Clinical Imaging 132 (2026) 110741

Contents lists available at ScienceDirect

Clinical Imaging

journal homepage: www.elsevier.com/locate/clinimag

Letter to the Editor

Comment on "Classifying the clinical significance of common breast pain symptoms using a large language model, ChatGPT (GPT-4)"

Dear Editor,

We commend the authors for exploring the use of a large language model (GPT-4) to triage the clinical significance of breast pain descriptions and for demonstrating encouraging sensitivity using a zero-shot prompting strategy.[1] The study provides a timely and valuable contribution to the emerging literature on LLM-assisted clinical triage. We would, however, like to raise a methodological concern that extends beyond model performance and touches on the alignment between model reasoning and clinical decision-making.

In the current study, GPT-4 is tasked with mapping free-text symptom descriptions directly to a binary clinical recommendation. While this end-to-end formulation is appealing from an engineering perspective, it departs fundamentally from how clinicians reason about breast pain. Clinical assessment typically proceeds through the identification and synthesis of intermediate symptom attributes, such as focality, and associated red-flag features, before arriving at a risk-stratified decision.[2] By bypassing these intermediate representations, the model is required to implicitly infer clinically salient attributes that are neither explicitly represented nor verifiable.

This design choice has important implications. As the authors note, a substantial proportion of misclassifications occur in cases where key attributes are absent or ambiguously described. More importantly, even when predictions are correct, the absence of explicit feature extraction makes it impossible to determine whether the model's decision is grounded in clinically valid cues or in spurious correlations. In a medical context, correct outcomes produced for the wrong reasons remain problematic, particularly for downstream auditing, error analysis, and safety assurance.

From a clinical decision-support standpoint, we believe that a structured, two-stage framework, where the model first extracts key symptom attributes in a standardized format, followed by a downstream rule-based or model-based decision, offers advantages that extend beyond interpretability alone. Such a framework would better mirror established diagnostic reasoning, enable targeted evaluation of extraction versus decision failures, and, critically, support uncertainty-aware deployment by allowing selective human review when extracted attributes are incomplete or uncertain.

We therefore encourage the authors to consider a comparative experiment evaluating their current zero-shot classification approach against a structured-extraction-first pipeline. Reporting both the accuracy and reliability of extracted clinical attributes, alongside downstream classification performance, would help clarify whether observed errors arise primarily from representational ambiguity or from decision thresholds. We believe this analysis would substantially strengthen the study's methodological contribution and enhance its relevance for real-world clinical implementation. Furthermore, it is recommended that the study be extended to other diseases, such as *Clonorchis sinensis*[3] infection, while also exploring the integration of encryption techniques to enhance the confidentiality of the model.[4] Such expansion would contribute to validating the generalizability of the proposed framework and its potential applicability across diverse clinical contexts.

CRediT authorship contribution statement

Yihan Hu: Writing – original draft.

Ethical approval

Not required.

Research registry number

Not applicable.

Clinical trial registration details/number

Not applicable, as this study does not report a clinical trial.

Human ethics and consent to participate declarations

Not applicable as no patient data were collected or analyzed in this Study.

Declaration of Generative AI and AI-assisted technologies in the writing process

Generative AI tools, including Paperpal and ChatGPT-4o, were used exclusively for language polishing, grammar correction, and stylistic refinement. These tools played no role in the conceptualization, data analysis, interpretation of results, or substantive content development of this manuscript. All intellectual contributions, data analysis, and scientific interpretations are solely attributed to the authors. The final content was rigorously reviewed and edited to ensure accuracy and originality. The authors assume full responsibility for the accuracy, originality, and integrity of the work presented.

Declaration of funding

No funding was received for this study.

https://doi.org/10.1016/j.clinimag.2026.110741
Received 4 January 2026; Accepted 6 January 2026
Available online 9 February 2026

Declaration of competing interest

The authors declare no conflict of interests relevant to this study.

Data availability

Not applicable, as no data were generated or analyzed in this study.

References

1 Haver H, Bahl M, Chung M. Classifying the clinical significance of common breast pain symptoms using a large language model, ChatGPT (GPT-4). Clin Imaging 2025;125.
2 Peskine Y, Korenčić D, Grubisic I, et al. Definitions matter: guiding GPT for multi-label classification[C]//EMNLP 2023. In: Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics; 2023.
3 Ma RH, Luo XB, Liu YP, et al. *Clonorchis sinensis* infection are associated with calcium phosphate gallbladder stones: a single-center retrospective observational study in China. Medicine (Baltimore) 2025;104(46):e45739.
4 Wu X, Zhang Y, Shi M, et al. An adaptive federated learning scheme with differential privacy preserving. Future Gener Comput Syst 2022;127:362–72.

Yihan Hu[a]
[a] *MRC Epidemiology Unit, University of Cambridge, Cambridge, CB2 0SL, UK*
E-mail address: yh623@cam.ac.uk.