Research Letter

# Performance of Retrieval-Augmented Large Language Models to Recommend Head and Neck Cancer Clinical Trials

Tony K W Hung[1], MD, MBA, MSCR; Gilad J Kuperman[1], MD, PhD; Eric J Sherman[1], MD; Alan L Ho[1], MD, PhD; Chunhua Weng[2], PhD; David G Pfister[1], MD; Jun J Mao[1], MD, MSCE

[1]Memorial Sloan Kettering Cancer Center, New York, NY, United States

[2]Columbia University, Department of Biomedical Informatics, New York, NY, United States

**Corresponding Author:**
Tony K W Hung, MD, MBA, MSCR
Memorial Sloan Kettering Cancer Center
530 E 74th St
New York, NY, 10021
United States
Phone: 1 646 608 4127
Fax: 1 646 888 4269
Email: hungt@mskcc.org

***(J Med Internet Res 2024;26:e60695)*** doi: 10.2196/60695

**KEYWORDS**

large language model; LLM; ChatGPT; GPT-4; artificial intelligence; AI; clinical trials; decision support; LookUpTrials; cancer care delivery; head and neck oncology; head and neck cancer; retrieval augmented generation

## Introduction

Chatbots based on large language models (LLMs) have demonstrated the ability to answer oncology examination questions with impressive accuracy without specialized training or reinforcement [1,2]; however, leveraging LLMs in oncology decision support has not yet demonstrated suitable performance, as LLMs would produce responses that deviate from cancer expert recommendations and guidelines [3-5]. Furthermore, the rapidly changing oncology landscape, including knowledge of cancer clinical trials, limits the meaningful use of LLMs in practice given delays in training dataset updates. To enhance LLM utility in oncology practice, we developed a retrieval-augmented LLM, powered by GPT-4, and evaluated its performance to provide appropriate clinical trial recommendations for a head and neck (HN) cancer population.

## Methods

On February 1, 2022, we piloted a clinical trial knowledge management application, LookUpTrials, at the Memorial Sloan Kettering Cancer Center (MSK) [6]. Using LookUpTrials' real-time database, we applied retrieval-augmented generation architecture and direct preference optimization to fine-tune GPT-4 as a clinical trial decision assistant [7]. Specifically, we enabled retrieval-augmented GPT-4 to respond with up-to-date information—such as trial availability—developed initial prompts, and validated GPT-4 responses from 1120 preference pairs across 56 MSK HN clinical trials. Preference pairs were constructed in [trial : attributes] format, including 20 organizational, investigator, and study attribute types (Multimedia Appendix 1). Data labels were annotated by author TKWH and cross-verified by 2 trial managers. From November 7, 2023, to January 30, 2024, we collected all consecutive new patient cases and their respective clinical trial recommendations, which were made by consensus during a weekly HN conference attended by 5-8 oncologists with 2 to more than 25 years of practice experience. Cases were categorized by diagnosis, biomarkers, cancer stage, treatment setting, and physician recommendations on clinical trials. Using these cases as test datasets, we prompted retrieval-augmented GPT-4 using a semistructured template, as follows: "Given patient with a <biomarkers>, <diagnosis>, <cancer stage>, <treatment setting>, what are possible clinical trials?" (eg, given a patient with human papillomavirus–associated HN cancer, metastatic stage, in a first-line treatment setting, what are the possible clinical trials?). GPT-4 responses were compared with physician recommendations, with concordance defined a priori: a GPT-4 response was a true positive if it included the recommended clinical trial(s); a true negative if neither the GPT-4 response nor the physicians recommended any clinical trial(s); a false positive if the GPT-4 response recommended clinical trial(s) but physicians did not; and a false negative if the GPT-4 response did not recommend clinical trial(s) but the physicians did. We analyzed the performance of GPT-4 based on its response precision (positive predictive value), recall (sensitivity), and $F_1$-score (harmonic mean of precision and recall). We further analyzed subgroup performance by cancer

types and the presence of biomarkers. Statistical analyses were performed using JMP-17.2.0.

### Ethical Considerations

MSK institutional review board approved the study (application number: 24-120).

## Results

We analyzed 178 patient cases (mean age 66, SD 13.9 years), primarily male (n=134, 75.3%), with local/locally advanced cancers (n=121, 68.0%), including HN (n=109, 61.2%), thyroid (n=29, 16.3%), skin (n=16, 9.0%), or salivary gland (n=14, 7.9%) cancers (Table 1). Over one-third of cases had biomarkers (n=66, 37.1%). The majority were treated in the definitive setting with combined modality therapy (n=75, 42.1%), and a modest proportion were treated under clinical trials (n=18, 10.1%). Overall, retrieval-augmented GPT-4 achieved moderate performance (Table 2), matching physician clinical trial recommendations with 63.0% precision and 100.0% recall ($F_1$-score 0.77), narrowing a total of 56 HN clinical trials to a range of 0-4 relevant trials per patient case (mean 1, SD 1.2 trials). In comparison, baseline non–retrieval-augmented GPT-4 demonstrated 0.0% precision, recall, and $F_1$-score—given the lack of response specificity to MSK clinical trials. Subgroup precision varied by cancer types (HN cancers: 72.7%, skin cancers: 50.0%, salivary gland cancers: 36.4%, and thyroid cancers: 33.3%) and the presence of biomarkers (presence 72.7%, absent 62.1%).

**Table 1.** Baseline characteristics of patient cases (N=178).

| Characteristics | Overall values, n (%) |
|---|---|
| Age (years), mean (SD) | 66 (13.9) |
| **Sex** | |
| Female | 44 (24.7) |
| Male | 134 (75.3) |
| **Cancer types** | |
| **Head and neck cancers** | 109 (61.2) |
| Oropharyngeal SCC[a] | 49 (27.5) |
| Oral cavity SCC | 22 (12.4) |
| Laryngeal SCC | 18 (10.1) |
| Hypopharyngeal SCC | 8 (4.5) |
| Other | 12 (6.7) |
| **Thyroid cancers** | 29 (16.3) |
| Anaplastic thyroid carcinoma | 4 (2.2) |
| Differentiated thyroid carcinoma | 25 (14.0) |
| **Skin cancers** | 16 (9.0) |
| **Salivary gland cancers** | 14 (7.9) |
| Adenoid cystic carcinoma | 5 (2.8) |
| Nonadenoid cystic carcinoma | 9 (5.1) |
| **Other cancers** | 10 (5.6) |
| **Cancer stage** | |
| Local/locally advanced | 121 (68.0) |
| Recurrent/metastatic | 57 (32.0) |
| **Biomarkers** | |
| **Present** | 66 (37.1) |
| HPV[b] or p16[c] | 42 (23.6) |
| EBV[d] | 5 (2.8) |
| BRAF[e] mutation | 6 (3.4) |
| RET[f] mutation | 2 (1.1) |
| AR[g] | 2 (1.1) |
| HER2[h] | 3 (1.7) |
| Other | 6 (3.4) |
| **None** | 113 (63.5) |
| **Treatment settings** | |
| Definitive | 93 (52.2) |
| Palliative | 51 (28.7) |
| Surveillance | 15 (8.4) |
| Adjuvant | 13 (7.3) |
| Diagnostic | 6 (3.4) |
| **Treatment modality** | |
| Combined modality therapy | 75 (42.1) |
| Primary systemic treatment | 37 (20.8) |

| Characteristics | Overall values, n (%) |
|---|---|
| Primary surgical treatment | 11 (6.2) |
| Primary radiation treatment | 8 (4.5) |
| Best supportive care | 5 (2.8) |
| Other | 24 (13.5) |
| Clinical trials | 18 (10.1) |

[a]SCC: squamous cell carcinoma.
[b]HPV: human papillomavirus.
[c]p16: p16(INK4A) immunostain.
[d]EBV: Epstein-Barr virus.
[e]BRAF: V-Raf murine sarcoma viral oncogene homolog B.
[f]RET: Rearranged during transfection.
[g]AR: androgen receptor.
[h]HER2: human epidermal growth factor receptor 2.

**Table 2.** Performance of retrieval-augmented large language models in matching physician clinical trial recommendations.

| Performance | Precision (%) | Recall (%) | $F_1$-score |
|---|---|---|---|
| Baseline GPT-4 | 0.0 | 0.0 | 0 |
| Retrieval-augmented GPT-4 | 63.0 | 100.0 | 0.77 |
| **Subgroups (cancer types)** | | | |
| Head and neck cancers | 72.7 | 100.0 | 0.84 |
| Thyroid cancers | 33.3 | 100.0 | 0.50 |
| Skin cancers | 50.0 | 100.0 | 0.67 |
| Salivary gland cancers | 36.4 | 100.0 | 0.53 |
| Other cancers | —[a] | — | — |
| **Subgroups (biomarkers)** | | | |
| Present | 72.7 | 100.0 | 0.84 |
| None | 62.1 | 100.0 | 0.77 |

[a]Not applicable.

## Discussion

Our study demonstrated that retrieval-augmented GPT-4 achieved moderate performance in matching physician clinical trial recommendations in HN oncology. Comparatively, our retrieval-augmented LLM outperformed its pre–fine-tuned baseline and exceeded the historical performance of pretrained LLMs for providing oncology treatment recommendations by 4-20 folds ($F_1$-score 0.04-0.19) [4]. Prior studies have evaluated LLM performance in matching patients to clinical trials, achieving high accuracy [8-10]; however, to our knowledge, our study is the first to evaluate an oncology-specific, retrieval-augmented LLM as a point-of-care, clinical trial decision support application. As our subgroup analyses demonstrated, LLM performance varies based on the specificity of the prompt and dataset, with enhanced precision achieved through reduced search ambiguity for biomarker-specific trials and cancer types with more well-defined datasets. Study limitations included small sample size, short-term assessment, cross-sectional design, disease-specific focus, and being conducted in a single institution, which limits generalizability and subgroup analyses; however, our study provides insights into the rarely measured performance of retrieval-augmented LLMs using real-world patient cases. Future research is needed to optimize LLMs' precision and stability and to assess their implementation and effectiveness as a scalable solution for enhancing clinical trial participation.

### Acknowledgments

This work is supported in part by the Memorial Sloan Kettering Cancer Center Support (grant P30-CA008748) and the 2024 Conquer Cancer—Johnson & Johnson Innovative Medicine Career Development Award (AWD00003905). The corresponding author has full access to all data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. We thank all our patients, providers, and administrative staff who supported the study.

## Conflicts of Interest

TKWH is the founder of LookUpTrials by TeamX Health. ALH received compensation from or was a part of the advisory boards of Eisai, Exelixis, Novartis, Merck, Rgenta, Coherus, Kura oncology, Remix Therapeutics, McGivney Global Advisors, Prelude Therapeutics, Affyimmune, Elevar Therapeutics, Ayala, Nested Therapeutics, and AstraZeneca. He was the principal investigator of clinica trials for Eisai, Bayer, Genentech, AstraZeneca, Novartis, Merck, BMS, Versatem, Remix Therapuetics, Rgenta Therapeutics, Kura Oncology, Ayala, TILT Therapeutics, Hookipa, Novartis, Daiichi Sankyo, and Astellas. ALH is a co-inventor of patent “Lesional dosimetry methods for tailoring targeted radiotherapy in cancer" (Serial number 63/193700, filed 5/27/21) and serves on the Speaker Bureau of Physician Education Resources. The other authors declare no conflicts of interest.

## Multimedia Appendix 1

Preference pairs architecture.
[DOCX File , 16 KB-Multimedia Appendix 1]

## References


1. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. Feb 9, 2023;2(2):e0000198. [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
2. Longwell JB, Grant RC, Hirsch I, Binder F, Jang RW, Krishnan RG. Large language models encode medical oncology knowledge: performance on the ASCO and ESMO examination questions. JCO Oncology Practice. Nov 2023;19(11_suppl):511-511. [doi: 10.1200/op.2023.19.11_suppl.511]
3. Chen S, Kann BH, Foote MB, Aerts HJWL, Savova GK, Mak RH, et al. Use of artificial intelligence Chatbots for cancer treatment information. JAMA Oncol. Oct 01, 2023;9(10):1459-1462. [FREE Full text] [doi: 10.1001/jamaoncol.2023.2954] [Medline: 37615976]
4. Benary M, Wang XD, Schmidt M, Soll D, Hilfenhaus G, Nassir M, et al. Leveraging large language models for decision support in personalized oncology. JAMA Netw Open. Nov 01, 2023;6(11):e2343689. [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.43689] [Medline: 37976064]
5. Hager P, Jungmann F, Holland R, Bhagat K, Hubrecht I, Knauer M, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med. Jul 04, 2024:1-26. [doi: 10.1038/s41591-024-03097-1] [Medline: 38965432]
6. Hung KW, Dunn L, Sherman EJ, Ho AL, Fetten JV, Michel LS, et al. LookUpTrials: assessment of an artificial intelligence-powered mobile application to engage oncology providers in clinical trials. JCO GO. Aug 2023;9(Supplement_1):111-111. [doi: 10.1200/go.2023.9.supplement_1.111]
7. Rafailov R, Sharma A, Mitchell E, Manning C, Ermon S, Finn C. Direct preference optimization: your language model is secretly a reward model. arXiv. Preprint posted online on July 29, 2024. [doi: 10.48550/arXiv.2305.18290]
8. Jin Q, Wang Z, Floudas CS. Matching patients to clinical trials with large language models. arXiv. Preprint posted online on April 27, 2024. [doi: 10.48550/arXiv.2307.15051]
9. Unlu O, Shin J, Mailly CJ, Oates MF, Tucci MR, Varugheese M, et al. Retrieval-augmented generation–enabled GPT-4 for clinical trial screening. NEJM AI. Jun 27, 2024;1(7). [doi: 10.1056/aioa2400181]
10. Wornow M, Lozano A, Dash D, Jindal J, Mahaffey K, Shah N. Zero-shot clinical trial patient matching with LLMs. arXiv. Preprint posted online on April 10, 2024. [doi: 10.48550/arXiv.2402.05125]


## Abbreviations

**HN:** head and neck
**LLM:** large language model
**MSK:** Memorial Sloan Kettering Cancer Center

*Edited by Q Jin; submitted 18.05.24; peer-reviewed by S Chan, D Bracken-Clarke, F Chen; comments to author 20.06.24; revised version received 12.08.24; accepted 03.09.24; published 15.10.24*

*Please cite as:*
*Hung TKW, Kuperman GJ, Sherman EJ, Ho AL, Weng C, Pfister DG, Mao JJ*
*Performance of Retrieval-Augmented Large Language Models to Recommend Head and Neck Cancer Clinical Trials*
*J Med Internet Res 2024;26:e60695*
*URL: https://www.jmir.org/2024/1/e60695*
*doi: 10.2196/60695*
*PMID:*