


COMMENTARY OPEN ACCESS

Agentic Large Language Models for Healthcare: Current Progress and Future Opportunities

Han Yuan 

Duke-NUS Medical School, National University of Singapore, Singapore

Correspondence: Han Yuan (yuan.han@u.duke.nus.edu)**Received:** 31 October 2024 | **Revised:** 16 November 2024 | **Accepted:** 12 December 2024**Funding:** The author received no specific funding for this work.**Keywords:** artificial intelligence agent | generative pre-trained transformer | human-computer interaction | large language models | natural language processing

1 | Agentic Large Language Models

Large language models (LLMs) have revolutionized the healthcare industry by advancing artificial intelligence (AI) to expert-level capabilities in understanding, reasoning, and generating human language [1, 2]. However, standard LLMs occasionally make basic errors, such as miscalculations in simple arithmetic problems [3], prompting the development of agentic LLMs (ALLMs) to address these limitations. ALLMs combine vanilla LLMs' robust reasoning, planning, and reflecting capabilities acquired from vast pre-trained datasets, supervised instructions, and reinforcement learning from human feedback (RLHF), with specialized tools such as calculators to enhance task-specific problem-solving [4]. The advanced reasoning, planning, and reflecting abilities of ALLMs enable them to independently execute complex, multistep clinical tasks, and surpass the performance of conventional LLMs, which rely passively on clinician-provided prompts to initiate and sustain the generation process. Specialized analytical tools integrated into ALLMs streamline intermediate analytical processes and complement information from other modalities, such as medical images and clinical signals [5], which presents challenges that traditional LLMs, limited by their pre-training on internet-derived text corpora, struggle to process [6]. Additionally, ALLMs incorporate short-term memory within individual conversation sessions, thereby retaining interaction information for iterative reasoning and planning, and long-term memory across multiple sessions, thereby enabling decision-

making recall in similar historic tasks. Augmenting ALLMs with both short-term and long-term memory significantly reduces the likelihood of oversights or errors [7], which is a capability that standard LLMs can only achieve through intricate prompt tuning and complex engineering interventions. Moreover, ALLMs integrate external knowledge to mitigate challenges such as hallucinations in vanilla LLMs, thereby enabling the generation and verification of evidence-based clinical recommendations by drawing on diverse medical knowledge sources [8]. Standard LLMs can acquire clinical knowledge through fine-tuning; however, the high computational costs associated with retraining hinder their ability to keep pace with rapidly evolving medical knowledge [9]. By contrast, ALLMs can access research findings, clinical case reports, and updated guidelines in a cost-effective manner without additional training.

These augmentations enable ALLMs to tackle complex tasks that necessitate iterative reasoning, planning, and action, which align them, compared with standard LLMs, more closely with the golden clinical procedures of gathering cues, generating and interpreting hypotheses, refining them, and repeating the process as needed [10]. Figure 1 depicts the general architecture of ALLMs, which integrate LLMs as their central intelligence unit, leverage tools to enhance problem-solving capabilities, use memory to support advanced planning and reflection, and apply knowledge systems to gather contextual information.

Abbreviations: AI, artificial intelligence; ALLM, agentic large language model; LLM, large language model; RLHF, reinforcement learning from human feedback.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2025 The Author(s). *Medicine Advances* published by John Wiley & Sons Ltd on behalf of Tsinghua University Press.

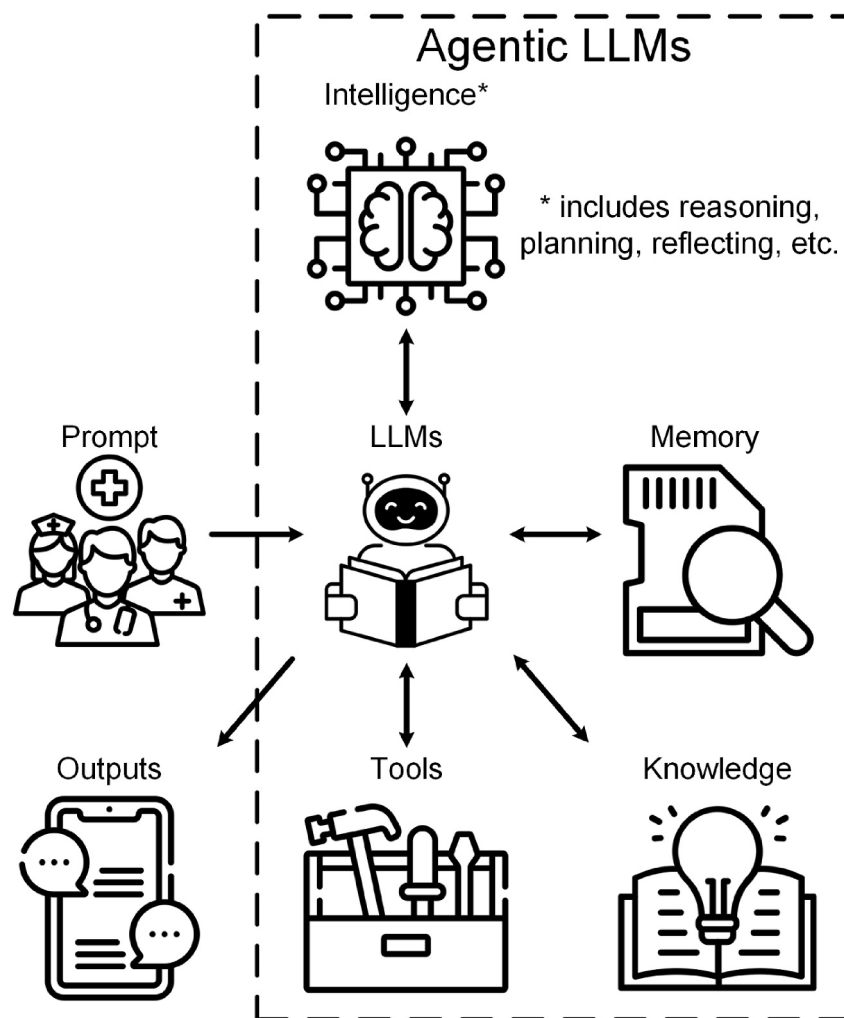


FIGURE 1 | General architecture of agentic LLMs. LLM, large language model.

2 | Current Progress

To elucidate the current landscape of ALLMs-assisted healthcare, we performed a comprehensive literature search in PubMed, as shown in Figure 2, and summarized the key takeaways in the following paragraphs. A pragmatic ALLMs-based system was proposed by Zhou et al. [11] for multi-omics analyses. It requires users to input the data path, data description, and task objectives, and then autonomously calls built-in LLMs to generate analytical plans and corresponding code for downstream execution. To mitigate errors or refine results, it retains a memory of previous inferences, actions, and outcomes. When user demands are unmet or errors are detected, another round of agentic analysis is initiated, with relevant logs reprocessed by LLM-based intelligence engines. Apart from memory logs accumulated through the past behaviors of ALLMs, external knowledge is leveraged from behavior science to improve the empathy and actionability of fitness coaching-oriented chatbots. Similarly, KNOWNET [12] augments LLMs by integrating knowledge graphs to improve their accuracy and facilitate the structured exploration of non-pharmaceutical interventions for Alzheimer's disease. Clinical web pages are queried and autonomously converted into document objects, which enables the retrieval of real-time updated online knowledge [13].

Although the single-agent LLMs discussed above have demonstrated remarkable capabilities, their limitations in handling highly complex or diverse healthcare tasks have prompted interest in more advanced approaches [14]. Consequently, researchers are shifting to multi-agent frameworks, wherein specialized agents collaborate to tackle distinct aspects of a problem, harnessing the collective intelligence of agents with unique capabilities [8]. A foundational two-agent system was exemplified by Alghamdi and Mostafa [15], consisting of one agent dedicated to generating medical guidance and a second agent responsible for validating the trustworthiness of generated responses. Similarly, a symptom-disease chatbot using a dual-agent system was developed by Ananta et al. [16]. The primary agent addresses user queries using a well-established knowledge graph. When the primary agent fails to retrieve relevant results, a secondary agent, fine-tuned on GPT-3, is activated to respond to user queries. A more complex system is established with five sequential agents responsible for medical guideline classification, question retrieval, matching evaluation, intelligent question answering, and results evaluation and source citation. In contrast to specific job assignments, agents in the study by Ghafarollahi and Buehler [17] are designated the general roles of user proxy, planner, assistant, critic, and group chat manager to collaboratively address sophisticated protein analysis and design.

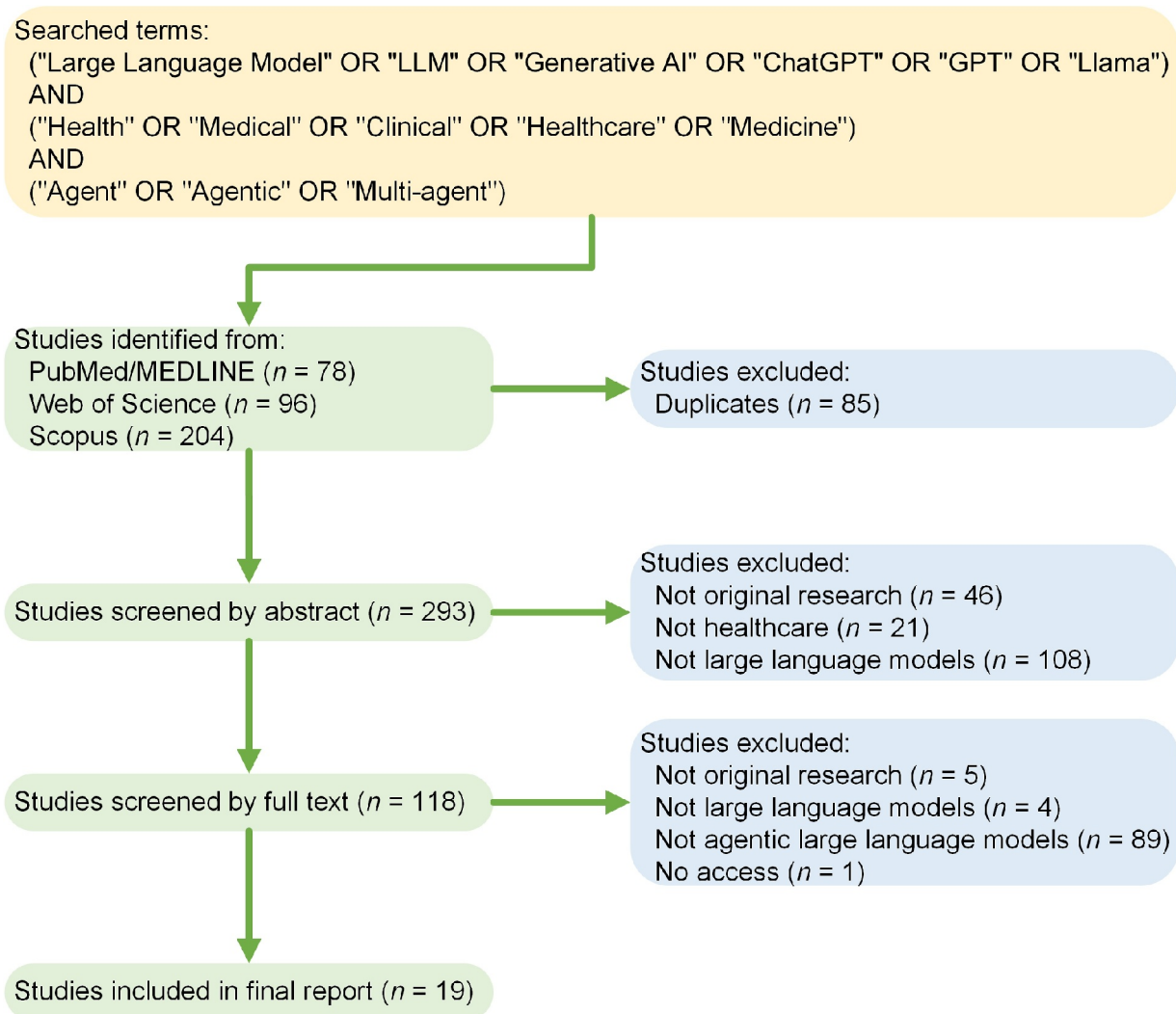


FIGURE 2 | Literature search pipeline in PubMed to identify relevant articles published from January 1, 2022 to October 19, 2024. We followed previous systematic reviews (1, 2) to design the groups of keywords: (1) “Large Language Model” “LLM” “Generative AI” “ChatGPT” “GPT” “Llama”; (2) “Health” “Medical” “Clinical” “Healthcare” “Medicine”; and (3) “Agent” “Agentic” “Multi-agent.”

In addition to modules centered on upgrading model performance, recent studies suggest that visual or audio interactions in ALLMs can enhance user engagement. In Refs. [18, 19], the authors demonstrated the critical role of avatar-like visual embodiment in achieving therapeutic success with ALLM-based psychotherapy. For audio interactions, in Refs. [20, 21], the authors developed chatbots based on GPT-3.5 tailored to elderly individuals, and embedded speech recognition and text-to-speech tools to enable human-like conversations. My Care Questionnaire [22] and Convai [23] further explore both visual and audio interactivity to facilitate health data entry for individuals with sensory impairments and patient encounter simulations. In advanced studies in which avatars were provided with diverse exteriors [24], multiple avatars were designed with distinct personas tailored to accommodate users' preferences in a digital assistant aimed at promoting physical activity [25].

Lastly, although ALLMs show promise, their reliability and safety require thorough validation [26], particularly for specialized medical applications, because of misdiagnosis, poor management, and lack of clinical knowledge [1, 27]. Despite

their integration with clinical knowledge-based guidelines, ALLMs still underperform human experts in detecting antimicrobial resistance mechanisms [28].

3 | Future Opportunities

As illustrated in Figure 2, only 19 studies on ALLMs were identified from the 118 abstract-screened articles on LLM-assisted healthcare, which demonstrates limited attention accorded to the advanced technique within the broad community. This forms the core of our commentary to provide clinical experts with a foundational understanding of the latest research trends and foster their collaboration with AI researchers, and identify actionable scenarios of ALLMs that suit clinical needs [29] rather than imaginary applications driven by cutting-edge techniques or over-engineered frameworks. Additionally, deploying ALLMs-based tools in healthcare requires aligning model behavior with both user preferences and strict adherence to clinical guidelines, which calls for further investigation into

whether general-purpose instruction tuning, RLHF, and LLMs' self-evolution inspired by DeepSeek can ensure compliance in real-world deployment. Moreover, based on clinician-conceptualized applications and AI researcher-developed methodology, dataset preparation requires the collaboration of AI researchers and clinicians to develop efficient annotation interfaces and organize panel discussions to design annotator reference materials, disagreement resolution approaches, and quality assessment protocols [30].

Author Contributions

Han Yuan: conceptualization, data curation, formal analysis, investigation, methodology, visualization, writing—original draft, writing—review and editing.

Acknowledgments

The author has nothing to report.

Ethics Statement

This study is exempted from review by the ethics committee as it does not involve human participants, animal subjects, or the collection of sensitive data.

Consent

The author has nothing to report.

Conflicts of Interest

The author declares no conflicts of interest.

Data Availability Statement

The author has nothing to report.

References

1. J. Haltaufderheide and R. Ranisch, "The Ethics of ChatGPT in Medicine and Healthcare: A Systematic Review on Large Language Models (LLMs)," *NPJ Digital Medicine* 7, no. 1 (2024): 183, <https://doi.org/10.1038/s41746-024-01157-x>.
2. H. Yuan, "Natural Language Processing for Chest X-Ray Reports in the Transformer Era: BERT-Like Encoders for Comprehension and GPT-Like Decoders for Generation," *iRADIOLOGY* (2025): 1–8, <https://doi.org/10.1002/ird3.115>.
3. K. M. Collins, A. Q. Jiang, S. Frieder, et al., "Evaluating Language Models for Mathematics Through Interactions," *Proceedings of the National Academy of Sciences of the United States of America* 121, no. 24 (2024): e2318124121, <https://doi.org/10.1073/pnas.2318124121>.
4. Z. Khan, V. Kumar, S. Schuler, and M. Chandraker, "Foundational Vision-LLM for AI Linkage and Orchestration," *NEC Technical Journal* 17, no. 2 (2024): 96–101.
5. H. Yuan, K. Yu, F. Xie, M. Liu, and S. Sun, "Automated Machine Learning With Interpretation: A Systematic Review of Methodologies and Applications in Healthcare," *Medicine Advances* 2, no. 3 (2024): 205–237, <https://doi.org/10.1002/meda.75>.
6. Y. Kang and J. Kim, "ChatMOF: An Artificial Intelligence System for Predicting and Generating Metal-Organic Frameworks Using Large Language Models," *Nature Communications* 15, no. 1 (2024): 4705, <https://doi.org/10.1038/s41467-024-48998-4>.
7. C. Gao, X. Lan, N. Li, et al., "Large Language Models Empowered Agent-Based Modeling and Simulation: A Survey and Perspectives," *Humanities and Social Sciences Communications* 11, no. 1 (2024): 1259, <https://doi.org/10.1057/s41599-024-03611-3>.
8. X. Li, S. Wang, S. Zeng, Y. Wu, and Y. Yang, "A Survey on LLM-Based Multi-Agent Systems: Workflow, Infrastructure, and Challenges," *Vicinagearth* 1, no. 1 (2024): 9, <https://doi.org/10.1007/s44336-024-00009-2>.
9. D. Truhn, G. Müller-Franzes, and J. N. Kather, "The Ecological Footprint of Medical AI," *European Radiology* 34, no. 2 (2024): 1176–1178, <https://doi.org/10.1007/s00330-023-10123-2>.
10. M. A. Jones, "Clinical Reasoning in Manual Therapy," *Physical Therapy* 72, no. 12 (1991): 875–884, <https://doi.org/10.1093/ptj/72.12.875>.
11. J. Zhou, B. Zhang, G. Li, et al., "An AI Agent for Fully Automated Multi-Omic Analyses," *Advanced Science* 11, no. 44 (2024): 2407094, <https://doi.org/10.1002/adv.202407094>.
12. Y. Yan, Y. Hou, Y. Xiao, R. Zhang, and Q. Wang, "KNowNET: Guided Health Information Seeking From LLMs via Knowledge Graph Integration," *IEEE Transactions on Visualization and Computer Graphics* 31, no. 1 (2024): 547–557, <https://doi.org/10.1109/TVCG.2024.3456364>.
13. J. Yang, L. Shu, H. Duan, and H. Li, "RDguru: A Conversational Intelligent Agent for Rare Diseases," *IEEE Journal of Biomedical and Health Informatics* (2024): 1–13, <https://doi.org/10.1109/JBHI.2024.3464555>.
14. A. Soltoggio, E. Ben-Iwhiwhu, V. Braverman, et al., "A Collective AI via Lifelong Learning and Sharing at the Edge," *Nature Machine Intelligence* 6, no. 3 (2024): 251–264, <https://doi.org/10.1038/s42256-024-00800-2>.
15. H. M. Alghamdi and A. Mostafa, "Towards Reliable Healthcare LLM Agents: A Case Study for Pilgrims During Hajj," *Information* 15, no. 7 (2024): 371, <https://doi.org/10.3390/info15070371>.
16. I. Ananta, S. Khetarpaul, D. Sharma, I. Ananta, S. Khetarpaul, and D. Sharma, "Symptoms-Disease Detecting Conversation Agent Using Knowledge Graphs," in *Proceedings of the 2024 Australasian Computer Science Week* (Sydney, NSW, Australia, 2024), 98–107, <https://doi.org/10.1145/3641142.3641165>.
17. A. Ghafarollahi and M. J. Buehler, "ProtAgents: Protein Discovery via Large Language Model Multi-Agent Collaborations Combining Physics and Machine Learning," *Digital Discovery* 3, no. 7 (2024): 1389–1409, <https://doi.org/10.1039/D4DD00013G>.
18. W. Vossen, M. Szymanski, K. Verbert, W. Vossen, M. Szymanski, and K. Verbert, "The Effect of Personalizing a Psychotherapy Conversational Agent on Therapeutic Bond and Usage Intentions," in *Proceedings of the 29th International Conference on Intelligent User Interfaces* (Greenville, SC, 2024), 761–771, <https://doi.org/10.1145/3640543.3645195>.
19. K. Hopman, D. Richards, and M. N. Norberg, "An Embodied Conversational Agent to Support Wellbeing After Injury: Insights From a Stakeholder Inclusive Design Approach," in *In Persuasive Technology: Proceedings of the 2024 Conference*, Edited by Editor(s) (Cham: Springer Nature Switzerland, 2024), 161–175, https://doi.org/10.1007/978-3-031-58226-4_13.
20. K. Shimizu, B. Ami, C. Dongeon, et al., "Exploring Photo-Based Dialogue Between Elderly Individuals and Generative AI Agents," *International Journal of Advanced Computer Science and Applications* 15, no. 7 (2024): 1–8, <https://doi.org/10.14569/ijacsa.2024.01507106>.
21. A. Alessa and H. Al-Khalifa, "Towards Designing a ChatGPT Conversational Companion for Elderly People," in *Proceedings of the 16th International Conference on Pervasive Technologies Related to Assistive Environments* (2023), <https://doi.org/10.1145/3594806.3596572>.

22. A. Cuadra, J. Breuch, S. Estrada, et al., "Digital Forms for All: A Holistic Multimodal Large Language Model Agent for Health Data Entry," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, no. 2 (2024): 1–39, <https://doi.org/10.1145/3659624>.
23. N. Sardesai, P. Russo, J. Martin, and A. Sardesai, "Utilizing Generative Conversational Artificial Intelligence to Create Simulated Patient Encounters: A Pilot Study for Anaesthesia Training," *Postgraduate Medical Journal* 100, no. 1182 (2024): 237–241, <https://doi.org/10.1093/postmj/qgad137>.
24. J. Llanes-Jurado, L. Gómez-Zaragozá, M. E. Minissi, M. Alcañiz, and J. Marín-Morales, "Developing Conversational Virtual Humans for Social Emotion Elicitation Based on Large Language Models," *Expert Systems with Applications* 246 (2024): 123261, <https://doi.org/10.1016/j.eswa.2024.123261>.
25. C. Vandelandotte, S. Trost, D. Hodgetts, et al., "Increasing Physical Activity Using an Just-in-Time Adaptive Digital Assistant Supported by Machine Learning: A Novel Approach for Hyper-Personalised mHealth Interventions," *Journal of Biomedical Informatics* 144 (2023): 104435, <https://doi.org/10.1016/j.jbi.2023.104435>.
26. H. Yuan, "Toward Real-World Deployment of Machine Learning for Health Care: External Validation, Continual Monitoring, and Randomized Clinical Trials," *Health Care Science* 3, no. 5 (2024): 360–364, <https://doi.org/10.1002/hcs2.114>.
27. M. Daher, J. Koa, P. Boufadel, J. Singh, M. Y. Fares, and J. A. Abboud, "Breaking Barriers: Can ChatGPT Compete With a Shoulder and Elbow Specialist in Diagnosis and Management?," *JSES International* 7, no. 6 (2023): 2534–2541, <https://doi.org/10.1016/j.jseint.2023.07.018>.
28. C. G. Giske, M. Bressan, F. Fiechter, et al., "GPT-4-Based AI Agents—The New Expert System for Detection of Antimicrobial Resistance Mechanisms?," *Journal of Clinical Microbiology* 62, no. 11 (2024): e0068924, <https://doi.org/10.1128/jcm.00689-24>.
29. H. Yuan, L. Kang, Y. Li, and Z. Fan, "Human-in-the-Loop Machine Learning for Healthcare: Current Progress and Future Opportunities in Electronic Health Records," *Medicine Advances* 2, no. 3 (2024): 318–322, <https://doi.org/10.1002/med4.70>.
30. J. Wu, X. Liu, M. Li, et al., "Clinical Text Datasets for Medical Artificial Intelligence and Large Language Models—A Systematic Review," *NEJM AI* 1, no. 6 (2024): e871–e878, <https://doi.org/10.1056/AIra2400012>.