npj | precision oncology

Perspective

Published in partnership with The Hormel Institute, University of Minnesota

https://doi.org/10.1038/s41698-024-00573-2

Large language models and multimodal foundation models for precision oncology

Daniel Truhn [1], Jan-Niklas Eckardt[2,3], Dyke Ferber [4,5] & Jakob Nikolas Kather[2,3,4,5]

The technological progress in artificial intelligence (AI) has massively accelerated since 2022, with far-reaching implications for oncology and cancer research. Large language models (LLMs) now perform at human-level competency in text processing. Notably, both text and image processing networks are increasingly based on transformer neural networks. This convergence enables the development of multimodal AI models that take diverse types of data as an input simultaneously, marking a qualitative shift from specialized niche models which were prevalent in the 2010s. This editorial summarizes these developments, which are expected to impact precision oncology in the coming years.

The volume of patient-specific data in oncology is rapidly expanding. This is due to the widespread introduction of electronic health records (EHRs), advances in medical imaging, and the integration of large-scale genomic analyses into clinical routine. Effective use of this large amount of data is important for ensuring the optimal treatment for cancer patients. Artificial intelligence (AI) and machine learning (ML) have shown promise in helping healthcare professionals to process such data.

AI's application in domains like oncology has experienced periodic surges of activity. Prior to 2012, AI technologies played a marginal role in oncology research. Computer-based data analysis studies mostly relied on classical ML algorithms with a modest model complexity. 2012 marked a turning point in computer-based data analysis: image processing, a notoriously difficult task, was suddenly made much easier with the advent of convolutional neural networks (CNNs). This technological shift was subsequently integrated into medical research[1], most notably evidenced by a 2017 publication demonstrating neural network performance on par with human experts across large image datasets[2]. In parallel, hardware improvements have gradually lowered the computational barriers for training and deploying resource-intensive models, thus broadening the user base capable of developing and refining AI algorithms[3].

Between 2012 and 2022, neural networks were applied in numerous studies that primarily focused on the analysis of oncological imaging or text[1]. Regulatory bodies in the United States and the European Union approved a number of specialized AI-based tools for cancer, particularly in radiological and pathological image analysis. However, high-profile and ambitious initiatives, such as IBM Watson, did not achieve their projected outcomes[4]. This decade can be viewed as a stabilization phase for AI applications in oncology, marked by incremental improvements and specialized uses, predominantly in image analysis[5]. This landscape changed in 2022 and 2023 with the advent of two key innovations: large language models (LLMs) and multimodal AI models.

LLMs are deep learning models that serve the purpose of both processing and generating primarily text-based data[6]. The training data for these models is a large and diverse amount of text, usually sourced from the internet and commercial data providers, and can include diverse types of medical data[7]. While potentially this purpose can also be fulfilled by various model architectures, the most successful models have recently relied on transformer-based architectures pertaining to their attention mechanisms[8]. Their training process is autoregressive, which means that the model is trained to predict subsequent tokens in a sequence (similar to words in a sentence). Notably, model performance scales with size, i.e. number of parameters, and larger models show emergent behavior[9]: They acquire an understanding of concepts underlying the training data without having been explicitly trained on this. When LLMs are applied on new tasks without explicit training, this is referred to as a "zero-shot" application. The LLM Generative Pre-trained Transformer (GPT) 3.5 by OpenAI gained widespread attention in 2022 because it was made available as a chatbot in the "chatGPT" user interface and demonstrated impressive conversational skills. It was later succeeded by GPT-4, which displayed much-improved knowledge retrieval and logical reasoning capacities with far fewer hallucinations[7].

[1]Department of Diagnostic and Interventional Radiology, University Hospital Aachen, Aachen, Germany. [2]Department of Internal Medicine I, University Hospital Carl Gustav Carus, Technical University Dresden, Dresden, Germany. [3]Else Kroener Fresenius Center for Digital Health, Technical University Dresden, Dresden, Germany. [4]National Center for Tumor Diseases (NCT), Heidelberg University Hospital, Heidelberg, Germany. [5]Department of Medical Oncology, Heidelberg University Hospital, Heidelberg, Germany.
e-mail: jakob-nikolas.kather@alumni.dkfz.de

THE HORMEL INSTITUTE
UNIVERSITY OF MINNESOTA

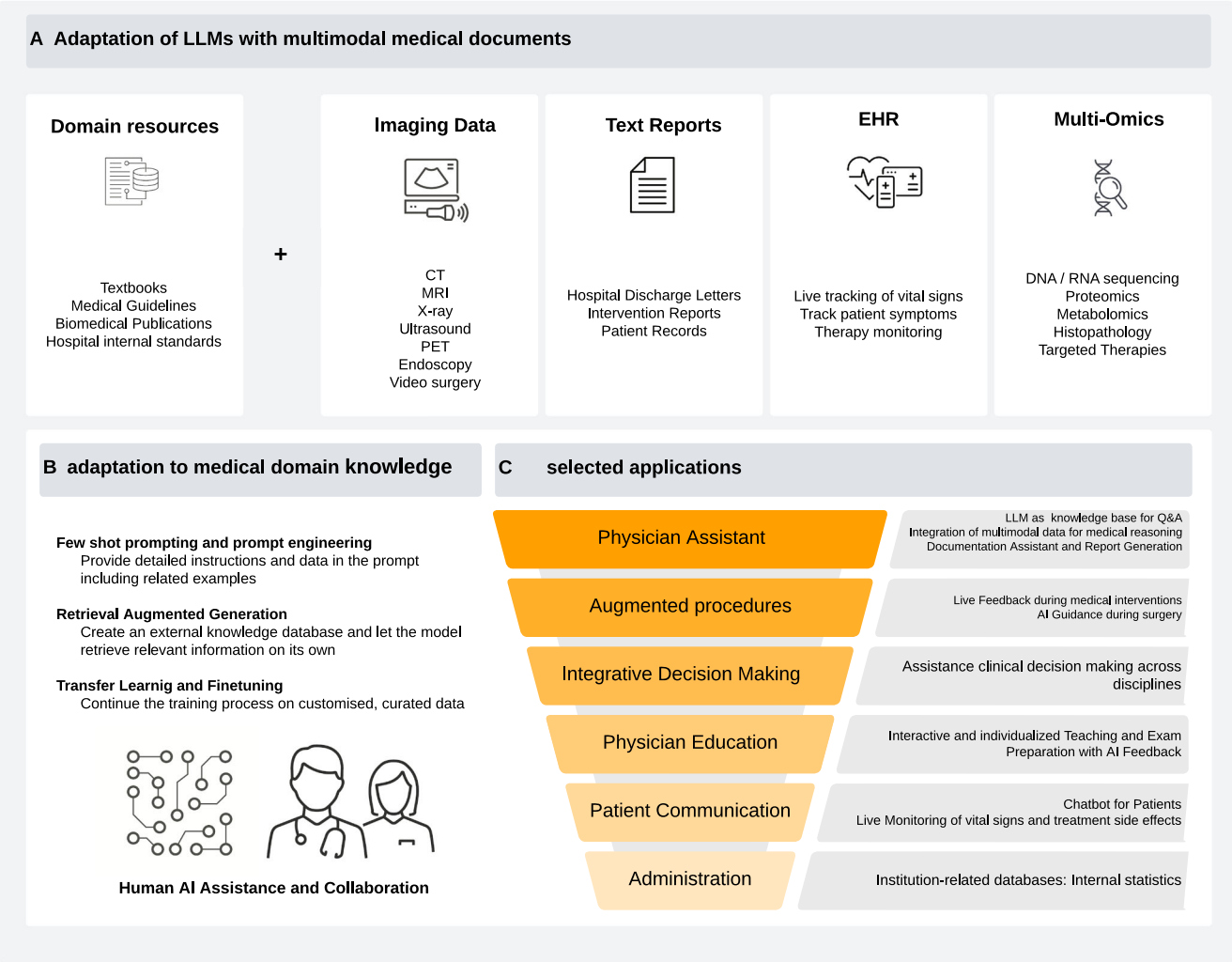

Fig. 1 | Overview of medical adaptations of LLMs. A Integrating information from various sources comprising domain knowledge from databases and patient related documents including different modalities (imaging, text, tabular, numerical data). **B** Possible model adaptation strategies include: Combining user prompts with relevant source documents and examples (few-shot prompting, top). Retrieval augmented generation (RAG), in which the information from A gets processed by a separate embedding model and is stored in a database. Relevant information can then be retrieved based on similarity measures with a user input (middle). Documents can be used to continue the training process and adapt the model parameters for specific use cases (bottom). Procedures are sorted by complexity and computational cost in increasing order. **C** Selected possible use cases for LLMs in clinical routine.

A common application area under exploration for these LLMs is healthcare[7]. LLMs can be applied to medical problems through different approaches. One approach is to train these models specifically on medical data (Fig. 1A). One of the first language models to be successfully transferred to the medical domain was Bio-BERT, which showed robust capabilities for biomedical text mining[10]. Also, Google's LLM "PALM" was fine-tuned on medical training data, resulting in Med-PaLM, which outperformed its previous version in medical use cases[11]. Recently, Google has introduced the next iteration, Med-PaLM 2, which scored a high 86.5% in the US Medical Licensing Exam (USMLE)[12]. Solving USMLE questions is a common benchmark test for LLMs, yet it is of limited use for real-world applications. However, fine-tuned LLMs have been shown to solve real-world problems such as predicting clinical outcomes just based on unstructured text data in EHRs[13]. Other ways to apply LLMs on medical problems do not require fine-tuning: Generalist LLMs can often be applied to medical tasks by just using a detailed input prompt. Another alternative is the use of "Retrieval Augmented Generation" (RAG) by which domain knowledge can be provided to a trained LLM in machine-readable format (Fig. 1B).

Today's LLMs are transformer neural networks. This network architecture is well suited for almost any type of data, and enables multimodality. Multimodal AI systems are capable of interpreting multiple types of data together, such as textual and image data. Their development and validation require collaborative efforts between a number of disciplines including but not limited to medical experts in diagnostic specialties such as radiology or pathology and specialties such as surgery or medicine as well as technology experts both in software and hardware. Multimodal AI systems have been evaluated for various applications in precision oncology, such as outcome predictions[14,15]. However, more scientific evidence is required to ensure that LLMs and multimodal models provide quantifiable benefits in oncology.

Whenever a model is pre-trained on large and diverse tasks, and is subsequently applied to specialized tasks, it can be referred to as a "foundation model"[16]. Foundation models reduce the data requirements for specialized tasks, for example in predicting diseases from retinal photographs[17]. For instance, linking images from chest X-rays to corresponding report text data, and foundation models can alleviate the need for time-consuming and laborious manual annotation while preserving human-level accuracies and outperforming supervised methods[18]. In clinical practice, such models may be deployed in the form of chatbot assistants that can aid diagnosis in an interactive manner[19]. Similar examples exist in pathology, where large image datasets are linked with contextual knowledge and case-specific information yielding high performance in disease

detection as well as biomarker prediction and can also inform further diagnostic procedures such as additional stains[20,21]. Furthermore, early generalist models have been introduced which show consistently high performance across a variety of medical domains and tasks integrating knowledge from diverse domains[22,23]. Given the resource-intensive nature of training models with parameter sizes in the billions, there is a trend towards improving model efficiency by reducing model size while retaining model performance. Recent advances with open-sourced models yield the perspective of de novo model training and development at much lower financial and computational burdens[24].

As foundation models diversify their capabilities, they open up new avenues for their potential application in oncology and cancer research, such as multimodal diagnostics and drug discovery. However, to fully unlock the potential of foundation models in oncology and cancer research, several challenges must be addressed. Firstly, the underlying data the model is trained on has to be carefully assessed for quality, quantity, and diversity[25]. Secondly, the design of systems which integrate foundation models should be guided not only by experts in computer science, but also by medical professionals and patient advocates as well as the broader scientific community. Thirdly, the integration of such models in operable clinical software systems faces legal and regulatory challenges because these models require approval as medical devices[26]. Fourthly, ongoing model evaluation, validation and improvement are important to maintain quality, safety, and usefulness in light of the accelerating pace with which scientific discoveries are translated into novel medicines and guidelines. Lastly, a prominent concern of AI models based on neural network architecture is their often criticized lack of interpretability which earned them the term 'black boxes'[27,28]. While substantial progress has been made in model explainability for image-related tasks, fewer studies address explainability in text processing or multimodal tasks in medicine[29].

Overall, the advancements in LLMs and multimodal models have the potential to impact the practice of oncology through many different applications (Fig. 1C). This "Collection" in "npj Precision Oncology" aims to collect articles that provide solid empirical evidence for applications of these models in precision oncology.

Received: 21 September 2023; Accepted: 12 March 2024;
Published online: 22 March 2024

References

1. Shmatko, A., Ghaffari Laleh, N., Gerstung, M. & Kather, J. N. Artificial intelligence in histopathology: enhancing cancer research and clinical oncology. *Nat. Cancer* **3**, 1026–1038 (2022).
2. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
3. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019).
4. Schmidt, C. M. D. Anderson breaks with IBM Watson, raising questions about artificial intelligence in oncology. *J. Natl Cancer Inst*. **109** (2017).
5. He, J. et al. The practical implementation of artificial intelligence technologies in medicine. *Nat. Med.* **25**, 30–36 (2019).
6. Clusmann, J. et al. The future landscape of large language models in medicine. *Commun. Med.* **3**, 141 (2023).
7. Bubeck, S. et al. Sparks of artificial general intelligence: early experiments with GPT-4. Preprint at https://arxiv.org/abs/2303.12712 (2023).
8. Vaswani, A. et al. Attention is all you need. Preprint at https://arxiv.org/abs/1706.03762 (2017).
9. Wei, J. et al. Emergent abilities of large language models. Preprint at https://arxiv.org/abs/2206.07682 (2022).
10. Lee, J. et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).
11. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
12. Singhal, K. et al. Towards expert-level medical question answering with large language models. Preprint at https://arxiv.org/pdf/2305.09617.pdf (2023).
13. Jiang, L. Y. et al. Health system-scale language models are all-purpose prediction engines. *Nature* **619**, 357–362 (2023).
14. Chen, R. J. et al. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell* **40**, 865–878.e6 (2022).
15. Lipkova, J. et al. Artificial intelligence for multimodal data integration in oncology. *Cancer Cell* **40**, 1095–1110 (2022).
16. Moor, M. et al. Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).
17. Zhou, Y. et al. A foundation model for generalizable disease detection from retinal images. *Nature* https://doi.org/10.1038/s41586-023-06555-x (2023).
18. Tiu, E. et al. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. *Nat. Biomed. Eng.* **6**, 1399–1406 (2022).
19. Thawkar, O. et al. XrayGPT: Chest radiographs summarization using medical vision-language models. Preprint at https://arxiv.org/abs/2306.07971 (2023).
20. Vorontsov, E. et al. Virchow: a million-slide digital pathology foundation model. Preprint at https://arxiv.org/abs/2309.07778 (2023).
21. Lu, M. Y. et al. A foundational multimodal vision language AI assistant for human pathology. Preprint at https://arxiv.org/abs/2312.07814 (2023).
22. Zhang, K. et al. BiomedGPT: a unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. Preprint at https://arxiv.org/abs/2305.17100 (2023).
23. Tu, T. et al. Towards generalist biomedical AI. Preprint at https://ai.nejm.org/doi/full/10.1056/AIoa2300138 (2023).
24. 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).
25. de Hond, A. A. H. et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. *NPJ Digit Med.* **5**, 2 (2022).
26. Gilbert, S., Harvey, H., Melvin, T., Vollebregt, E. & Wicks, P. Large language model AI chatbots require approval as medical devices. *Nat. Med*. https://doi.org/10.1038/s41591-023-02412-6 (2023).
27. Castelvecchi, D. Can we open the black box of AI? *Nature* **538**, 20–23 (2016).
28. Savage, N. Breaking into the black box of artificial intelligence. *Nature* https://doi.org/10.1038/d41586-022-00858-1 (2022).
29. Tjoa, E. & Guan, C. A survey on explainable artificial intelligence (XAI): toward medical XAI. *IEEE Trans. Neural Netw. Learn Syst.* **32**, 4793–4813 (2021).

Author contributions

All authors co-wrote the manuscript and made the decision to submit this article for publication.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

J.N.K. declares consulting services for Owkin, France; DoMore Diagnostics, Norway; Panakeia, UK; Scailyte, Switzerland; Mindpeak, Germany; and MultiplexDx, Slovakia. Furthermore, he holds shares in StratifAI GmbH, Germany, has received a research grant by GSK, and has received honoraria by AstraZeneca, Bayer, Eisai, Janssen, MSD, BMS, Roche, Pfizer and Fresenius. Furthermore, J.N.K. is a Deputy

Editor at npj Precision Oncology. J.-N.E. declares research funding from Novartis Oncology, honoraria from Janssen, AstraZeneca, Amgen, and co-ownership of Cancilico GmbH. No other potential conflicts of interest are disclosed by any of the authors.

Additional information

Correspondence and requests for materials should be addressed to Jakob Nikolas Kather.