Benchmarking Large Language Models for Clinical Decision Support in Breast Cancer Care: A Multi-Institutional Expert Evaluation

Zunairah Shah, Safa Afridi, Mulham Ombada, Arya Mariam Roy, Alexis LeVee, Sara Premji, Varsha Gupta, Nerea Lopetegui, Dionisia Quiroga, Ruth. L. Sacks, Saba Shaikh, Yara Abdou, Han Yu, Anant Madabhushi, Ravi Parikh, Lubna. N. Chaudhary, Ellis Levine, Matteo Lambertini, Kevin Kalinsky, Sheheryar Kabraji, Shipra Gandhi

ROSWELL PARK COMPREHENSIVE CANCER CENTER

BACKGROUND

- AI and **large language models (LLMs)** are emerging tools for clinical decision support in oncology.
- Validation in complex **breast cancer scenarios (BCCS)** remains limited.
- Objective: To benchmark three LLMs for clinical accuracy, usability, and readiness in real-world breast oncology practice.

METHODS

AI Models
1.ChatGPT-4o (GPT)
2. DeepSeek-R1 (DS)
3.OpenEvidence (OE)

Reviewers
16 Breast medical oncologists from 9 academic centers

Questions
Q1–9: treatment decisions
Q10: image interpretation

Evaluation
5-point Likert scale
(1=poor, 5=excellent) for accuracy, clarity, relevance, usability.

Analysis
Repeated measures ANOVA + Tukey's post hoc comparisons

RESULTS

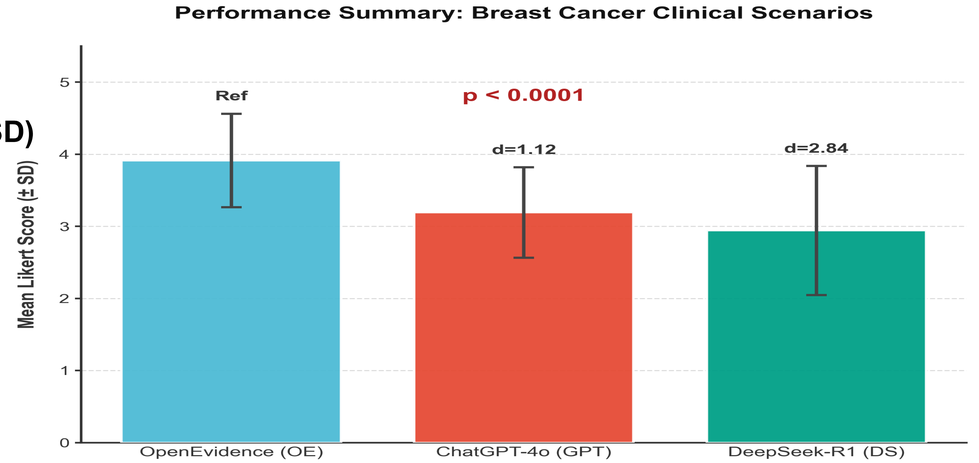
OE achieved significantly higher accuracy than GPT and DS across Q1–9 ($p < 0.0001$)
GPT-4o outperformed DS for multimodal image interpretation (Q10)

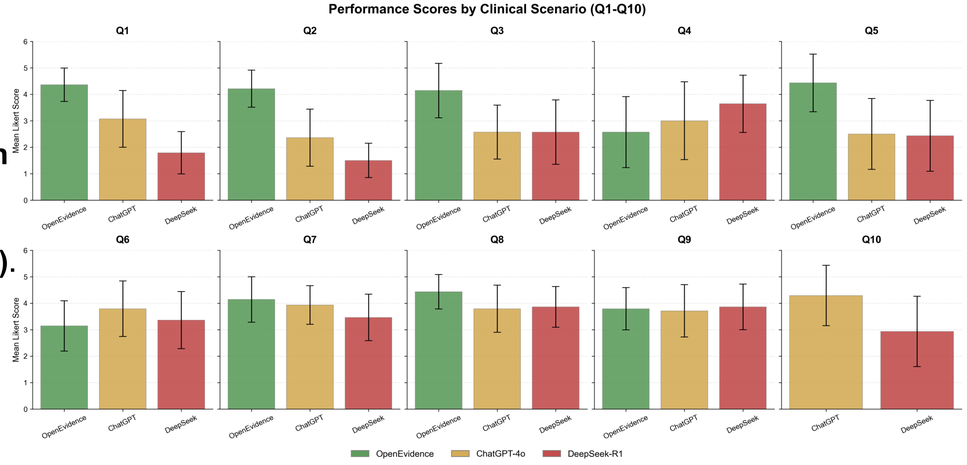
Breast cancer clinical scenarios (Q1–9)
Mean Likert score (± SD)

- **OE**: 3.91 ± 0.48
- **GPT**: 3.19 ± 0.67
- **DS**: 2.93 ± 0.54



Image used for multimodal image-interpretation evaluation in the breast cancer clinical scenario (Q10)
Mean Likert score (± SD).
GPT (vision-enabled)
4.29 ± 1.14
DS (OCR-based)
2.93 ± 1.33
OE: No MMI support





CONCLUSIONS

- First **head-to-head** evaluation of **GPT-4o, DS, and OE in BCCS.**
- OE produced the most guideline-concordant and clinically accurate recommendations.
- GPT-4o showed promise for multimodal image integration.
- DS lagged in context comprehension and accuracy.

IMPLICATIONS

LLMs hold potential to **augment** rather than replace clinical judgment.

These findings **highlight** the promise of **LLMs in breast oncology** and the need for further refinement to ensure reliability and real-world applicability.

CONTACTS

Zunairah Shah, MD
𝕏 @zunairahshah
Zunairah.Shah@RoswellPark.org
Shipra Gandhi, MD
𝕏 @Gandhi_Shipra
Shipra_Gandhi@Emory.edu