Retrieval-Augmented GPT-4 Improves NCCN-Concordant Breast Cancer Treatment Recommendations

Bradley Callas[1], Connor Yost[2], Sarah Osborn[3], Sarah Monick[4], Natalie Ertz[5], Ryan Nguyen[6], Arturo Loaiza[7]

1. Creighton University School of Medicine-Arizona, 2. Creighton University Arizona, Department of Internal Medicine, 3. St. George's University School of Medicine, 4. Mayo Clinic - Arizona Department of Oncology Hematology, 5. Honor Health Department of Oncology Hematology, 6. University of Illinois, Department of Oncology Hematology, 7. St. Luke's Hospital Philadelphia Department of Oncology Hematology

Creighton UNIVERSITY

Abstract

Large language models (LLMs), such as ChatGPT, offer a potential tool to support clinical decision-making. However, hallucinations and deviation from evidence-based guidelines remain significant concerns. This study evaluated whether retrieval-augmented generation (RAG) using National Comprehensive Cancer Network (NCCN) guidelines improves performance in treatment recommendation tasks. Fifty clinical vignettes were created from published breast cancer case reports. Three models—a baseline GPT-4, GPT-5 RAG, and OpenEvidence—were prompted. Our results show that the GPT-5 RAG model improved readability, reliability and total composite score when compared to GPT-4 and OpenEvidence. The reduced incidence of severe hallucinations, increase in guideline-concordant recommendations, and enhanced readability findings support further development of guideline-anchored LLMs for clinical use.

BACKGROUND

The swift development of biomarker-driven breast cancer therapies has increasingly complicated the process of guideline-concordant treatment selection. Large language models (LLMs), such as ChatGPT and OpenEvidence, are potential tools to support clinician efforts to utilize the most recent recommendations and guidelines. Despite this potential, LLMs face significant concerns about accuracy, privacy, and regulatory compliance. This study aimed to evaluate whether retrieval-augmented generation (RAG) using National Comprehensive Cancer Network (NCCN) guidelines improves responses to treatment recommendation tasks.

METHODS

Design & Models

Output 1: Baseline GPT-4 with open retrieval

Output 2: GPT-5 RAG model restricted to NCCN breast cancer guidelines

Output 3: Open Evidence (NEJM/JAMA-anchored medical LLM)

Cases

50 clinical vignettes were created from published case reports to reflect diverse scenarios, including HER2-low, triple-negative, and BRCA-mutated disease across all stages.

Prompting

Structured template for: treatment plan, rationale, needed work-up/biomarkers, and trial/clinical-study options

Scoring

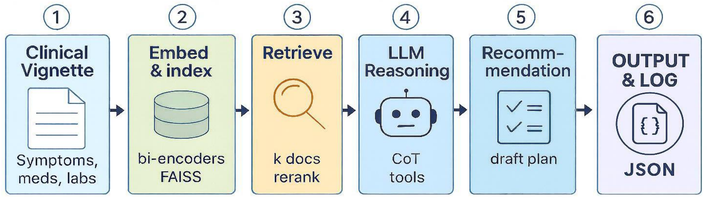
Guideline Concordance (0.00–1.00 mGPS component): alignment with current NCCN CNS recommendations for disease, line, and key qualifiers (e.g., BRCA, HER2, PR, age)

Hallucination Severity (0.00 to –1.00 penalty): factual accuracy of drugs, doses, trials, and guideline statements

Readability & Rationale (1–5 Likert): clarity, coherence, and explicit linking of recommendations to evidence

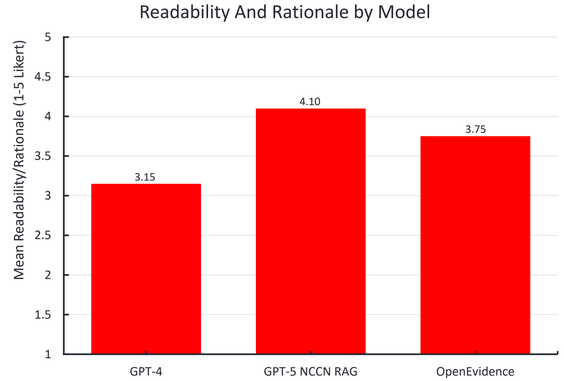
Total Composite Score: Defined as guideline concordance score plus hallucination severity score

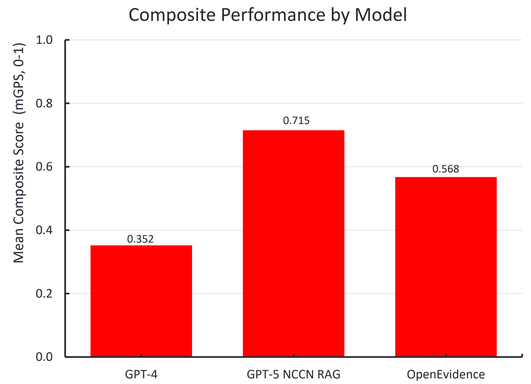
Process of Retrieval-Augmented Generation (RAG) for Clinical Decision Support



RESULTS

- Comparative analysis was performed among three large-language-model outputs – baseline GPT-4, GPT-5 RAG anchored to NCCN guidelines, and OpenEvidence – across 50 clinical vignettes of diverse breast cancer cases.
- GPT-5 RAG significantly improved overall performance compared to GPT-4 and OpenEvidence, producing both higher composite mGPS scores as well as improved readability and rationale.
- Composite scores increased dramatically when comparing GPT-5 RAG and GPT-4 (mean difference = 0.363; $t = 8.2$; $p < 0.0001$), indicating markedly fewer hallucinations and stronger concordance with NCCN guidelines.
- Comparing composite scores between GPT-5 RAG and OpenEvidence showed similar results (mean difference = 0.148; $t = 3.0$; $p = 0.004$), further demonstrating the increased reliability of the GPT-5 RAG model.
- Readability and rationale also improved when using GPT-5 RAG compared to GPT-4 (mean difference = 0.950; $t = 12.7$; $p < 0.0001$) as well as when compared to Open Evidence (mean difference = 0.350; $t = 3.6$; $p = 0.0006$), confirming that vector-based retrieval improved accuracy and interpretability.
- OpenEvidence also showed improvement over GPT-4 in both readability and rationale (mean difference = 0.6; $t = 7$; $p < 0.0001$) and composite score (mean difference = 0.215; $t = 3.9$; $p = 0.0002$).





DISCUSSION

- A GPT-5 RAG model constrained to breast cancer NCCN guidelines substantially improves total composite score over baseline GPT-4 performance demonstrating fewer hallucinations and increased consistency with NCCN guidelines for the treatment of breast malignancies.
- The GPT-5 RAG model also showed statistically significant improvement in readability and rationale over the GPT-4 model, supporting the benefit of using augmented large language models.
- Compared with OpenEvidence, the GPT-5 RAG model exhibited a higher composite score and enhanced readability and rationale, further indicating the improved guideline concordance, decreased hallucinations, superior reliability of the NCCN-RAG model.
- Some limitations of this study include a smaller sample size and the use of published case reports as clinical vignettes; real-world validation and further fine-tuning remain necessary.
- While further refinement is needed, the reduced incidence of severe hallucinations, increase in guideline-concordant recommendations, and enhanced readability findings support further development and prospective validation of guideline-anchored LLMs for clinical use.

REFERENCES

1 Baclig NV, McCann KE. Updates in the treatment of HR+HER2- breast cancer. *Curr Opin Obstet Gynecol*. 2024;36(1):57-63. doi:10.1097/GCO.0000000000000925

2 Wong EYT, Verlingue L, Aldea M, et al. ESMO guidance on the use of Large Language Models in Clinical Practice (ELCAP). Ann Oncol. Published online December 31, 2025. doi:10.1016/j.annonc.2025.09.001

3 Liang EN, Pei S, Staibano P, van der Woerd B. Clinical applications of large language models in medicine and surgery: A scoping review. *J Int Med Res*. 2025;53(7):3000605251347556. doi:10.1177/03000605251347556

4 Benary M, Wang XD, Schmidt M, et al. Leveraging Large Language Models for Decision Support in Personalized Oncology. JAMA Netw Open. 2023;6(11):e2343689. doi:10.1001/jamanetworkopen.2023.43689.

5 Park YJ, Pillai A, Deng J, et al. Assessing the research landscape and clinical utility of large language models: a scoping review. *BMC Med Inform Decis Mak*. 2024;24(1):72. Published 2024 Mar 12. doi:10.1186/s12911-024-02459-6