Editorial

Redefining Virtual Assistants in Health Care: The Future With Large Language Models

Emre Sezgin[1,2], PhD

[1]The Abigail Wexner Reseach Institute at Nationwide Children's Hospital, Columbus, OH, United States

[2]The Ohio State University College of Medicine, Columbus, OH, United States

Corresponding Author:
Emre Sezgin, PhD
The Abigail Wexner Reseach Institute at Nationwide Children's Hospital
700 Children's Drive
Columbus, OH, 43205
United States
Phone: 1 6147223179
Email: emre.sezgin@nationwidechildrens.org

Abstract

This editorial explores the evolving and transformative role of large language models (LLMs) in enhancing the capabilities of virtual assistants (VAs) in the health care domain, highlighting recent research on the performance of VAs and LLMs in health care information sharing. Focusing on recent research, this editorial unveils the marked improvement in the accuracy and clinical relevance of responses from LLMs, such as GPT-4, compared to current VAs, especially in addressing complex health care inquiries, like those related to postpartum depression. The improved accuracy and clinical relevance with LLMs mark a paradigm shift in digital health tools and VAs. Furthermore, such LLM applications have the potential to dynamically adapt and be integrated into existing VA platforms, offering cost-effective, scalable, and inclusive solutions. These suggest a significant increase in the applicable range of VA applications, as well as the increased value, risk, and impact in health care, moving toward more personalized digital health ecosystems. However, alongside these advancements, it is necessary to develop and adhere to ethical guidelines, regulatory frameworks, governance principles, and privacy and safety measures. We need a robust interdisciplinary collaboration to navigate the complexities of safely and effectively integrating LLMs into health care applications, ensuring that these emerging technologies align with the diverse needs and ethical considerations of the health care domain.

(J Med Internet Res 2024;26:e53225) doi: 10.2196/53225

KEYWORDS

large language models; voice assistants; virtual assistants; chatbots; conversational agents; health care

Virtual assistants (VAs)—mostly voice assistants, chatbots, and dialogue-based interactive applications—have been leading conversational technologies, being used for health care communications and remote monitoring [1-4]. However, their accuracy and reliability in understanding and responding to medical questions have been limitations [5], which have been slowly improving over the years [6]. Large language models (LLMs) offer scalable and customizable solutions for these limitations of VAs. The body of literature demonstrating the capabilities of LLMs in medicine and health care has been growing, and a number of studies benchmarking LLMs' performance against each other or humans' medical knowledge, decision-making processes, and empathic responses have been published [7-11]. Furthermore, LLM-based services improve equitable access to information and reduce language barriers via contextual and culturally aware systems [12] and privacy-preserving, local solutions for low-resource settings [13-16]. This research and evidence give a glimpse at the future of personalized VAs.

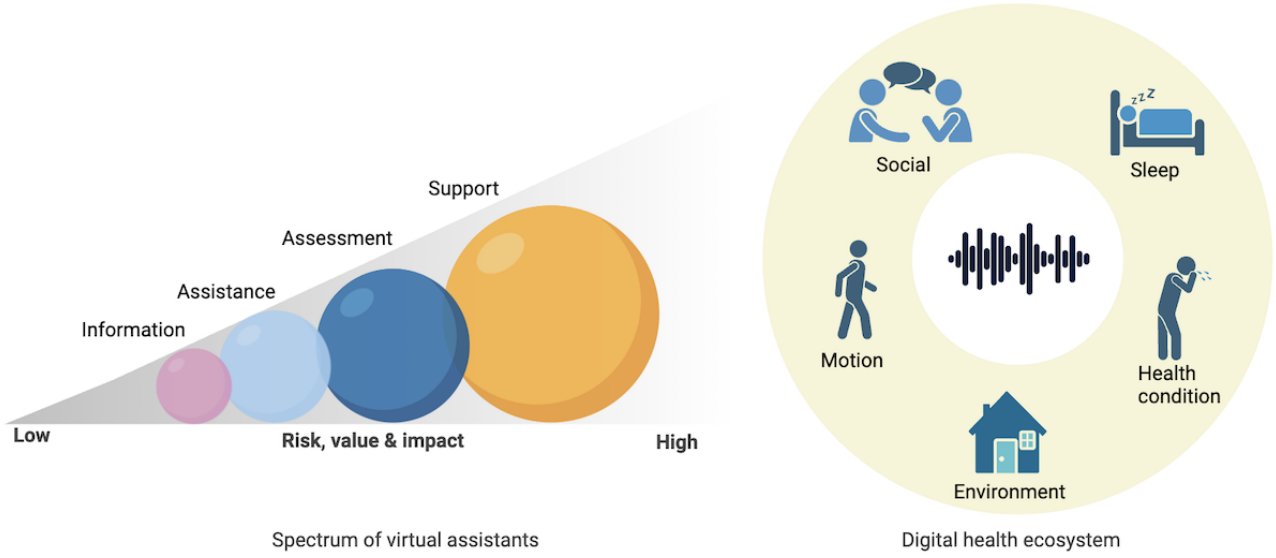
In the context of mental health and health information–seeking behavior, we investigated the performance of VAs in responding to postpartum depression–related frequently asked questions. The evidence from our two studies (conducted in 2021 with VAs [17] and 2023 with LLMs [18]) provides comparable findings on the 2-year difference in technology; illuminates the evolving roles of artificial intelligence, natural language processing, and LLMs in health care; and shows the promise of a more accurate and reliable digital health landscape.

In our first study in 2021 [17], we investigated the clinical accuracy of Google Assistant, Amazon Alexa, Microsoft Cortana, and Apple Siri voice assistant responses to postpartum depression–related questions. In our second study in 2023 [18], we replicated our research by using LLMs, and the new study

showed significant improvements in the accuracy and clinical relevance of responses. Specifically, GPT-4's responses to all postpartum depression–related questions were more accurate and clinically relevant in contrast to the VAs, of which the proportion of clinically relevant responses did not exceed the 29% reported in our earlier study. In addition, the interrater reliability score for LLMs (GPT-4: $\kappa=1$; $P<.05$) was higher than that for VAs ($\kappa=0.87$; $P<.001$), underscoring LLMs' consistency in clinically relevant responses, which is vital to achieve for health care applications. LLMs also recommended consultations with health care providers—a feature that adds an extra layer for safety—which was not observed in our earlier study with VAs. This dramatic improvement suggests a paradigm shift in the capabilities of digital health tools for health information–seeking activities. The high clinical accuracy and reliability of LLMs point toward a promising future for their integration into existing VA platforms. LLMs can offer dynamic adaptability for VAs via custom applications and decentralized LLM architectures [19,20]. Given their capabilities for open-source developments and collaborations, LLMs could serve as cost-effective and inclusive frameworks for collaborative developments (ie, among technology providers, patients, and clinical experts) in fine-tuning and training VAs for specific medical purposes.

The empirical data from our studies, as well as the literature [21], indicate a compelling trajectory toward LLMs being used to potentially improve the clinical and instructional capabilities of conversational technologies. This suggests a shift in our earlier spectrum model for VAs in health care (Figure 1) [22], in which we proposed 4 service levels for a spectrum of VA use that were associated with the risk, value, and impact of VAs. These levels were the "information" (eg, asking Amazon Alexa to start self-care guidance), "assistance" (eg, setting up reminders for medication or self-therapy), "assessment" (eg, identification, detection, prediction with digital biomarkers, and management), and "support" (prescribing, substituting, or supplementing medication and therapy tools) levels. In 2020, the evidence on the utilization of VAs in health care indicated that VAs were at the "information" and "assistance" levels [22]. However, LLMs are opening up opportunities for VAs, potentially toward the "assessment" and "support" levels. As Figure 1 shows, the level of a service and the associated risk, value, and impact of the service can change based on the targeted problems and solutions. A digital health ecosystem represents the ecosystem where we may envision a future of support from VAs enhanced by LLMs with speech interaction and audio-based sensing capabilities [23]. Such enhancements may include quantifying human behavior–related factors beyond VA engagement, such as social engagement, emotions, neurodevelopmental and behavioral health, sleep health (snoring, heart rate, and movements), respiratory symptoms (sneezing and coughing), and motion (gait, exercise, and sedentary behavior).

Figure 1. Spectrum of virtual assistants (outlines the risk, value, and impact in health care services) and applications in digital health ecosystems. These can change based on the targeted problems and solutions. This figure was created with BioRender.com (BioRender).



Despite this promising horizon, we need to approach cautiously. As LLMs are becoming highly appealing tools that can be used as VAs in health care, it is imperative to establish a platform that facilitates democratized access and interdisciplinary collaboration during the development of such applications [19,24]. This platform should be designed to bring together a diverse range of stakeholders, including technologists, ethicists, researchers, health care professionals, and patients. This would ensure that the development and integration of VAs are guided by a balanced perspective that considers ethical guidelines and regulatory oversight [25,26], governance principles [27,28], privacy and safety measures [29], feasibility, efficacy, and patient-centric approaches and assessment methods [30,31]. By prioritizing such collaborative and inclusive dialogues, we can better navigate the complex challenges and harness the full potential of these advanced technologies in health care, ensuring that they are developed responsibly, ethically, and in alignment with the diverse needs of all users.

Acknowledgments

The author thanks Yasemin Sezgin for her constructive feedback. Figure 1 was created with BioRender.com (BioRender).

Conflicts of Interest

ES serves on the editorial board of JMIR Publications.

References

1. Sezgin E, Huang Y, Ramtekkar U, Lin S. Readiness for voice assistants to support healthcare delivery during a health crisis and pandemic. NPJ Digit Med. Sep 16, 2020;3:122. [FREE Full text] [doi: 10.1038/s41746-020-00332-0] [Medline: 33015374]
2. Corbett CF, Combs EM, Chandarana PS, Stringfellow I, Worthy K, Nguyen T, et al. Medication adherence reminder system for virtual home assistants: mixed methods evaluation study. JMIR Form Res. Jul 13, 2021;5(7):e27327. [FREE Full text] [doi: 10.2196/27327] [Medline: 34255669]
3. Xu L, Sanders L, Li K, Chow JCL. Chatbot for health care and oncology applications using artificial intelligence and machine learning: systematic review. JMIR Cancer. Nov 29, 2021;7(4):e27850. [FREE Full text] [doi: 10.2196/27850] [Medline: 34847056]
4. Sawad AB, Narayan B, Alnefaie A, Maqbool A, Mckie I, Smith J, et al. A systematic review on healthcare artificial intelligent conversational agents for chronic conditions. Sensors (Basel). Mar 29, 2022;22(7):2625. [FREE Full text] [doi: 10.3390/s22072625] [Medline: 35408238]
5. Palanica A, Thommandram A, Lee A, Li M, Fossat Y. Do you understand the words that are comin outta my mouth? Voice assistant comprehension of medication names. NPJ Digit Med. Jun 20, 2019;2:55. [FREE Full text] [doi: 10.1038/s41746-019-0133-x] [Medline: 31304401]
6. Palanica A, Fossat Y. Medication name comprehension of intelligent virtual assistants: a comparison of Amazon Alexa, Google Assistant, and Apple Siri between 2019 and 2021. Front Digit Health. May 19, 2021;3:669971. [FREE Full text] [doi: 10.3389/fdgth.2021.669971] [Medline: 34713143]
7. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. Aug 2023;620(7972):172-180. [FREE Full text] [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
8. Thirunavukarasu AJ, Hassan R, Mahmood S, Sanghera R, Barzangi K, El Mukashfi M, et al. Trialling a large language model (ChatGPT) in general practice with the Applied Knowledge Test: observational study demonstrating opportunities and limitations in primary care. JMIR Med Educ. Apr 21, 2023;9:e46599. [FREE Full text] [doi: 10.2196/46599] [Medline: 37083633]
9. 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. Aug 22, 2023;25:e48659. [FREE Full text] [doi: 10.2196/48659] [Medline: 37606976]
10. Sharma A, Lin IW, Miner AS, Atkins DC, Althoff T. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. Nat Mach Intell. Jan 23, 2023;5(1):46-57. [doi: 10.1038/s42256-022-00593-2]
11. 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. Jun 1, 2023;183(6):589-596. [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
12. Yao B, Jiang M, Yang D, Hu J. Empowering LLM-based machine translation with cultural awareness. arXiv. Preprint posted online on May 23, 2023. [FREE Full text]
13. Wiest IC, Ferber D, Zhu J, van Treeck M, Meyer SK, Juglan R, et al. From text to tables: a local privacy preserving large language model for structured information retrieval from medical documents. medRxiv. Preprint posted online on Dec 8, 2023. [FREE Full text] [doi: 10.1101/2023.12.07.23299648]
14. Cai W. Feasibility and prospect of privacy-preserving large language models in radiology. Radiology. Oct 2023;309(1):e232335. [doi: 10.1148/radiol.232335] [Medline: 37815443]
15. Yang R, Tan TF, Lu W, Thirunavukarasu AJ, Ting DSW, Liu N. Large language models in health care: development, applications, and challenges. Health Care Science. Jul 24, 2023;2(4):255-263. [FREE Full text] [doi: 10.1002/hcs2.61]
16. Ogueji K, Zhu Y, Lin J. Small data? No problem! Exploring the viability of pretrained multilingual language models for low-resourced languages. In: Ataman D, Birch A, Conneau A, Firat O, Ruder S, Sahin GG, editors. Proceedings of the 1st Workshop on Multilingual Representation Learning. Kerrville, TX. Association for Computational Linguistics; 2021;116-126.
17. Yang S, Lee J, Sezgin E, Bridge J, Lin S. Clinical advice by voice assistants on postpartum depression: cross-sectional investigation using Apple Siri, Amazon Alexa, Google Assistant, and Microsoft Cortana. JMIR Mhealth Uhealth. Jan 11, 2021;9(1):e24045. [FREE Full text] [doi: 10.2196/24045] [Medline: 33427680]
18. Sezgin E, Chekeni F, Lee J, Keim S. Clinical accuracy of large language models and Google search responses to postpartum depression questions: cross-sectional study. J Med Internet Res. Sep 11, 2023;25:e49240. [FREE Full text] [doi: 10.2196/49240] [Medline: 37695668]

19. Xu B, Liu X, Shen H, Han Z, Li Y, Yue M, et al. Gentopia: a collaborative platform for tool-augmented LLMs. arXiv. Preprint posted online on Aug 8, 2023. [FREE Full text] [doi: 10.18653/v1/2023.emnlp-demo.20]
20. Chen C, Feng X, Zhou J, Yin J, Zheng X. Federated large language model: a position paper. arXiv. Preprint posted online on Jul 18, 2023. [FREE Full text]
21. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. Aug 2023;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
22. Sezgin E, Militello LK, Huang Y, Lin S. A scoping review of patient-facing, behavioral health interventions with voice assistant technology targeting self-management and healthy lifestyle behaviors. Transl Behav Med. Aug 7, 2020;10(3):606-628. [doi: 10.1093/tbm/ibz141] [Medline: 32766865]
23. Agbavor F, Liang H. Predicting dementia from spontaneous speech using large language models. PLOS Digit Health. Dec 22, 2022;1(12):e0000168. [FREE Full text] [doi: 10.1371/journal.pdig.0000168] [Medline: 36812634]
24. AMA issues new principles for AI development, deployment and use. American Medical Association. Nov 28, 2023. URL: https://www.ama-assn.org/press-center/press-releases/ama-issues-new-principles-ai-development-deployment-use [accessed 2023-12-25]
25. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med. Jul 6, 2023;6(1):120. [FREE Full text] [doi: 10.1038/s41746-023-00873-0] [Medline: 37414860]
26. Hacker P, Engel A, Mauer M. Regulating ChatGPT and other large generative AI models. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. New York, NY. Association for Computing Machinery; Jun 12, 2023;1112-1123.
27. Mökander J, Schuett J, Kirk HR, Floridi L. Auditing large language models: a three-layered approach. AI Ethics. May 30, 2023 [FREE Full text] [doi: 10.1007/s43681-023-00289-2]
28. Liao F, Adelaine S, Afshar M, Patterson BW. Governance of clinical AI applications to facilitate safe and equitable deployment in a large health system: key elements and early successes. Front Digit Health. Aug 24, 2022;4:931439. [FREE Full text] [doi: 10.3389/fdgth.2022.931439] [Medline: 36093386]
29. De Angelis L, Baglivo F, Arzilli G, Privitera GP, Ferragina P, Tozzi AE, et al. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. Front Public Health. Apr 25, 2023;11:1166120. [FREE Full text] [doi: 10.3389/fpubh.2023.1166120] [Medline: 37181697]
30. Ding H, Simmich J, Vaezipour A, Andrews N, Russell T. Evaluation framework for conversational agents with artificial intelligence in health interventions: a systematic scoping review. J Am Med Inform Assoc. Dec 09, 2023 Epub ahead of print. [doi: 10.1093/jamia/ocad222] [Medline: 38070173]
31. Chang Y, Wang X, Wang J, Wu Y, Yang L, Zhu K, et al. A survey on evaluation of large language models. arXiv. Preprint posted online on Dec 29, 2023. [FREE Full text]

Abbreviations

LLM: large language model
VA: virtual assistant

Edited by T Leung, T de Azevedo Cardoso; this is a non–peer-reviewed article. Submitted 29.09.23; accepted 02.01.24; published 19.01.24.

Please cite as:
Sezgin E
Redefining Virtual Assistants in Health Care: The Future With Large Language Models
J Med Internet Res 2024;26:e53225
URL: https://www.jmir.org/2024/1/e53225
doi: 10.2196/53225
PMID: 38241074