

Coordinated AI agents for advancing healthcare

In the format provided by the
authors and unedited

Supplementary Information

Full list of references to relevant literature, grouped per topic and ordered according to the narrative of the Comment.

Examples of specialist AI models being developed or deployed for tasks in healthcare.

1. Rajpurkar, P., Chen, E., Banerjee, O. & Topol, E. J. AI in health and medicine. *Nat. Med.* **28**, 31–38 (2022).
2. Kraljevic, Z. *et al.* Foresight - Deep Generative Modelling of Patient Timelines using Electronic Health Records.
3. Yang, Z., Mitra, A., Liu, W., Berlowitz, D. & Yu, H. TransformEHR: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records. *Nat. Commun.* **14**, 7857 (2023).
4. Wornow, M., Thapa, R., Steinberg, E., Fries, J. A. & Shah, N. H. EHRSHOT: An EHR Benchmark for Few-Shot Evaluation of Foundation Models. Preprint at <https://arxiv.org/abs/2307.02028> (2023).
5. Boussina, A. *et al.* Impact of a deep learning sepsis prediction model on quality of care and survival. *Npj Digit. Med.* **7**, 1–9 (2024).
6. Lin, C.-S. *et al.* AI-enabled electrocardiography alert intervention and all-cause mortality: a pragmatic randomized clinical trial. *Nat. Med.* 1–10 (2024) doi:10.1038/s41591-024-02961-4.
7. MD, J. N., Wang, J. & Bhattacharyya, A. Where Generative AI Meets Healthcare: Updating The Healthcare AI Landscape. *AI Checkup* <https://aichckup.substack.com/p/where-generative-ai-meets-healthcare> (2023).

Foundation models with generalist capabilities that dynamically adapt to novel tasks and that accommodate flexible multimodal inputs.

8. Moor, M. *et al.* Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).
9. Wu, C., Zhang, X., Zhang, Y., Wang, Y. & Xie, W. Towards Generalist Foundation Model for Radiology by Leveraging Web-scale 2D&3D Medical Data. Preprint at <https://doi.org/10.48550/arXiv.2308.02463> (2023).
10. Tu, T. *et al.* Towards Generalist Biomedical AI. Preprint at <https://doi.org/10.48550/arXiv.2307.14334> (2023).
11. Chen, Z. *et al.* CheXagent: Towards a Foundation Model for Chest X-Ray Interpretation. Preprint at <https://arxiv.org/abs/2401.12208> (2024).
12. Hamamci, I. E. *et al.* A foundation model utilizing chest CT volumes and radiology reports for supervised-level zero-shot detection of abnormalities. Preprint at <https://arxiv.org/abs/2403.17834> (2024).

LLM-based agentic and multi-agentic systems outperforming standard and more sophisticated LLM prompting techniques in complex tasks.

13. Yao, S. *et al.* ReAct: Synergizing Reasoning and Acting in Language Models. Preprint at <https://doi.org/10.48550/arXiv.2210.03629> (2023).
14. Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K. & Yao, S. Reflexion: Language Agents with Verbal Reinforcement Learning.
15. Liu, N. *et al.* From LLM to Conversational Agent: A Memory Enhanced Architecture with Fine-Tuning of Large Language Models. Preprint at <https://doi.org/10.48550/arXiv.2401.02777> (2024).
16. Shi, W. *et al.* EHRAgent: Code Empowers Large Language Models for Few-shot Complex Tabular Reasoning on Electronic Health Records. Preprint at <https://doi.org/10.48550/arXiv.2401.07128> (2024).
17. Wang, Z. *et al.* Unleashing the Emergent Cognitive Synergy in Large Language Models: A Task-Solving Agent through Multi-Persona Self-Collaboration. Preprint at <https://arxiv.org/abs/2307.05300> (2024).
18. Zhu, Y. *et al.* Prompting Large Language Models for Zero-Shot Clinical Prediction with Structured Longitudinal Electronic Health Record Data. Preprint at <https://arxiv.org/abs/2402.01713> (2024).

Performance scaling of agentic systems.

19. Li, J., Zhang, Q., Yu, Y., Fu, Q. & Ye, D. More Agents Is All You Need. Preprint at <https://doi.org/10.48550/arXiv.2402.05120> (2024).
20. Snell, C., Lee, J., Xu, K. & Kumar, A. Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters. Preprint at <https://arxiv.org/abs/2408.03314> (2024).

21. Learning to Reason with LLMs. <https://openai.com/index/learning-to-reason-with-llms/>.
22. Akyürek, E. *et al.* The Surprising Effectiveness of Test-Time Training for Abstract Reasoning. Preprint at <https://arxiv.org/abs/2411.07279> (2024).

Models for analysing medical images to detect signs of disease and to automatically generate a report. See also ref. 11.

23. Hamamci, I. E., Er, S. & Menze, B. CT2Rep: Automated Radiology Report Generation for 3D Medical Imaging. Preprint at <https://arxiv.org/abs/2403.06801> (2024).
24. Hyland, S. L. *et al.* MAIRA-1: A specialised large multimodal model for radiology report generation. Preprint at <https://arxiv.org/abs/2311.13668> (2024).

Expert-level answers to medical questions via conversational agents powered by a LLM or a vision–language model. See also ref. 10.

25. Singhal, K. *et al.* Towards Expert-Level Medical Question Answering with Large Language Models. Preprint at <https://arxiv.org/abs/2305.09617> (2023).
26. Moor, M. *et al.* Med-Flamingo: a Multimodal Medical Few-shot Learner. Preprint at <https://arxiv.org/abs/2307.15189> (2023).
27. Li, C. *et al.* LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. Preprint at <https://arxiv.org/abs/2306.00890> (2023).
28. Luo, Y. *et al.* BioMedGPT: Open Multimodal Generative Pre-trained Transformer for BioMedicine. Preprint at <https://arxiv.org/abs/2308.09442> (2023).
29. Han, T. *et al.* MedAlpaca – An Open-Source Collection of Medical Conversational AI Models and Training Data. Preprint at <https://doi.org/10.48550/arXiv.2304.08247> (2023).
30. Chen, Z. *et al.* MEDITRON-70B: Scaling Medical Pretraining for Large Language Models. Preprint at <https://arxiv.org/abs/2311.16079> (2023).
31. Labrak, Y. *et al.* BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. Preprint at <https://arxiv.org/abs/2402.10373> (2024).
32. Saab, K. *et al.* Capabilities of Gemini Models in Medicine. Preprint at <https://arxiv.org/abs/2404.18416> (2024).

Challenges in the coordination between decentralized multi-agentic systems.

33. Agashe, S., Fan, Y., Reyna, A. & Wang, X. E. LLM-Coordination: Evaluating and Analyzing Multi-agent Coordination Abilities in Large Language Models. Preprint at <https://arxiv.org/abs/2310.03903> (2024).
34. Kim, Y. *et al.* Adaptive Collaboration Strategy for LLMs in Medical Decision Making. Preprint at <https://arxiv.org/abs/2404.15155> (2024).

Multi-agentic frameworks in domains other than healthcare.

35. Xi, Z. *et al.* The Rise and Potential of Large Language Model Based Agents: A Survey. Preprint at <https://doi.org/10.48550/arXiv.2309.07864> (2023).
36. Wu, Q. *et al.* AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. Preprint at <https://arxiv.org/abs/2308.08155> (2023).
37. Durante, Z. *et al.* Agent AI: Surveying the Horizons of Multimodal Interaction. Preprint at <https://doi.org/10.48550/arXiv.2401.03568> (2024).
38. Shavit, Y. *et al.* Practices for Governing Agentic AI Systems.
39. Park, J. S. *et al.* Generative Agents: Interactive Simulacra of Human Behavior. Preprint at <https://doi.org/10.48550/arXiv.2304.03442> (2023).
40. Hong, S. *et al.* MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework. Preprint at <https://arxiv.org/abs/2308.00352> (2023).
41. Qian, C. *et al.* Communicative Agents for Software Development. Preprint at <https://doi.org/10.48550/arXiv.2307.07924> (2023).
42. Masterman, T., Besen, S., Sawtell, M. & Chao, A. The Landscape of Emerging AI Agent Architectures for Reasoning, Planning, and Tool Calling: A Survey. arXiv, <https://arxiv.org/abs/2404.11584> (2024).

Example of a multi-agentic system for healthcare.

43. Mukherjee, S. *et al.* Polaris: A Safety-focused LLM Constellation Architecture for Healthcare. Preprint at <https://arxiv.org/abs/2403.13313> (2024).

Algorithmic monoculture.

- 44. Kleinberg, J. & Raghavan, M. Algorithmic monoculture and social welfare. *Proc. Natl. Acad. Sci.* **118**, e2018340118 (2021).
- 45. Bommasani, R., Creel, K. A., Kumar, A., Jurafsky, D. & Liang, P. Picking on the Same Person: Does Algorithmic Monoculture lead to Outcome Homogenization? Preprint at <https://doi.org/10.48550/arXiv.2211.13972> (2022).

Communication overhead in federated learning.

- 46. Conquering the Communication Constraints to Enable Large Pre-Trained Models in Federated Learning. <https://arxiv.org/html/2210.01708v3>.

Improvement of factual accuracy and reasoning.

- 47. Du, Y., Li, S., Torralba, A., Tenenbaum, J. B. & Mordatch, I. Improving Factuality and Reasoning in Language Models through Multiagent Debate. Preprint at <https://arxiv.org/abs/2305.14325> (2023).
- 48. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine | NEJM. <https://www.nejm.org/doi/10.1056/NEJMSr2214184>.
- 49. Johri, S. *et al.* Guidelines For Rigorous Evaluation of Clinical LLMs For Conversational Reasoning. <https://medrxiv.org/lookup/doi/10.1101/2023.09.12.23295399> (2023) doi:10.1101/2023.09.12.23295399.

Maximization of the effectiveness of agentic systems and minimization of physician burnout.

- 50. Jha, A. K. *et al.* A crisis in health care: a call to action on physician burnout, <https://content.sph.harvard.edu/wwwhsph/sites/21/2019/01/PhysicianBurnoutReport2018FINAL.pdf>
- 51. Agarwal, N., Moehring, A., Rajpurkar, P. & Salz, T. Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology.
- 52. Tang, J. S. N. *et al.* Impact of Different Artificial Intelligence User Interfaces on Lung Nodule and Mass Detection on Chest Radiographs. *Radiol. Artif. Intell.* **5**, e220079 (2023).
- 53. Yu, F. *et al.* Heterogeneity and predictors of the effects of AI assistance on radiologists. *Nat. Med.* **30**, 837–849 (2024).

Regulation of AI systems and assessment of their biases.

- 54. Meskó, B. & Topol, E. J. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *npj Digit. Med.* **6**, 120 (2023).
- 55. Yao, S., Shinn, N., Razavi, P. & Narasimhan, K. τ -bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains. *arXiv.org* <https://arxiv.org/abs/2406.12045v1> (2024).
- 56. Schmidgall, S. *et al.* AgentClinic: a multimodal agent benchmark to evaluate AI in simulated clinical environments. *arXiv.org* <https://arxiv.org/abs/2405.07960v3> (2024).

Empathy with AI agents.

- 57. Topol, E. J. Machines and empathy in medicine. *The Lancet* **402**, 1411 (2023).
- 58. Tu, T. *et al.* Towards Conversational Diagnostic AI. Preprint at <https://arxiv.org/abs/2401.05654> (2024).
- 59. AI, H. Hippocratic AI Announces Collaboration with NVIDIA to Develop Super-Low-Latency “Empathy Inference” for One of the World’s First Generative AI-Powered Healthcare Agents. *GlobeNewswire News Room* <https://www.globenewswire.com/news-release/2024/03/18/2848236/0/en/Hippocratic-AI-Announces-Collaboration-with-NVIDIA-to-Develop-Super-Low-Latency-Empathy-Inference-for-One-of-the-World-s-First-Generative-AI-Powered-Healthcare-Agents.html> (2024).

Affective visual avatars for agents.

- 60. Hasan, M., Ozel, C., Potter, S. & Hoque, E. SAPIEN: Affective Virtual Agents Powered by Large Language Models. in *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)* 1–3 (2023). doi:10.1109/ACIIW59127.2023.10388188.
- 61. Loizou, M. *et al.* Designing, implementing and testing an intervention of affective intelligent agents in nursing virtual reality teaching simulations—a qualitative study. *Front. Digit. Health* **6**, (2024).

62. Chen, Z., Cao, J., Chen, Z., Li, Y. & Ma, C. EchoMimic: Lifelike Audio-Driven Portrait Animations through Editable Landmark Conditions. arXiv, <https://doi.org/10.48550/arXiv.2407.08136> (2024).
63. Digital Twin APIs, <https://www.tavus.io>.