Performance of Retrieval-Augmented Language Model to Recommend Head and Neck Cancer Clinical Trials

Tony Kin Wai Hung, Gilad J. Kuperman, Eric J Sherman, Alan L. Ho, Chunhua Weng, David G. Pfister, Jun J. Mao

Submitted to: Journal of Medical Internet Research
on: May 18, 2024

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Performance of Retrieval-Augmented Language Model to Recommend Head and Neck Cancer Clinical Trials

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

[1]Memorial Sloan Kettering Cancer Center New York US
[2]Columbia University, Department of Biomedical Informatics New York US

Corresponding Author:
Tony Kin Wai Hung MD, MBA, MSCR
Memorial Sloan Kettering Cancer Center
530 E 74th St
New York
US

Abstract

In this study, we evaluated the performance of a retrieval-augmented language model, powered by GPT-4, to recommend appropriate clinical trial recommendations for a head & neck cancer population at the Memorial Sloan Kettering Cancer Center. We demonstrated that retrieval-augmented LLM could achieve moderate performance, exceeding the historical performance of untrained LLMs to provide oncology treatment recommendations by 4-20 folds. Our study provided insights into the rarely measured performance of retrieval-augmented LLM using real-world patient cases in comparison to physician expert recommendations.

(JMIR Preprints 18/05/2024:60695)
DOI: https://doi.org/10.2196/preprints.60695

Preprint Settings

Original Manuscript

Performance of Retrieval-Augmented Language Model to Recommend Head and Neck Cancer Clinical Trials

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

[1] Memorial Sloan Kettering Cancer Center

[2] Columbia University, Department of Biomedical Informatics

* Corresponding Author: Tony K. W. Hung, MD, MBA, MSCR

Memorial Sloan Kettering Cancer Center | 530 East 74th Street, New York, NY 10021

(646) 608-4127 | HungT@mskcc.org | Twitter: @tonyhungmdmba

Presentation: ASCO Annual Meeting: June 3rd, 2024. 9:00 AM-12:00 PM CDT. Abstract 11081.

Acknowledgments: This work is supported in part by the Memorial Sloan Kettering Cancer Center Support Grant (P30-CA008748). 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 staffs who supported the study.

Disclosure: T.K.W.H is founder of LookUpTrials by TeamX Health.

Keywords: Large Language Model; ChatGPT; Artificial Intelligence; Clinical Trials; Decision Support; LookUpTrials; Cancer Care Delivery; Head and Neck Oncology

Word Count: 600 words

INTRODUCTION

Chatbots based on large language model (LLM) have demonstrated the ability to pass the United States Medical Licensing Examination (USMLE) and answer specialized medical oncology examination questions with impressive accuracy without any training or reinforcement;[1, 2] however, leveraging LLMs in oncology-decision support have not yet demonstrated suitable performance, noting untrained LLMs would produce responses that deviate from cancer expert recommendations and the National Comprehensive Cancer Network (NCCN) guideline.[3, 4] Furthermore, the rapidly changing and complex landscape of oncology – including knowledge on current cancer clinical trials – limits the meaningful use of LLM in practice given delay in training dataset update. To bridge 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 & neck (HN) cancer population.

METHODS

In Feb 2022, we developed and piloted a clinical trial knowledge management application, LookUpTrials, at the Memorial Sloan Kettering Cancer Center (MSK).[5] Using LookUpTrials real-time database, we applied retrieval augmented generation architecture and direct preference optimization to further fine-tune GPT-4 as an in-app assistant.[6] From Nov-7-2023 to Jan-30-2024, we then collected consecutive, new patient cases and their respective clinical trial recommendations from physicians in the HN medical oncology service at MSK. Cases were categorized by diagnosis, cancer stage, treatment setting, and physician recommendation on clinical trials. Using these cases, GPT-4 was prompted by a semi-structured template: "Given patient with a <diagnosis>, <cancer stage>, <treatment setting>, what are possible clinical trials?" GPT-4 responses were compared with physician recommendations with concordance a priori defined: GPT-4 response was a true positive if it included the physician recommended clinical trial(s); true negative if response did not include the physician recommended clinical trial(s); false positive if response recommended clinical trial(s) but physicians did not; and false negative if response did not recommend clinical trial(s) but physicians did. We analyzed the performance of GPT-4 based on its response precision (positive predictive value), recall (sensitivity), and F1 score (harmonic mean of precision and recall). We further analyzed subgroup performance by cancer types. Statistical analyses were performed using JMP-17.2.0. MSK institutional review board approved the study.

RESULTS

We analyzed 178 patient cases, mean age 65.6 (SD 13.9), primarily male (75%) with local/locally advanced (68%) HN (61%), thyroid (16%), skin (9%), or salivary (8%) cancers (*Table 1*). Majority were treated in the definitive setting with combined modality therapy (42%), and a modest proportion were treated under clinical trials (10%). Overall, retrieval-augmented GPT-4 achieved moderate performance (*Table 2*), matching physician clinical trial recommendations with 63% precision and 100% recall (F1 score 0.77), narrowing a total list of 56 HN clinical trials to a range of 0-4 relevant trials per patient case (mean 1, SD 1.2). Subgroup performance of precision varied by cancer types: HN cancers (73%), skin cancers (50%), salivary gland cancers (36%), and thyroid cancers (33%)

DISCUSSION

Our study demonstrated that retrieval-augmented LLM can achieve moderate performance in matching physician clinical trial recommendations in HN oncology. Comparatively, our retrieval-augmented LLM exceeded historical performance of untrained LLMs to provide oncology treatment recommendations by 4-20 folds (F1 score 0.04 - 0.19).[4] Our results suggest the potential of retrieval-augmented LLM to reduce clinician cognitive-load in clinical trial search, although its performance can varied based on dataset specificity. Our study is limited to sample size, cross-sectional, disease-specific, and single-institutional design; however, it provides insights into, rarely measured, performance of retrieval-augmented LLM using real-world patient cases in comparison to physician expert recommendations. Future research is needed to optimize the precision and stability of LLM and to assess its effectiveness as a scalable solution to enhance clinical trial search and participation.

Table 1. Baseline Characteristic of Patient Cases

Characteristic	Overall, No. (%)
All cases	178
Age, mean (SD), y	66 (13.9)
Sex	
Female	44 (25)
Male	134 (75)
Cancer Types	
Head and Neck Cancers	109 (61)
OPC	*49 (28)*
OCC	*22 (12)*
Laryngeal SCC	*18 (10)*
Hypopharyngeal SCC	*8 (4)*
Other	*12 (7)*
Thyroid Cancers	29 (16)
ATC	*4 (2)*
DTC	*25 (14)*
Skin Cancers	16 (9)
Salivary Gland Cancers	14 (8)
ACC	*5 (3)*
Non-ACC	*9 (5)*
Other Cancers	10 (6)
Cancer Stage	
Local or Local Advanced	121 (68)
Recurrent/Metastatic	57 (32)
Biomarkers	
HPV+ or p16+	42 (24)
EBV	5 (3)
BRAF	6 (3)
RET	2 (1)
AR+	2 (1)
HER2+	3 (2)
Other	6 (3)
None	113 (63)
Treatment Settings	
Definitive	93 (52)
Palliative	51 (29)
Surveillance	15 (8)
Adjuvant	13 (7)
Diagnostic	6 (3)
Treatment Modality	
Combined modality therapy	75 (42)
Primary systemic treatment	37 (21)
Primary surgical treatment	11 (6)
Primary radiation treatment	8 (5)
Best supportive care	5 (3)
Other	24 (13)
Clinical trials	18 (10)

Table 2. Performance of Retrieval-Augmented LLM in Matching Physician Clinical Trial Recommendations

Performance	**Precision**	**Recall**	**F1 Score**
Overall	63%	100%	0.77
Subgroups			
Head and Neck Cancers	73%	100%	0.84
Thyroid Cancers	33%	100%	0.50
Skin Cancers	50%	100%	0.67
Salivary Gland Cancers	36%	100%	0.53
Other Cancers	--	--	--

REFERENCES

1. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2: e0000198.

2. Longwell JB, Grant RC, Hirsch I, Binder F, Jang RW-J, Krishnan RG. Large language models encode medical oncology knowledge: Performance on the ASCO and ESMO examination questions. JCO Oncology Practice. 2023;19: 511-511.

3. Chen S, Kann BH, Foote MB, et al. Use of Artificial Intelligence Chatbots for Cancer Treatment Information. JAMA Oncology. 2023;9: 1459-1462.

4. Benary M, Wang XD, Schmidt M, et al. Leveraging Large Language Models for Decision Support in Personalized Oncology. JAMA Network Open. 2023;6: e2343689-e2343689.

5. Hung KW, Dunn L, Sherman EJ, et al. LookUpTrials: Assessment of an artificial intelligence-powered mobile application to engage oncology providers in clinical trials. JCO Global Oncology. 2023;9: 111-111.

6. Rafailov R, Sharma A, Mitchell E, Manning CD, Ermon S, Finn C. Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems. 2024;36.