npj | digital medicine

Editorial

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01941-3

Incorporating large language models as clinical decision support in oncology: the Woollie model

Check for updates

Integrating large language models (LLMs) into oncology holds promise for clinical decision support. Woollie is an LLM recently developed by Zhu et al., fine-tuned using radiology impression notes from Memorial Sloan Kettering Cancer Center and externally validated on UCSF oncology datasets. This methodology prioritizes data accuracy, preempts catastrophic forgetting, and demonstrates unparalleled rigor in predicting the progression of various cancer types. This work establishes a foundation for reliable, scalable, and equitable applications of LLMs in oncology.

The effectiveness of oncology treatments depends on how cancer responds, as observed through radiological or pathological assessments. Tracking tumor regression in response to chemotherapy via serial radiologic imaging is critical for assessing treatment efficacy and guiding ongoing clinical management[1]. However, these important data points are frequently documented as real-world data in non-standardized and unstructured formats, making them difficult to access and interpret—especially when leveraging Large Language Models (LLMs) in oncology[2]. Besides non-standardized and unstructured formats of documenting tumor progression, subspecialty knowledge barriers and privacy concerns related to the deployment of closed-source LLMs in clinical settings complicate the integration of LLMs in oncology[3].

Still, the incorporation of LLMs in oncology could serve as a helpful decision support tool for clinicians. Rapidly expanding medical literature presents a challenge to oncologists seeking optimized and targeted cancer therapies for their patients. Equally, in gathering large-scale data about tumor progression, LLMs can facilitate large-scale, systematic analysis of tumor progression data, potentially informing both individualized care and public health strategies by identifying patterns in metastasis and treatment efficacy[4]. Developing and facilitating prompts for LLMs to derive clinical factors has been proven efficient in extracting and collating crucial information from large medical records[5]. In streamlining data extraction from clinical reports (such as radiology interpretations or progress notes) in a reproducible manner, LLMs reduce manual labor and thereby alleviate time constraints on clinicians[6].

The Woollie training model

Considering privacy concerns surrounding the implementation of LLMs in clinical decision-making, the increasing need for specialized oncology knowledge, and the complexity of extracting and collating real-world data, Menglei Zhu et al. developed Woollie, a dedicated LLM that is trained on real-world data from Memorial Sloan Kettering Cancer Center and

is thereby specialized for interpreting oncological radiology reports[2]. In their paper entitled "Large Language Model Trained on Clinical Oncology Data Predicts Cancer Progression," Zhu et al. demonstrate that this model surpasses existing LLMs in terms of medical knowledge benchmarks, including PubMedQA, MedMCQA, and USMLE[7]. In addition, Zhu et al. extended their validation to include an independent dataset of 600 radiology impressions involving 600 unique patients from the University of California, San Francisco medical center. Given the complexity and breadth of oncological knowledge, the authors enhanced Woollie's analytical skills through a stacked alignment process. Through this process, the LLM is trained on various and interdependent fields of understanding cancer care and cancer progression: the LLM is first trained on a foundational model and then fine-tuned with increasingly domain-specific databases and validation tools, such as the most recent medical benchmarks and external datasets. This approach is necessary given the complexity and depth of oncological data and ensures that increased specialization preserves foundational knowledge. Resultingly, the model demonstrated excellent performance for predicting the progression of various cancer types, including lung, breast, and prostate cancer (AUROC 0.97 and 0.88 on internal and external validation data, respectively).

A strategy against catastrophic forgetting

Catastrophic forgetting occurs when an LLM loses previously learned knowledge while acquiring new knowledge for achieving a satisfactory performance in downstream tasks, especially in fields requiring complex subspecialty knowledge[8]. A core contribution of Zhu et al.'s work is their meticulous training approach: by employing the aforementioned stacked alignment process, they minimize catastrophic forgetting while expanding the model's specialized knowledge base[8]. This training model ensures that Woollie LLMs preserves general domain competencies of reasoning, conversation, and information extraction while building upon each successive model iteration to enhance medical domain proficiency. This capacity is critical in incorporating LLMs in oncology, where confidently delivered but incorrect information can have severe consequences on patient care and trust. When combined with persistent attainment of clinical performance benchmarks, this robust training strategy suggests the safe integration of AI models into clinical decision-making for cancer progression prediction.

Potential for scalable cancer data

Given recent advances in training LLMs to retain broad knowledge while gaining domain-specific expertise in oncology, models like Woollie could be utilized alongside other LLMs to scale and systematize knowledge about cancer progression across different cancer centers[9]. Verification and validation of Woollie data across multiple cancer sites will strengthen the generalizability of this model. When aligned with established AI governance frameworks for clinical care, operations, and research in oncology, the Woollie model can be scaled responsibly and ethically across multiple institutions, both nationally and globally[10]. By thoroughly integrating a wide range of clinical protocols and perspectives across cancer care institutions, scaled LLMs can limit non-evidence-based variation in clinical recommendations and therefore promote more equitable cancer care worldwide. From a public health standpoint, access to such systematized data provides an opportunity to enhance population-level insights into cancer progression.

Conclusion

Efforts to support and systematize clinical decision-making must preserve clinical accuracy, as sustaining patients' trust and confidence is critical, especially in cancer care. Due to both robust model training and external validation against UCSF datasets, Woollie is an LLM that prioritizes data accuracy and safeguards against catastrophic forgetting in delivering clinical decision support for predicting cancer progression. This advancement paves the way for making cancer care more scalable and equitable.

Data availability

No datasets were generated or analysed during the current study.

Kimia Heydari[1] ✉, **Elizabeth J. Enichen**[1], **Ben Li**[2,3] **& Joseph C. Kvedar**[1,4]
[1]Harvard Medical School, Boston, MA, USA. [2]Division of Vascular Surgery, University of Toronto, Toronto, ON, Canada. [3]Temerty Centre for Artificial Intelligence Research and Education in Medicine, University of Toronto, Toronto, ON, Canada. [4]Massachusetts General Hospital, Harvard University, Boston, MA, USA.
✉e-mail: kimiaheydari@hms.harvard.edu

Received: 31 July 2025; Accepted: 8 August 2025
Published online: 18 August 2025

References

1. Ko, CC., Yeh, LR. & Kuo, YT. et al. Imaging biomarkers for evaluating tumor response: RECIST and beyond. *Biomark Res.* **9**, 52, https://doi.org/10.1186/s40364-021-00306-8 (2021).
2. Zhu, M. et al. Large language model trained on clinical oncology data predicts cancer progression. *Npj Digit. Med.* **8**, 397 (2025).
3. Chen, S. et al. Use of artificial intelligence Chatbots for cancer treatment information. *JAMA Oncol.* **9**, 1459–1462 (2023).
4. Fountzilas, E., Pearce, T., Baysal, M. A., Chakraborty, A. & Tsimberidou, A. M. Convergence of evolving artificial intelligence and machine learning techniques in precision oncology. *NPJ Digit. Med.* **8**, 75 (2025).
5. Choi, H. S., Song, J. Y., Shin, K. H., Chang, J. H. & Jang, B. S. Developing prompts from large language model for extracting clinical information from pathology and ultrasound reports in breast cancer. *Radiat. Oncol. J.* **41**, 209–216 (2023).
6. Huang, J. et al. A critical assessment of using ChatGPT for extracting structured data from clinical notes. *npj Digit. Med.* **7**, 106 (2024).
7. Singhal, K. et al. Toward expert-level medical question answering with large language models. *Nat. Med.* **31**, 943–950 (2025).
8. Mermillod, M., Aurélia, B. & Patrick, B. The stability-plasticity dilemma: Investigating the continuum from catastrophic forgetting to age-limited learning effects. *Front. Psychol.* **4**, 504 (2013).
9. Lavery, J. A. et al. A scalable quality assurance process for curating oncology electronic health records: the project GENIE biopharma collaborative approach. *JCO Clin. Cancer Inform.* **6**, e2100105 (2022).
10. Stetson, P. D. et al. Responsible artificial intelligence governance in oncology. *npj Digit. Med.* **8**, 407 (2025).

Acknowledgements

This editorial did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author contributions

K.H. wrote the first draft of the manuscript. E.J.E., B.L. and J.C.K. provided time-sensitive, critical revisions. All authors have read and approved the final manuscript.

Competing interests

J.C.K. is the editor-in-chief of npj Digital Medicine. All other authors declare no competing interests.