The application of multimodal large language models in medicine

Jianing Qiu,[a] *Wu Yuan,*[a] *and Kyle Lam*[b,*]

[a]Department of Biomedical Engineering, The Chinese University of Hong Kong, Hong Kong SAR, China
[b]Department of Surgery and Cancer, St Mary's Hospital, Imperial College London, United Kingdom

The Lancet Regional Health - Western Pacific 2024;45: 101048

Published Online xxx
https://doi.org/10.1016/j.lanwpc.2024.101048

In September 2023, OpenAI released GPT-4V,[1] a multimodal foundation model[2,3] connecting large language models (LLMs) with vision input. Foundation models, defined as large AI models which are trained upon vast datasets that can be later adapted to a range of downstream tasks,[4] represent the latest wave in AI research. Differing from specific AI models which are trained for a single function, foundation models are designed to be multi-purpose. The most widely known of these are the GPT models, powering ChatGPT, which previously could input language alone.

Language represents only a proportion of the data encountered within healthcare and previously this limitation of 'unimodal' AI including LLMs has meant that vital data sources such as radiology, endoscopic images and laboratory investigations have not been incorporated. However, OpenAI's latest offering, GPT-4V, allows image input; in conjunction with Whisper,[5] an automatic speech recognition system, and text-to-speech generation techniques, this now means that ChatGPT can see, hear, and speak.[6] ChatGPT's leap into multimodality therefore opens new horizons for clinical work processes and applications.

Here, we highlight four example areas where multimodal LLMs can benefit clinicians in an example scenario of a patient presenting with small bowel obstruction (Fig. 1) and across varying specialties (Supplement).

Multimodal LLMs empower LLMs further through seamless transcription and summarisation of speech data, allowing generation of clinical records or letters directly from the doctor-patient consult (Fig. 1); this could reduce the burden of clinical documentation significantly. Secondly, multimodal LLMs build upon existing AI image interpretation through their ability to integrate existing information including the patient's history, indications for imaging, and comparisons with previous imaging, and by offering recommendations (Fig. 1). They can reduce the need for large datasets through their few-shot or zero-shot learning abilities (completing a task through limited or no training examples respectively) and support visual prompting to refine the prediction (for example, a user can manually indicate the region of interest within an image).[7,8] Thirdly, optical character recognition empowers multimodal LLMs to detect numerical and text (irrespective of the language used) from image input (Fig. 1). Finally, capabilities in video understanding could allow automatic documentation of procedural notes, improving efficiency and accuracy of documentation. Scene understanding–for example identification of anatomical landmarks–could open doors to procedure assistance, augment clinician capabilities, and ultimately lead to improved clinical outcomes.

These newfound capabilities of multimodal LLMs also pose newfound challenges for their adoption within healthcare. Hallucinations, where the model outputs incorrect or nonsensical information, are a fundamental issue. In our example, the multimodal ChatGPT outputs an incorrect interpretation of an ECG which is convincing at face-value (Fig. 1). Exploratory studies[1,8] have also shown that GPT-4V can hallucinate while responding to vision-based queries, secondary to either incorrect reasoning from the underlying LLM or incorrect recognition of visual content. Input of increasing numbers of clinical data types is a concern as broader expertise will be required to determine ground truth resulting in greater challenges in identifying the source of hallucinations. Reliability must therefore be improved in order to meet the high threshold required for translation into clinical practice.

Secondly, increasing data modalities will lead to greater privacy concerns. The growing size of foundation models threatens the accidental exposure of patient data from the data used within its training process. Data modalities such as speech and video threaten not only the privacy of patients but also clinicians themselves.

Finally, regulation of multimodal LLMs presents a significant challenge. While task specific AI requires validation only for the task it is designed for, the emergent intelligence of foundation models (where future capabilities of the model are still to be discovered) demands a rethink for regulators in the approach taken both to test models and mitigate against AI failure. It is likely that foundation models will not fit neatly into existing regulation and require novel custom solutions. This should be a key priority for translation as innovation is likely to outpace regulation. One potential

*Corresponding author.
E-mail address: k.lam@imperial.ac.uk (K. Lam).

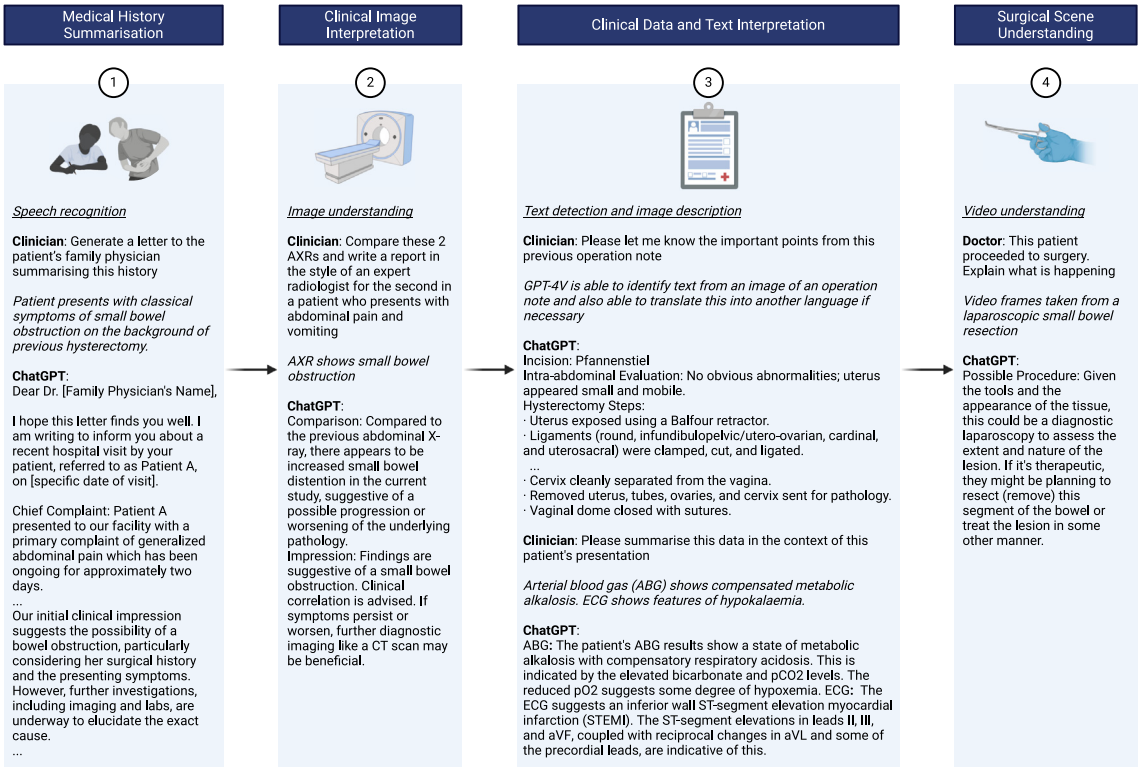

Fig. 1: The multimodal ChatGPT can be capitalised upon at all stages of the patient pathway. This preliminary proof-of-concept example demonstrates how multiple data types can be inputted to ChatGPT to assist clinicians from presentation to treatment in a patient with small bowel obstruction. Outputs are shortened for the purpose of presentation of this figure.

approach is to adapt existing regulation for anticipated downstream applications from the foundation model and monitor for emerging functions, risks, and failures. However, multimodal LLMs are trained on huge amounts of data taken from the internet, and it is difficult for users to know what data are being used to train them. This calls into question the traditional approach of validating models upon public benchmarks as the data used to train the multimodal LLMs may have included these benchmarks. Thus, regulators need to establish isolated validation datasets which are inaccessible to model developers, and conduct independent examinations using such datasets to ensure trustworthy and objective validation.

Despite these challenges, multimodal AI powered by foundation models offers significant promise in augmenting the medical workforce in clinical decision-making and management. The release of GPT-4V will spark future endeavours in the responsible development, use, and regulation of multimodal medical AI, and the improvement of AI trustworthiness and accessibility in medicine.

Declaration of interests
The authors declare no competing interests.

Acknowledgements
Funding: Funding and infrastructural support was provided by the NIHR Imperial Biomedical Research Centre. Kyle Lam is supported by a NIHR Academic Clinical Fellowship.

Appendix A. Supplementary data
Supplementary data related to this article can be found at https://doi.org/10.1016/j.lanwpc.2024.101048.

References
1 OpenAI. *GPT-4V(ision) system card*; 2023. https://cdn.openai.com/papers/GPTV_System_Card.pdf. Accessed October 18, 2023.
2 Qiu J, Li L, Sun J, et al. Large AI models in health informatics: applications, challenges, and the future. *IEEE J Biomed Health Inform*. 2023:1–14.
3 Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence. *Nature*. 2023;616(7956):259–265.
4 Bommasani R, Hudson DA, Adeli E, et al. On the opportunities and risks of foundation models. *arXiv*. 2021. https://doi.org/10.48550/arXiv.2108.07258.
5 Radford A, Kim JW, Xu T, Brockman G, Mcleavey C, Sutskever I. Robust speech recognition via large-scale weak supervision. In: Andreas K, Emma B, Kyunghyun C, Barbara E, Sivan S, Jonathan S, eds. *Proceedings of the 40th international conference on machine learning. Proceedings of machine learning Research*. PMLR; 2023:28492–28518.
6 OpenAI. *ChatGPT can now see, hear, and speak*; 2023. https://openai.com/blog/chatgpt-can-now-see-hear-and-speak.
7 Kirillov A, Mintun E, Ravi N, et al. Segment anything. *arXiv*. 2023. https://doi.org/10.48550/arXiv.2304.02643.
8 Yang Z, Li L, Lin K, et al. The dawn of LMMs: preliminary explorations with GPT-4V(ision)2023. https://ui.adsabs.harvard.edu/abs/2023arXiv230917421Y. Accessed October 18, 2023.