npj | digital medicine

Brief communication

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01671-6

Patient perceptions of empathy in physician and artificial intelligence chatbot responses to patient questions about cancer

David Chen[1,2], Kabir Chauhan[1], Rod Parsa[3], Zhihui Amy Liu[4,5], Fei-Fei Liu[1,6], Ernie Mak[7,8], Lawson Eng[9,10], Breffni Louise Hannon[7,11], Jennifer Croke[1,6], Andrew Hope[1,6], Nazanin Fallah-Rad[9], Phillip Wong[1,6] & Srinivas Raman[1,6] ✉

Artificial intelligence chatbots can draft empathetic responses to cancer questions, but how patients perceive chatbot empathy remains unclear. Here, we found that people with cancer rated chatbot responses as more empathetic than physician responses. However, differences between patient and physician perceptions of empathy highlight the need for further research to tailor clinical messaging to better meet patient needs. Chatbots may be effective in generating empathetic template responses to patient questions under clinician oversight.

Large language models (LLM) applications serve as promising artificial intelligence (AI) tools to address administrative burden and support clinical decision-making in medicine[1]. Conversational LLM chatbots can provide quality and empathetic responses to questions in general medicine[2] and oncology[3,4], as evaluated by clinicians. As chatbots are deployed in patient-facing roles, there remains debate about whether patients also perceive that chatbots can demonstrate empathy, a core competency in medicine[5]. Empathy, defined as the ability to understand and share the feelings of others, is central to establishing trustworthy patient-provider relationships which have been linked to improved patient outcomes[6]. However, patients, rather than clinicians, should serve as the benchmark for determining whether their experiences have been understood, shared, and addressed[7].

State-of-the-art methods to design empathetic chatbots primarily involve integrating emotional intelligence into LLMs by employing specialized training regimens, model architectures, and attention mechanisms to improve context-dependent empathetic reasoning[8–12]. These approaches have demonstrated significant advancements in generating contextually appropriate and emotionally attuned chatbot outputs but remain limited due to the computationally prohibitive nature of training and fine-tuning large-scale LLMs. Moreover, training increasingly complex LLMs on larger datasets may be subject to scaling laws of diminishing improvements in empathic responses[13]. An alternative approach to design more empathetic LLM applications involves multi-step processing of emotional dialogue that involves recognition of emotion in user input followed by integration of appropriate emotions in the generated response[14]. Adapting chain of thought prompting to elicit emotional reasoning may be an effective method to improve the human perception of empathy of foundational LLMs while decreasing resource demands, but this has not been evaluated by real-world patients[15].

This study evaluated the empathy of chatbots compared to physicians in responding to oncology-related patient questions from the perspective of people with cancer and tested the chain-of-thought prompting method to elicit empathy in chatbot responses.

In total, 45 patient participants completed the survey. Survey participants were primarily White (40/45, 88.89%), identified as male (33/45, 73.33%), were above 65 years old (32/45, 71.11%), and post-secondary educated (35/45, 77.78%) (Supplementary Table 2). Descriptive statistics are shown in Supplementary Table 3.

Responses generated by Claude V1, Claude V2, and Claude V2 with CoT were all rated by participants as higher in empathy compared to physician responses (Fig. 1). The best-performing AI chatbot, Claude V2 with CoT (mean, 4.11 [95% CI, 3.99–4.22]), was rated as more empathetic

[1]Princess Margaret Cancer Centre, Radiation Medicine Program, Toronto, ON, Canada. [2]Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada. [3]Michael G. DeGroote School of Medicine, McMaster University, Hamilton, ON, Canada. [4]Department of Biostatistics, Princess Margaret Cancer Centre, University Health Network, Toronto, ON, Canada. [5]Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada. [6]Department of Radiation Oncology, University of Toronto, Toronto, ON, Canada. [7]Department of Supportive Care, University Health Network, Toronto, ON, Canada. [8]Department of Family & Community Medicine, University of Toronto, Toronto, ON, Canada. [9]Division of Medical Oncology and Hematology, Department of Medicine, Princess Margaret Cancer Centre/University Health Network Toronto, Toronto, ON, Canada. [10]Division of Medical Oncology, Department of Medicine, University of Toronto, Toronto, ON, Canada. [11]Department of Medicine, University of Toronto, Toronto, ON, Canada. ✉e-mail: Srinivas.Raman@bccancer.bc.ca

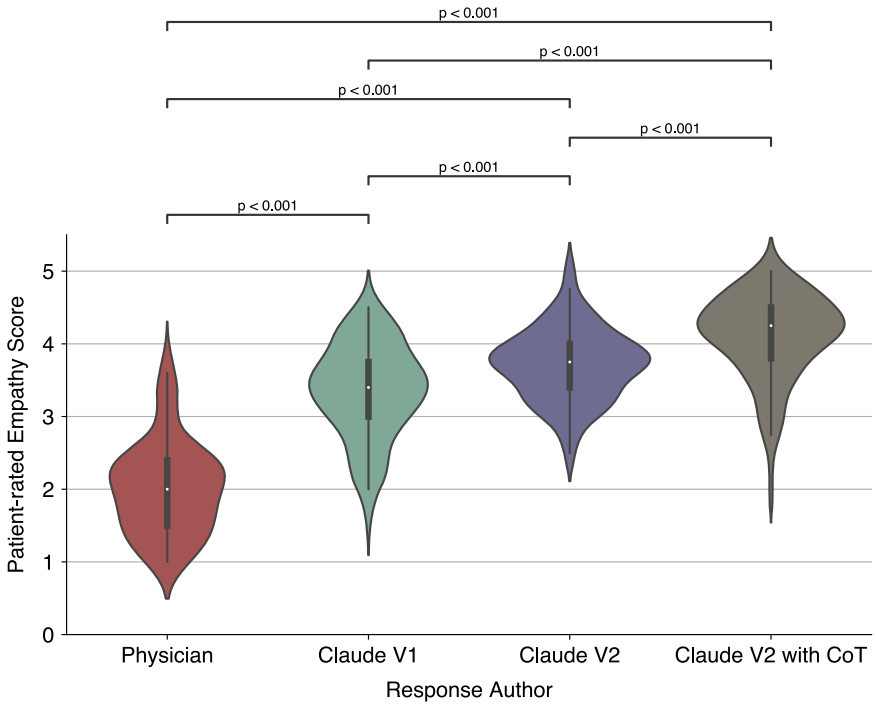

Fig. 1 | Empathy rating of physician and chatbot responses to patient questions about cancer by survey participants. People with cancer ($n = 45$) rated the overall empathy of both physician and chatbot (Claude V1, Claude V2, Claude V2 with CoT) responses.

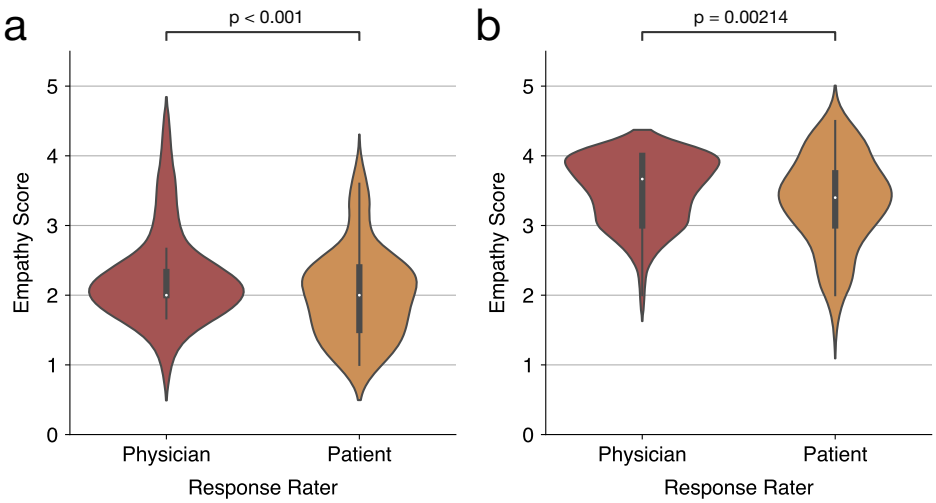

Fig. 2 | Empathy rating of physician and Claude V1 chatbot responses to patient questions about cancer by people with cancer ($n = 45$) and physicians ($n = 3$). **A** Empathy rating of physician responses by people with cancer and physicians. **B** Empathy rating of Claude V1 chatbot responses by people with cancer and physicians.

than Claude V2 (mean, 3.72 [95% CI, 3.62–3.81]; $P < 0.001$; $d = 3.46$), Claude V1 (mean, 3.35 [95% CI, 3.23–3.48]; $P < 0.001$; $d = 3.01$), and physicians (mean, 2.01 [95% CI, 1.88-2.13]; $P < 0.001$; $d = 2.11$) (Fig. 1, Supplementary Fig. 1).

Participant ratings of empathy were consistently lower than physician ratings of empathy for the same set of responses generated by physicians (2.01 [95% CI, 1.88–2.13] vs mean, 2.24 [95% CI, 2.11–2.37]; $P = 0.00214$; $d = -0.36$) (Fig. 2A) and Claude V1 (mean, 3.35 [95% CI, 3.23–3.48] vs 3.51 [95% CI, 3.41–3.60]; $P < 0.001$; $d = -0.28$) (Fig. 2B). Participant-rated empathy was moderately correlated with physician-rated empathy of Claude V1 and physician responses (Supplementary Fig. 2).

The word count of Claude V1 (mean, 192.45 [95% CI, 183.35–201.55]; $P < 0.001$), Claude V2 (mean, 152.31 [95% CI, 147.46–157.16]; $P < 0.001$), and Claude V2 with CoT (mean, 186.72[95% CI, 183.08–190.36]; $P < 0.001$) responses were increased compared to physician responses (mean, 99.71 [95% CI, 74.34–125.08]) (Supplementary Fig. 3). Word count was associated with participant-rated empathy for physician and ClaudeV1 but not Claude V2 and Claude V2 with CoT responses based on correlation (Supplementary Fig. 4) and OLS analyses (Supplementary Table 4).

The reading grade level of Claude V1 (mean, 9.66 [95% CI, 9.19–10.12]; $P < 0.001$), Claude V2 (mean, 8.79 [95% CI,8.49–9.08]; $P < 0.001$), and Claude V2 with CoT (mean, 8.55 [95% CI, 8.27–8.82]; $P < 0.001$) responses were increased compared to physician responses (mean, 8.13 [95% CI, 7.50–8.76]) (Supplementary Fig. 5). Readability of physician or chatbot responses was not correlated with participant-rated empathy (Supplementary Fig. 6).

In this cross-sectional study, we observed that patient participants rated responses authored by Claude V1 as more empathetic than physicians; this is consistent with previous results from the physician perspective[3]. We hypothesize that chatbots can consistently provide empathetic responses by appropriately responding to emotional cues in patient questions and offering supportive language without the time pressures of clinical workload or emotional variability due to human-oriented stressors that physicians may experience. We caution that LLMs generate outputs that reflect

perceived empathy through linguistic mimicry based on probabilistic text prediction from learned patterns, rather than emotional cognition or empathic experiences inherent to human interactions[16]. Fine-tuning of chatbot-generated outputs to prioritize empathy may inadvertently impact medical accuracy. However, in the same dataset used in this study, we highlight that Chen et al. found that physicians rated responses generated by chatbots as higher quality, more empathetic, and more readable than responses generated by physicians[3]. Our current study extends these findings by assessing empathy from the perspective of people with cancer, confirming that both physicians and people with cancer may perceive chatbots as more empathetic compared to physicians.

Compared to physicians, participants rated the same set of Claude V1 and physician responses as less empathetic, suggesting that there may be differences in perception of empathy from the patient and physician perspective. Given previous findings of discordance between physician and patient perceptions of empathy[17], we speculate that physicians and patients may prioritize different elements of message content and delivery in clinical care. Further research is needed to characterize the prioritization of messaging elements across diverse patient demographics to evaluate how patient-facing chatbots convey empathy in real-world clinical oncology scenarios[18].

We assessed the Claude LLM due to prior evidence demonstrating its superior empathy for cancer-related inquiries[3] and to minimize confounding when comparing LLMs with distinct architectures and training regimens. Despite our standalone assessment of one family of LLMs, benchmarks of chain of thought prompting compared to the baseline LLMs for complex reasoning have shown that it can be a model-agnostic approach that may be broadly generalizable to other LLMs[19]. The superior empathy of Claude V2 with CoT prompting compared to other chatbots and physicians is promising evidence in support of prompt engineering techniques to optimize chatbot outputs with limited technical expertise required. Application of prompt engineering techniques have seen state-of-the-art success in encoding clinical knowledge[20], motivating the design of structured prompts that encode human psychosocial cues to facilitate chatbot responses to patient emotion in messaging.

This study was limited to static, single-time point interactions, restricting insights into longitudinal or real-time dynamics of patient-chatbot interactions. Future research can employ longitudinal designs, real-time conversational analysis, and established physiological or psychometric assessments to systematically explore LLM empathy and its clinical implications in medicine. Translating beyond research benchmarks towards clinician adoption of chatbots may require clinician education about how to design prompts to steer LLM responses, each with their own capabilities and limitations[21,22].

While AI systems have demonstrated the ability to generate responses perceived as empathetic[23,24] —potentially enhancing patient engagement and alleviating clinician workload— their deployment raises critical concerns. These include safeguarding patient privacy, ensuring informed consent, establishing oversight and liability for AI-generated outputs, mitigating biases to promote health equity, and amending changes to clinical workflows involving AI that may impact the patient-provider relationship[25]. The growing popularity of AI tools that mimic humanistic traits like empathy raises concerns about misinformation and misidentification as an authoritative expert or empathetic peer that should be considered in future real-world assessments[26]. Addressing these issues is essential for the responsible integration of AI technologies in oncology, ensuring they augment rather than diminish the empathetic communication central to effective patient care.

The primary limitations of this study include the use of isolated interactions at a single time point on an online forum to model physician-patient interactions, chatbot responses were longer than physician responses on average despite instruction to limit word count response which may confound the length of response and perceived empathy, and the biased demographic representation of survey participants who were primarily white, male, well educated, and high income that reflect the patient population available at our recruitment site. Using Reddit-derived data as a proxy for real-world oncology consultations since online anonymity may alter the patient–physician relationship by removing nonverbal cues, reducing trust and accountability, and encouraging different self-presentation compared to real-world oncology consultations.

Participant subgroup characteristics, including having former or current diagnosis of cancer, were not collected since this pilot study aimed to capture a broad range of perspectives from individuals with cancer experience. Given evidence that biological and social contexts may contribute to differences in empathy between genders[27] among other sociocultural factors, the biased demographics of the study sample may have limited the ability to generalize our results of empathic perception to non-represented populations, motivating future investigations into the association between subgroup characteristics such as demographic factors and perceived empathy of chatbot-patient interactions. This study exclusively assessed the Claude LLM, which may limit the generalizability of our findings to other LLMs. There may be differences between perceived empathy in written responses compared to real-world clinical settings that include non-verbal cues of empathy and constraints based on time and clinical workload.

Participant subgroup characteristics, including having former or current diagnosis or cancer, were not collected since this pilot study aimed to capture a broad range of perspectives from individuals with cancer experience, motivating future study evaluations of the association between participant subgroups and perceptions of chatbot empathy.

Methods

Dataset

In this prospective, cross-sectional study, we surveyed oncology patients at a tertiary cancer center about the empathy of physician and chatbot responses to patient questions related to cancer. Two authors (D.C, R.P.) reviewed the external database[3] in duplicate and included randomly sampled patient questions if they mentioned a cancer diagnosis and contained clinical context, such as symptom descriptions and diagnostic details, that may be typical of oncological concerns observed in clinical settings ($n = 100$). Patient questions were originally posted on Reddit's r/AskDocs from January 1, 2018 to May 31, 2023. The survey collected participant demographic information and empathy ratings between May 1, 2024 and October 31, 2024. This study followed the STROBE reporting guidelines. This study was approved by the UHN Research Ethics Board.

Procedure

We generated responses from 3 AI chatbots (Claude V1, Claude V2, and Claude V2 with chain of thought prompt engineering) to each patient question with an additional prompt to limit the word length of the chatbot response to the mean physician response word count (mean, 100). Claude V1 and Claude V2 were prompted using the CoT-1 prompt while Claude V2 with CoT was prompted with CoT-1, CoT-2, and CoT-3 prompts in succession as described in Supplementary Table 1.

Survey

Patient participants were included if they were above 18 years of age, could read and understand English, and had a former or current diagnosis of cancer. Eligible participants were approached in clinics, consented to the study, and completed the survey on REDCap, a digital survey platform. We originally planned a sample size of 200 participants, but after 45 participants, an interim analysis demonstrated a significant and meaningful effect size in empathy ratings between physician and chatbot responses that prompted early study termination. Each participant was randomly assigned 10 of the 100 patient questions in the database. For each question, participants rated perceived empathy for three chatbot responses and one physician response, presented in a random order and blinded. For a random sample of 10 out 100 total patient questions about cancer, participants rated the perceived empathy of three chatbot and one physician response to each patient question (Supplementary Note 1). Empathy was scored on a Likert scale,

ranging from 1 to 5 (1: very poor, 2: poor, 3: acceptable, 4: good, and 5: very good). Participant empathy scores for each response were averaged to determine a consensus participant empathy score. Physician empathy scores rated in triplicate for each response were sourced from Chen et al.[3] and averaged to determine a consensus physician empathy score.

Statistical analysis

Using the Wilcoxon test with Benjamini-Hochberg correction, we compared the participant-rated mean empathy scores for responses generated by chatbots and physicians. We also compared our participant-rated mean empathy scores with published physician-rated mean empathy scores for the same set of physician and Claude V1 responses[3]. Physician-rated empathy scores of Claude V2 were not analyzed because Claude V2 was not released at the time of publication of this external dataset[3]. Spearman correlation and Ordinary Least Squares (OLS) regression was used to measure the association between word count and Flesch Kincaid Reading Grade Level (FKRGL), a measure of readability, with participant-rated empathy. Cohen's d was used to measure effect size differences between groups. Statistical analyses were conducted with Python 3.8.9 and scipy 1.11.3.

Conclusion

This study found that oncology patients, like oncology physicians, perceive chatbot responses as more empathetic than physician responses to patient questions about cancer. However, patients may prioritize different elements of clinical messages in their evaluation of empathy than physicians. Further research is required to optimize the integration of empathy in clinical messaging and evaluate the implementation and scope of patient-facing chatbots.

Data availability

All data generated in this study is available at the online repository: https://github.com/davidchen0420/Empathy-Chatbot-Project-Data-and-Code.

Code availability

All code applied in this study is available at the online repository: https://github.com/davidchen0420/Empathy-Chatbot-Project-Data-and-Code

Received: 6 December 2024; Accepted: 24 April 2025;
Published online: 13 May 2025

References

1. Thirunavukarasu, A. J. et al. Large language models in medicine. *Nat. Med.* **29**, 1930–1940 (2023).
2. Ayers, J. W. et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern. Med.* **183**, 589 (2023).
3. Chen, D. et al. Physician and artificial intelligence chatbot responses to cancer questions from social media. *JAMA Oncol.* **10**, 956 (2024).
4. Chen, D. et al. Performance of multimodal artificial intelligence chatbots evaluated on clinical oncology cases. *JAMA Netw. Open* **7**, e2437711 (2024).
5. Liu, B. & Sundar, S. S. Should machines express sympathy and empathy? Experiments with a health advice chatbot. *Cyberpsychol. Behav. Soc. Netw.* **21**, 625–636 (2018).
6. Lelorain, S., Gehenne, L., Christophe, V. & Duprez, C. The association of physician empathy with cancer patient outcomes: a meta-analysis. *Psychooncology* **32**, 506–515 (2023).
7. Cadiente, A., Chen, J. & Pilkington, B. Machine-Made Empathy? Why medicine still needs humans. *JAMA Intern. Med.* **183**, 1278 (2023).
8. Gao, P., Han, D., Zhou, R., Zhang, X. & Wang, Z. CAB: empathetic dialogue generation with cognition, affection and behavior. Preprint at https://doi.org/10.48550/arXiv.2302.01935 (2023).
9. Hamad, O., Hamdi, A. & Shaban, K. ASEM: enhancing empathy in chatbot through attention-based sentiment and emotion modeling. Preprint at https://doi.org/10.48550/ARXIV.2402.16194 (2024).
10. Zandie, R. & Mahoor, M. H. EmpTransfo: a multi-head transformer architecture for creating empathetic dialog systems. Preprint at https://doi.org/10.48550/arXiv.2003.02958 (2020).
11. Zaranis, E., Paraskevopoulos, G., Katsamanis, A. & Potamianos, A. EmpBot: a T5-based empathetic chatbot focusing on sentiments. Preprint at https://doi.org/10.48550/arXiv.2111.00310 (2021).
12. Zhou, H., Huang, M., Zhang, T., Zhu, X. & Liu, B. Emotional chatting machine: emotional conversation generation with internal and external memory. Preprint at https://doi.org/10.48550/arXiv.1704.01074 (2018).
13. Kaplan, J. et al. Scaling laws for neural language models. Preprint at https://doi.org/10.48550/arXiv.2001.08361 (2020).
14. Zhao, W. et al. Is ChatGPT equipped with emotional dialogue capabilities? Preprint at https://doi.org/10.48550/arXiv.2304.09582 (2023).
15. Lee, Y. K., Lee, I., Shin, M., Bae, S. & Hahn, S. Chain of empathy: enhancing empathetic response of large language models based on psychotherapy models. *Korean J. Cognit. Sci.* **35**, 23–48 (2024).
16. Bender, E. M. & Koller, A. Climbing towards NLU: on meaning, form, and understanding in the age of data. In *Proc. 58th Annual Meeting of the Association for Computational Linguistics* 5185–5198 (Association for Computational Linguistics, Online,). https://doi.org/10.18653/v1/2020.acl-main.463. (2020).
17. Abdulkader, R. S. et al. The intricate relationship between client perceptions of physician empathy and physician self-assessment: lessons for reforming clinical practice. *J. Patient Exp.* **9**, 23743735221077537 (2022).
18. Tanco, K. et al. Patient perception of physician compassion after a more optimistic vs a less optimistic message: a randomized clinical trial. *JAMA Oncol.* **1**, 176 (2015).
19. Wei, J. et al. Chain-of-thought prompting elicits reasoning in large language models. Preprint at https://doi.org/10.48550/arXiv.2201.11903 (2023).
20. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
21. Meskó, B. Prompt engineering as an important emerging skill for medical professionals: tutorial. *J. Med. Internet Res.* **25**, e50638 (2023).
22. Chen, D. & Gorla, J. The need to develop digital health competencies for medical learners. *Med. Teach.* **45**, 790–791 (2023).
23. Chen, Z., Zhao, X., Hua, M. & Xu, J. Building bonds through bytes: the impact of communication styles on patient-chatbot relationships and treatment adherence in ai-driven healthcare. HCI International 2024 – Late Breaking Papers (ed. Duffy, V. G.) vol. 15376 32–52 (Springer Nature Switzerland, 2025).
24. Rokhshad, R. et al. Efficacy and empathy of AI chatbots in answering frequently asked questions on oral oncology. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol.* S2212440325000021 https://doi.org/10.1016/j.oooo.2024.12.028 (2025).
25. Kann, B. H., Hosny, A. & Aerts, H. J. W. L. Artificial intelligence for clinical oncology. *Cancer Cell* **39**, 916–927 (2021).
26. Lawson McLean, A. & Hristidis, V. Evidence-based analysis of AI chatbots in oncology patient education: implications for trust, perceived realness, and misinformation management. *J. Cancer Educ*. https://doi.org/10.1007/s13187-025-02592-4 (2025).
27. Christov-Moore, L. et al. Empathy: gender effects in brain and behavior. *Neurosci. Biobehav. Rev.* **46**, 604–627 (2014).

Acknowledgements

This work was partially supported by the CARO CROF studentship and the Robert L. Tundermann and Christine E. Couturier philanthropic funds.

Author contributions

D.C. and S.R. conceived the study. D.C., K.C., and R.P. designed the study and collected data. D.C. and Z.A.L. conducted data analyses. D.C. drafted the manuscript. D.C., K.C., R.P., Z.A.L., F.L., E.M., L.E., B.F.H., J.C., A.H.,

N.F., P.W., and S.R. revised, read, and approved the manuscript. S.R. supervised the study.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-01671-6.

Correspondence and requests for materials should be addressed to Srinivas Raman.