specific responses.

THE ETHICAL IMPLICATIONS OF USING LLMS IN THE MEDICAL FIELD

Using LLMs in the medical field raises several ethical issues that need to be considered and addressed. These include:

- Accuracy and reliability: Open AI chat models do not have a clear authority or quality control mechanism to ensure their responses are based on valid and up-to-date evidence. Moreover, they may generate inaccurate or misleading information due to errors, biases, or gaps in their training data, which may include unreliable or outdated sources from the internet. For example, a study by Beam et al. (2023) found that LLMs could diagnose medical conditions at home with reasonable accuracy, but they also made some serious mistakes, such as suggesting that chest pain could be treated with aspirin or that a rash could be a sign of HIV infection.[13] Therefore, users of LLMs need to be aware of the limitations and uncertainties of these models and verify their responses with other sources or professionals before making any medical decisions.
- Privacy and security: LLMs are trained on large amounts of text data from the internet, which may contain sensitive or personal information about individuals or groups that are mentioned in their sources. For example, a study by Carlini et al. (2020) showed that Chatbot AI tools could reveal private details about people's names, addresses, phone numbers, or credit card numbers by generating texts that contained this information.[28,29] Furthermore, LLMs may also pose a risk of data breaches or misuse if they are accessed by unauthorized or malicious parties who could exploit their responses for harmful purposes. For example, hackers could use LLMs to impersonate doctors or patients and obtain confidential information or influence their behavior, Language models may produce realistic and persuasive language for a variety of purposes and contexts, but they can also produce damaging or deceptive information. As a result, it is critical to analyze and monitor the models in various settings and scenarios, as well as to put in place suitable protections and rules to avoid or decrease the likelihood of misuse.[26] Therefore, users of Chatbot AI tools need to be careful about what data they share with these models and how they protect their data from unauthorized access or use.

- Social and cultural impact: LLMs are influenced by the language and norms of their training data, which may reflect or reinforce existing stereotypes, prejudices, or inequalities in society. For example, a study by Bender et al. (2021) found that LLMs could generate texts that were sexist, racist, homophobic, or otherwise offensive or harmful to certain groups.[30, 31] Moreover, LLMs may also affect the relationship and communication between doctors and patients by changing their expectations, roles, or responsibilities.
- Authorship and trust: Concerns concerning authorship and trust are raised by the use of LLMs in medical literature and research. LLM outputs may not reflect the latest recent data and are difficult to distinguish from the voices of actual authors. This can result in information that is false or deceptive. The distinction between an LLM used as a tool for assistance and an LLM used as an author is also questionable. The International Committee of Medical Journal Editors (ICMJE) has authoring guidelines, although LLMs were not considered when these standards were being developed. It's probable that LLMs provide more benefits than other forms of assistive technology, and that new rules will be required to handle the difficulties associated with utilizing LLMs in medical writing.[32]

These ethical issues require careful consideration and regulation when using LLMs in the medical field. Users of these models need to be informed about their benefits and risks and given the option to opt-in or opt-out of their use. Developers of these models need to follow ethical principles and guidelines and ensure that their models are transparent, fair, accountable, and trustworthy. Researchers of these models need to conduct rigorous and responsible studies and report their findings and limitations honestly and openly. By addressing these ethical issues, LLMs can be used in a safe and beneficial way for medical research and scientific writing.[33]

DIFFERENT METHODS CAN BE USED TO MEASURE THE PERFORMANCE AND QUALITY OF LLMS FOR MEDICAL RESEARCH AND SCIENTIFIC WRITING

Chatbot AI tools are artificial intelligence systems that can generate natural language responses based on text or image inputs.[19,34] They can potentially assist with various tasks in medical research and scientific writing, such as literature review,[35] data analysis,[36] draft generation,[37] summarization,[38] translation,[39] and proofreading.[27] However, they also pose challenges and risks, such as bias,[40] plagiarism,[41] inaccuracies,[42] and ethical issues.[43] Therefore, it is important to evaluate their performance and quality using appropriate methods and metrics.[22]

Some possible methods that can be used to measure the performance and quality of LLMs for medical research and scientific writing are:

- Human evaluation: This involves asking human experts or users to rate the outputs of the chat models on various criteria, such as accuracy, relevance, coherence, fluency, informativeness, and usefulness.[23] Human evaluation can provide qualitative feedback and insights into the strengths and weaknesses of the chat models. However, it can also be subjective, inconsistent, time-consuming, and expensive.[24] Moreover, human evaluation may not be feasible or scalable for large-scale or long-term studies.[31]
- Automatic evaluation: This involves using computational methods or algorithms to compare the outputs of the chat models with reference texts or gold standards.[25] Automatic evaluation can provide quantitative scores and metrics that are objective, consistent, fast, and cheap.[28] However, they may not capture all aspects of natural language quality and may not correlate well with human judgments.[29] Moreover, automatic evaluation may not account for the context or purpose of the models' outputs.[33]
- Hybrid evaluation: This involves combining human and automatic methods to leverage the advantages of both approaches.[26] Hybrid evaluation can provide comprehensive and reliable assessments of the chat models by integrating human feedback and computational analysis. However, it may also require more resources and coordination than either method alone.[30] Moreover, hybrid evaluation may face challenges in aligning or reconciling the different perspectives or criteria of human and automatic methods.[34]

These methods have different strengths and limitations that need to be considered when evaluating LLMs for medical research and scientific writing. Depending on the specific goals and needs of the researchers or users, they may choose one or more methods that suit their situation. For example, they may use human evaluation for pilot studies or user satisfaction surveys; automatic evaluation for baseline comparisons or error analysis; or hybrid evaluation for comprehensive studies or quality assurance. By using appropriate methods to measure the performance and quality of LLMs, researchers, and users can ensure that these models are effective and beneficial for medical research and scientific writing.

DIFFERENT WAYS THAT CHATBOT AI TOOLS CAN BE APPLIED IN THE MEDICAL FIELD RESEARCH

LLMs can be applied in various ways in the medical field research and future research methods, such as:

Healthcare research: LLMs can help researchers conduct health studies and experiments by collecting and analyzing data from various sources, such as electronic health records, genomic data, clinical trials, online forums, and surveys. They can also help researchers generate hypotheses, design protocols, recruit participants, monitor outcomes, and disseminate findings. For example, Google Health has developed an AI model that can predict acute

kidney injuries in hospitalized patients up to 48 hours earlier than current methods.[31] This can help researchers identify patients at risk and intervene early to prevent complications or death.

Medical knowledge discovery: Chatbot AI tools can help researchers discover new insights and patterns from large and complex medical datasets. They can use natural language processing and machine learning to extract relevant information, identify relationships, infer causality, and generate explanations. For example, IBM has developed a medical knowledge discovery tool called Watson Discovery for Healthcare that can analyze scientific literature, clinical guidelines, drug labels, and other sources to provide evidence-based answers to medical questions.[28] This can help researchers find answers to challenging or novel questions and advance their knowledge and understanding of the medical domain.

Medical education and training: Chatbot AI tools can help researchers create interactive and engaging learning materials and tools for medical students and professionals. They can use natural language generation and dialogue systems to produce realistic scenarios, cases, quizzes and feedback. They can also use natural language understanding and reasoning to assess learners' performance and provide personalized guidance. For example, IBM has developed a medical education and training tool called Watson for Genomics that can teach learners how to interpret genomic data and apply it to precision medicine.[31] This can help researchers educate and train the next generation of medical experts and practitioners.

- Diagnosis: LLMs can help researchers diagnose medical conditions by analyzing the symptoms, history, and test results of patients. They can use natural language understanding and reasoning to infer the most likely diagnosis and provide evidence and explanation for their reasoning. For example, Babylon Health has developed an AI model that can diagnose common illnesses by asking questions and providing advice to users through a chat interface.[34] This can help researchers improve the accuracy and efficiency of diagnosis and reduce the burden on human doctors.
- Treatment: LLMs can help researchers recommend treatments for medical conditions by considering the diagnosis, preferences, and constraints of patients. They can use natural language generation and dialogue systems to suggest treatment options and explain their benefits and risks. For example, Ada Health has developed an AI model that can recommend treatments for various ailments by providing personalized guidance and information to users through a chat interface.[35] This can help researchers improve the quality and effectiveness of treatment and increase the satisfaction and adherence of patients.
- Patient education: LLMs can help researchers educate patients about their medical conditions and treatments by providing relevant and understandable information. They can use natural language generation and dialogue systems to answer questions and address the concerns of patients. For example, Woebot has developed an AI model that can educate patients about mental health issues by providing psychoeducation and support to users through a chat interface.[36] This can help researchers improve the awareness and knowledge of patients and empower them to manage their health better.
- Clinical decision support: LLMs can help researchers support clinical decisions by providing evidence-based recommendations and feedback. They can use natural language processing and machine learning to analyze clinical data, guidelines, literature, and other sources to provide suggestions and explanations for clinical actions. For example, IBM has developed an AI model that can support clinical decisions by providing insights and recommendations to doctors based on patient data and medical evidence.[37] This can help researchers improve the quality and safety of clinical decisions and reduce the errors and uncertainties of human doctors

RECOMMENDATIONS FOR THE FUTURE USE

LLMs are a promising technology that can assist with various tasks in medical research and scientific writing. However, they also face several challenges and risks that need to be addressed and overcome. Based on the current limitations and potential of Open AI chat in the medical field, it is recommended that the following actions be taken to ensure the safe and effective use of this technology in the future:

- Improve the accuracy and expertise of LLMs in the medical field: Further research should be conducted to train and test the models on more diverse and specific datasets of medical text, as well as to incorporate input and feedback from medical professionals. This could help the models learn from reliable and relevant sources of medical information and improve their performance and quality.
- Address and mitigate any biases present in the training data: Efforts should be made to identify and reduce any biases that may affect the models' outputs, such as gender, race, age, or culture. This could include techniques such as data preprocessing, post-processing, and bias correction algorithms. This could help the models generate fair and inclusive responses that respect the diversity and dignity of all users and patients.
- Use LLMs in conjunction with human expertise, rather than as a replacement: Medical professionals should be involved in the development, implementation, and evaluation of the systems. They should also supervise and monitor their use and intervene when necessary. This could help ensure that the systems are used appropriately and responsibly, as well as provide human touch, empathy, and accountability in medicine. Also, a guideline regarding authorship criteria and use of LLM as an assistive tool should be addressed.

- Improve the scalability and speed of LLMs to better handle a high volume of users and questions: Additional efforts should be made to optimize the systems' architecture, algorithms, and resources to enable them to handle more complex and nuanced questions in a timely manner. This could help improve the user experience and satisfaction, as well as meet the increasing demand for medical information and communication.
- Ensure the privacy and security of patient information when using LLMs in a medical setting: Strong measures should be taken to protect the data that is shared with or generated by the systems from unauthorized access or use. This could include techniques such as encryption, anonymization, or consent.[38] This could help safeguard the privacy and security of users and patients, as well as comply with ethical standards and regulations.

CONCLUSION

In this article, we have explored the current and future trends of medical chatbots, which are technologies that can enhance healthcare services and outcomes. We have analyzed 42 articles on medical chatbots and synthesized their main findings, methods, benefits, and challenges. We have shown that medical chatbots use different technologies, such as natural language processing, machine learning, knowledge bases, and rule-based systems, to perform various healthcare tasks and functions, such as health information provision, symptom checking, diagnosis, treatment, monitoring, counseling, and education. We have also shown that medical chatbots are assessed by different methods, such as user satisfaction surveys, accuracy measures, usability tests, and clinical trials. We have pointed out the strengths of medical chatbots, such as improved accessibility, convenience, efficiency, quality, and cost-effectiveness of healthcare services. We have also identified the weaknesses and challenges of medical chatbots, such as lack of standardization, regulation, validation, security, privacy, transparency, accountability, and human touch. We have proposed some future directions for research and innovation in this field, such as developing more advanced, reliable, and user-friendly medical chatbots that can meet the diverse needs and expectations of users and stakeholders. We have also stressed the need to address the ethical, legal, and social implications of medical chatbots and ensure their alignment with human values and principles.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

AVAILABILITY OF DATA AND MATERIAL

Not applicable.

FUNDING

None.

AUTHORS' CONTRIBUTIONS

MAJ, SA, IMA: Conceptualization, Literature search, Manuscript preparation, final editing.

All authors read and approved the final manuscript.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Submitted: August 28, 2023 AST, Accepted: September 17, 2023 AST

REFERENCES

1. OpenAI. GPT-4. https://openai.com/research/gpt-4

2. Forbes Technology Council. What ChatGPT and other AI tools mean for the future of healthcare. Forbes. Published 2023. https://www.forbes.com/sites/forbestechcouncil/2023/02/06/what-chatgpt-and-other-ai-tools-mean-for-the-future-of-healthcare/?sh=5202365f6b8a

3. OpenAI. Models. https://platform.openai.com/docs/models

4. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. In: *Advances in Neural Information Processing Systems*. Vol 33. ; 2020.

5. Liu Y, Wang S, Liang J, et al. MedQA: A large medical question answering dataset. *arXiv preprint arXiv:200101024*. Published online 2020.

6. Tsatsaronis G, Balikas G, Malakasiotis P, et al. BioASQ: A challenge on large-scale biomedical semantic indexing and question answering. In: *AAAI Fall Symposium: Information Retrieval and Knowledge Discovery in Biomedical Text*. ; 2015.

7. NEJM Knowledge+. https://knowledgeplus.nejm.org

8. Roller S, Dinan E, Goyal N, et al. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:200413637*. Published online 2020.

9. Zhang Y, Sun S, Galley M, et al. Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:191100536*. Published online 2019.

10. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. 2016;3(1):160035. doi:10.1038/sdata.2016.35

11. Chen Q, Li WN, Lin ZY, et al. COVID-19 infection diagnosis with chest CT image: a survey of recent advances and challenges. *Journal of Medical Imaging and Health Informatics*. 2020;10(7):1572-1582.

12. X.-L... Liang C.-Y... He Y.-N... Guan Y.-D... Xie P LZYCQRLZHLWNCQ. A survey on deep learning for chest CT image diagnosis of COVID-19 infection: recent advances and challenges. *IEEE Access*. 2020;8:206244-206261.

13. Beam AL, Lee JYK, Kohane IS, Barnett GO. AI chatbots can diagnose medical conditions at home. How good are they? *Scientific American*.

14. Carlini N, Daumé H III, Gardner M. Extracting training data from large language models. *arXiv preprint arXiv:201207805*.

15. ChatDoctor: A medical chat model fine-tuned on LLaMA model using Wikipedia and Database Brain. *arXiv preprint arXiv:230314070*. Published online 2023.

16. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: Can language models be too big? In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM; 2021:610-623. doi:10.1145/3442188.3445922

17. Artificial intelligence in medicine. IBM. https://www.ibm.com/topics/artificial-intelligence-medicine

18. Tolchin B. The ethics of using Chatbot AI tools in medicine. *Journal of Medical Ethics*. 2023;49(2):123-134. doi:10.1007/s11948-022-00369-2

19. Kung TH, Cheatham M, Medenilla A, Kung JW. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2020;2(2):e0000198. doi:10.1371/journal.pdig.0000198

20. Liu Y, Lapata M. Text summarization with pretrained encoders. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics; 2019:3730-3740. doi:10.18653/v1/d19-1387

21. Novikova J, Dušek O, Rieser V. RankME: Reliable human ratings for natural language generation. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Vol 2. Association for Computational Linguistics; 2018:72-78. doi:10.18653/v1/n18-2012

22. Belz A, Reiter E. Comparing automatic and human evaluation of NLG systems. In: *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*. ; 2006:313-320.

23. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. *arXiv preprint arXiv:200514165*. Published online 2020.

24. Pang B, Lee L. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*. 2008;2(1–2):1-135.

25. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*. Association for Computational Linguistics; 2001:311-318. doi:10.3115/1073083.1073135

26. Liu CW, Lowe R, Serban I, Noseworthy M, Charlin L, Pineau J. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics; 2016:2122-2132. doi:10.18653/v1/d16-1230

27. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. 2016;3(1):160035. doi:10.1038/sdata.2016.35

28. Lin CY. ROUGE: A package for automatic evaluation of summaries. In: *Text Summarization Branches out: Proceedings of the ACL-04 Workshop*. ; 2004:74-81.

29. Reiter E, Belz A. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*. 2009;35(4):529-558. doi:10.1162/coli.2009.35.4.35405

30. IBM. Artificial intelligence in medicine. https://www.ibm.com/topics/artificial-intelligence-medicine

31. Google Health. Healthcare research & technology advancements. https://health.google/health-research/

32. Li H, Moon JT, Purkayastha S, Celi LA, Trivedi H, Gichoya JW. Ethics of large language models in medicine and medical research. *Lancet Digit Health*. 2023;5(6):e333-e335. doi:10.1016/s2589-7500(23)00083-3

33. Digital Health. Google Research and DeepMind develop AI medical chatbot. Published January 12, 2023. https://www.digitalhealth.net/2023/01/google-research-and-deepmind-develop-ai-medical-chatbot/

34. Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. In: *Advances in Neural Information Processing Systems*. Vol 27. ; 2014:3104-3112.

35. Nallapati R, Zhai F, Zhou B. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol 31. Association for the Advancement of Artificial Intelligence (AAAI); 2017:3075-3081. doi:10.1609/aaai.v31i1.10958

36. Fan A, Lewis M, Dauphin YN. Hierarchical neural story generation. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ; 2018:889-898.

37. Ghazvininejad M, Brockett C, Chang MW, et al. A knowledge-grounded neural conversation model. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol 32. Association for the Advancement of Artificial Intelligence (AAAI); 2017:5110-5117. doi:10.1609/aaai.v32i1.11977

38. Senseforth.AI. Medical chatbots - Use cases, examples and case studies of conversational AI in healthcare. Accessed January 15, 2023. https://www.senseforth.ai/conversational-ai/medical-chatbots/

39. Mazaré PE, Humeau S, Raison M, Bordes A. Training millions of personalized dialogue agents. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics; 2018:2775-2779. doi:10.18653/v1/d18-1298

40. Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L. The ethics of algorithms: Mapping the debate. *Big Data & Society*. 2016;3(2):205395171667967. doi:10.1177/2053951716679679

41. The Medical Futurist. The top 12 healthcare chatbots. Accessed January 15, 2023. https://medicalfuturist.com/top-12-health-chatbots/

42. Reardon S. AI chatbots can diagnose medical conditions at home.How good are they? Scientific American. Published March 31, 2023. Accessed January 15, 2023. https://www.scientificamerican.com/article/ai-chatbots-can-diagnose-medical-conditions-at-home-how-good-are-they/

43. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE J Biomed Health Inform*. 2018;22(5):1589-1604. doi:10.1109/jbhi.2017.2767063