JAMA Network Open

Original Investigation | Oncology

Performance of Large Language Models on Medical Oncology Examination Questions

Jack B. Longwell, HBSc; Ian Hirsch, MD, MSc; Fernando Binder, MD, MPH; Galileo Arturo Gonzalez Conchas, MD; Daniel Mau, HBSc; Raymond Jang, MD, MSc; Rahul G. Krishnan, PhD; Robert C. Grant, MD, PhD

Abstract

IMPORTANCE Large language models (LLMs) recently developed an unprecedented ability to answer questions. Studies of LLMs from other fields may not generalize to medical oncology, a high-stakes clinical setting requiring rapid integration of new information.

OBJECTIVE To evaluate the accuracy and safety of LLM answers on medical oncology examination questions.

DESIGN, SETTING, AND PARTICIPANTS This cross-sectional study was conducted between May 28 and October 11, 2023. The American Society of Clinical Oncology (ASCO) Oncology Self-Assessment Series on ASCO Connection, the European Society of Medical Oncology (ESMO) Examination Trial questions, and an original set of board-style medical oncology multiple-choice questions were presented to 8 LLMs.

MAIN OUTCOMES AND MEASURES The primary outcome was the percentage of correct answers. Medical oncologists evaluated the explanations provided by the best LLM for accuracy, classified the types of errors, and estimated the likelihood and extent of potential clinical harm.

RESULTS Proprietary LLM 2 correctly answered 125 of 147 questions (85.0%; 95% CI, 78.2%-90.4%; *P* < .001 vs random answering). Proprietary LLM 2 outperformed an earlier version, proprietary LLM 1, which correctly answered 89 of 147 questions (60.5%; 95% CI, 52.2%-68.5%; *P* < .001), and the best open-source LLM, Mixtral-8x7B-v0.1, which correctly answered 87 of 147 questions (59.2%; 95% CI, 50.0%-66.4%; *P* < .001). The explanations provided by proprietary LLM 2 contained no or minor errors for 138 of 147 questions (93.9%; 95% CI, 88.7%-97.2%). Incorrect responses were most commonly associated with errors in information retrieval, particularly with recent publications, followed by erroneous reasoning and reading comprehension. If acted upon in clinical practice, 18 of 22 incorrect answers (81.8%; 95% CI, 59.7%-94.8%) would have a medium or high likelihood of moderate to severe harm.

CONCLUSIONS AND RELEVANCE In this cross-sectional study of the performance of LLMs on medical oncology examination questions, the best LLM answered questions with remarkable performance, although errors raised safety concerns. These results demonstrated an opportunity to develop and evaluate LLMs to improve health care clinician experiences and patient care, considering the potential impact on capabilities and safety.

JAMA Network Open. 2024;7(6):e2417641. doi:10.1001/jamanetworkopen.2024.17641

Key Points

Question What medical oncology knowledge is encoded by large language models (LLMs)?

Findings In this cross-sectional study evaluating 8 LLMs, proprietary LLM 2 correctly answered 85.0% of examination-style multiple-choice questions from the American Society of Oncology, the European Society of Medical Oncology, and an original set from the authors, outperforming proprietary LLM 1 and open-source models. However, 81.8% of incorrect answers were rated as having a medium or high likelihood of moderate to severe harm if acted upon in practice.

Meaning These findings suggest that LLMs can accurately answer questions requiring advanced knowledge of medical oncology, although errors may cause harm.

+ Supplemental content

Author affiliations and article information are listed at the end of this article.

Introduction

Large language models (LLMs) have the potential to transform health care.[1,2] LLMs are deep learning systems trained using massive corpora of text from the internet to predict the next word in a sentence, which are then fine-tuned to perform specific tasks such as answering questions with human-like responses.[3] Potential applications of LLMs in oncology are broad, ranging from assisting clinicians with administrative tasks or decision-making to interacting with patients to provide medical or emotional counselling.[4,5]

LLMs encode a remarkable amount of medical knowledge without specific training. Just like with medical trainees, we can evaluate the medical knowledge of LLMs using standardized tests. LLMs can pass the US Medical Licensing Examination[6,7] and provide a rationale for answers that display comprehension of the question, recall of the relevant knowledge, and reasoning to arrive at the solution. However, the performance on examinations across different medical subspecialties has varied.[8-12]

Dynamic and specialized knowledge characterize medical oncology. For example, the average rate of US Food and Drug Administration approvals surged from 7.4 per year from 2000 to 2004 to 56 per year from November 2017 to October 2022.[13] Knowledge evolves swiftly, and a single trial can change practice compared with all preceding literature. Furthermore, the volume of published knowledge on cancer is vast, with over 3 million PubMed studies about cancer.[14] Whether the performance of LLMs transfers to medical oncology remains largely unknown, with current research generally limited to specific cancer types[15-18] or general questions.[19]

In this study, we evaluated the medical oncology knowledge encoded in LLMs using standardized examination-style questions from the American Society of Clinical Oncology (ASCO), the European Society for Medical Oncology (ESMO), and original examination-style questions created by our team. Our goal was to understand the accuracy and limitations of LLMs applied to medical oncology to guide future research and applications.

Methods

We conducted this cross-sectional study between May 28 and October 11, 2023. The study was exempt from research ethics board approval and the need for informed consent in accordance with 45 CFR §46, given the lack of involvement of human participants. We followed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guideline for cross-sectional studies.

Questions

ASCO

ASCO's Oncology Self-Assessment Series on ASCO Connection is a publicly accessible question bank hosted on the organization's official website.[20] Questions are multiple-choice with one correct answer and cover various topics. The website provides the correct answer and an explanation referencing pertinent published papers for each question. The bank includes 52 questions, with 49 questions containing 4 multiple-choice options and 4 having 5 options.

ESMO

ESMO provides multiple-choice questions with a single correct answer from the previous 2 years of ESMO's Examination Trial Questions from the ESMO Academy (2021 and 2022, type A).[21] These questions are "specifically designed for medical oncologists in training who intend to take the ESMO examination. [They are] also suitable for oncology professionals who are in need of a refresher course"[22]. We discarded questions based on images because of the LLM's capability at the time, leaving 75 of 80 questions for the final set. Of the chosen questions, 74 contained 5 multiple-choice options, and 1 included 4. ESMO provided the right answers, but explanations and references were developed by 2 medical oncologists (I.H. and F.B.).

Original Questions

The training data sources used by proprietary LLMs are not publicly available and could theoretically include these examinations. To remove any potential impact of questions from ASCO or ESMO included during the training of the LLMs, we tested performance on unseen original questions. Three oncologists (I.H., F.B., and G.A.G.C.) developed 20 original questions. The questions resembled ASCO and ESMO, maintaining the multiple-choice format, with 16 questions containing 4 options and 4 with 5 options. We justified each answer with references to the literature. One question was miswritten and sent through a more recent LLM.

Large Language Models

ChatGPT-3.5 (proprietary LLM 1; May 28, 2023 version; OpenAI) and ChatGPT-4 (proprietary LLM 2; May 28, 2023 version; OpenAI) were accessed through ChatGPT's website.[23] We presented all questions to proprietary LLM 1 and proprietary LLM 2 from May 28 to July 19, 2023, except the miswritten question, which was presented on October 11, 2023. We labeled the options for the ESMO multiple-choice answers A through E for consistency with ASCO and original questions. LLMs were prompted by copying and pasting the questions and potential answers into the user interface. Additional prompting techniques are described in the eMethods in Supplement 1.

For comparison, we evaluated 6 open-source LLMs with publicly available weights ranked highly on Chatbot Arena[24]: Mistral-7B-Instruct-v0.2,[25] Mixtral-8x7B-v0.1,[26] Llama-2-13b-chat,[27] Nous-Hermes-Llama2-70b,[28] openchat-3.5-1210,[29] and BioMistral-7B DARE.[30] BioMistral-7B DARE is tailored for biomedical domains. Technical details on the use of these LLMs are described in the eMethods in Supplement 1.

Data Coding

Each response was recorded, including the chosen letter answer and prose explanation. Two oncologists (I.H. and F.B.) classified the prose that accompanied answers into a 4-level ordinal error scale: wrong with major errors, wrong with minor errors, right with minor errors, and right with no errors. Major errors were defined as explanations showing a limited understanding of the core concepts required to answer the question in relation to the scientific and clinical consensus. Minor errors were defined as explanations consistent with an understanding of the core concepts, although the final answer was incorrect. Discussion with a third medical oncologist (R.G.) solved disagreements. Following a previously published framework, the oncologists evaluated answers for agreement with the scientific and clinical consensus, the likelihood and extent of possible harm, reading comprehension, recall of relevant clinical knowledge, and manipulation of knowledge via valid reasoning.[7] A medical oncologist (R.G.) categorized questions based on the year of the most recent publication needed to arrive at the correct answer. If the question did not require information from a publication after 2018, no specific year was assigned, and these questions were deemed as not requiring recent information. Finally, the oncologists marked questions where the answer provided by ESMO or ASCO may be incorrect in light of more current literature or may reflect regional variations.

Statistical Analysis

Our primary analysis tested whether the best model provided more accurate answers than random chance alone using the binomial distribution. McNemar test compared the distribution of correct answers between LLMs, while Fisher test compared the distribution of errors within explanations associated with incorrect answers. Weighted κ with linear weights[31] assessed if there was an agreement between the 2 oncologists evaluating the accuracy of the best model's explanations. A Wilcoxon rank sum test compared the distribution of reference years among right and wrong answers.[32] A 2-sided *P* value of .05 indicated statistical significance. We conducted statistical analysis in R version 4.3.0 (R Project for Statistical Computing).[33]

Results

We evaluated LLMs across 147 examination questions, including 52 from ASCO, 75 from ESMO, and 20 original questions (**Table**; eTable 1 in Supplement 1). The most common category was hematology (22 of 147 [15.0%]), although the questions covered a broad range of topics. ESMO included more general questions, such as considering the mechanisms and toxic effects of systemic therapies. From the entire question set, 41 of 147 (27.9%) required knowledge of evidence published from 2018 onwards. LLMs provided answers in prose to all questions (**Box**), with proprietary LLM 2 requiring a prompt to select a specific answer in 33 of 147 questions (22.4%).

Proprietary LLM 2 was most accurate among LLMs, correctly answering 125 of 147 questions (85.0%; 95% CI, 78.2%-90.4%; *P* < .001 vs random answering). Prompt engineering did not enable proprietary LLM 2 to correctly answer any of the questions that were incorrectly answered without specific prompting. Twelve of 22 incorrect answers (54.5%; 95% CI, 32.2%-75.6%) were corrected when proprietary LLM 2 was offered a second attempt. Performance on each of the specific sets of questions was similar, with 42 of 52 on ASCO (80.8%; 95% CI, 67.5%-90.4%; *P* < .001 vs random answering), 66 of 75 on ESMO (88.0%; 95% CI, 78.4%-94.4%; *P* < .001), and 17 of 20 on the original questions (85.0%; 95% CI, 62.1%-96.8%; *P* < .001) (**Figure 1**).

Proprietary LLM 1 correctly answered 89 of 147 questions (60.5%; 95% CI, 39.5%-67.8%; *P* < .001 vs random answering) of the questions, which was inferior to proprietary LLM 2 (χ^2_1 = 24.596; *P* < .001). Mixtral-8x7B-v0.1 was best among open-source LLMs (eFigure in Supplement 1), correctly answering 87 of 147 questions (59.2%; 95% CI, 50.0%-66.4%), which was not significantly different from proprietary LLM 1 (χ^2_1 = 0.021; *P* > .88) but was inferior to proprietary LLM 2 (χ^2_1 = 23.564; *P* < .001). BioMistral-7B DARE, an LLM tuned for the biomedical domain, answered 50 of 147 questions (33.6%; 95% CI, 26.0%-41.7%) correctly, inferior to both proprietary LLM 1 (χ^2_1 = 22.737; *P* < .001) and proprietary LLM 2 (χ^2_1 = 60.012; *P* < .001).

The prose provided by proprietary LLM 2 answered most questions correctly without errors when evaluated qualitatively by clinicians according to the scientific and clinical consensus (123 of 147 [83.7%]; 95% CI, 76.7%-89.3%) (**Figure 2**). There was a high concordance between clinician ratings

Table. Characteristics of the Examination-Style Oncology Questions Used to Evaluate Models

	No. of questions (%)			
Characteristic	ASCO Oncology Self-Assessment (n = 52)	ESMO Examination Trial questions (n = 75)	Original questions (n = 20)	Total (N = 147)
Topic				
Breast	6 (4.1)	9 (6.1)	0	15 (10.0)
General/rarer tumor	0	13 (8.8)	1 (0.7)	14 (9.5)
Gastrointestinal	5 (3.4)	10 (6.8)	4 (2.7)	19 (12.9)
Genitourinary	5 (3.4)	9 (6.1)	3 (2.0)	17 (11.6)
Hematology	10 (6.8)	8 (5.4)	4 (2.7)	22 (15.0)
Other	26 (17.7)	26 (17.7)	8 (5.4)	60 (40.8)
Year the most recent evidence required for the correct answer was published				
Before 2018	32 (21.8)	65 (44.2)	9 (6.1)	106 (72.1)
2018	5 (3.4)	6 (4.1)	0	11 (7.5)
2019	5 (3.4)	2 (1.3)	2 (1.3)	9 (6.1)
2020	7 (4.8)	2 (1.3)	6 (4.1)	15 (10)
2021	3 (2.0)	0	1 (0.7)	4 (2.7)
2022	0	0	2 (1.3)	2 (1.3)

Abbreviations: ASCO, American Society of Clinical Oncology; ESMO, European Society for Medical Oncology.

(weighted κ 83.4%; 95% CI, 80.0%-87.0%). Among incorrect answers, 13 of 22 (59.1%; 95% CI, 36.4%-79.3%) contained minor errors, and 9 of 22 (40.9%; 95% CI, 20.7%-63.6%) had major errors. We did not observe any hallucinations, which is when an LLM writes content that cannot be verified or conflicts with facts.[34]

Incorrect answers were more common when the questions required knowledge of recent publications (Wilcoxon test *P* = .02) (**Figure 3**): 12 of 106 questions (11.3%; 95% CI, 6.0%-18.9%) requiring only information before 2018 were incorrect, compared with 3 of 20 questions during 2018 and 2019 (15.0%; 95% CI, 3.2%-37.9%), and 7 of 21 questions after 2019 (33.3%; 95% CI, 14.6%-57.0%). Proprietary LLM 2 was trained on data until September 2021, and incorrectly

Box. Two Selected Questions From the ASCO Oncology Self-Assessment, With Response From Large Language Model (LLM) 2

Question 1

A 62-y-old woman is receiving palliative chemotherapy for metastatic breast cancer to the lung and bone. She presents to the clinic before her next round of chemotherapy with sudden-onset shortness of breath and a cough. No one in her home has had similar symptoms. She has not been febrile, and her cough is nonproductive. Two days ago, she traveled by car from her relative's house, which is six hours away. This morning, she noticed that she had more swelling in her legs than usual and she reports right mid-back pain when she takes a deep breath. A complete blood count shows her white blood count and differential to be within normal limits. A CT angiogram shows right lower lobe pulmonary embolus. Which of the following treatments would be most appropriate?

A. Subcutaneous heparin

B. Warfarin

C. Low molecular weight heparin, followed by oral warfarin

D. Low molecular weight heparin or a direct oral anticoagulant

Proprietary LLM 2 Response

The patient's presentation of sudden-onset shortness of breath, right mid-back pain with deep breath, and leg swelling, along with the finding of a right lower lobe pulmonary embolus on CT angiogram, is suggestive of acute pulmonary embolism. The most likely cause in this patient is venous thromboembolism secondary to her metastatic cancer and recent prolonged travel.

In the setting of cancer, low molecular weight heparin (LMWH) or a direct oral anticoagulant (DOAC) is recommended as first-line treatment for acute venous thromboembolism, including pulmonary embolism. This is due to the fact that cancer patients have a higher rate of recurrent venous thromboembolism and a higher risk of bleeding on warfarin compared to patients without cancer. Therefore, the correct answer is D. Low molecular weight heparin or a direct oral anticoagulant. Subcutaneous heparin and warfarin are not the preferred treatments in this setting. Bridging with LMWH followed by warfarin (option C) is also less preferable because warfarin has a higher risk of bleeding in cancer patients and is less effective in preventing recurrent VTE compared with LMWH or DOAC.

Question 2

A 61-y-old woman presents to the oncology clinic for follow-up after completing eight cycles of treatment for right-sided colon cancer metastatic to the liver and lung. She was treated with 5-fluorouracil, leucovorin, and oxaliplatin (FOLFOX) and bevacizumab for 4 mos. Prior CT scans showed a partial response to treatment with reduction in disease burden in her liver and lungs. She reports developing fatigue that prevents her from hiking, as well as tingling neuropathy that prevents her from gardening and preparing meals. Molecular testing of her tumor shows microsatellite stable disease, wild-type RAS, and a *BRAF* V600E mutation.

Which of the following is the most appropriate next step?

A. Continue FOLFOX and bevacizumab

B. Switch to capecitabine and bevacizumab

C. Switch to encorafenib and cetuximab

D. Switch to fluorouracil and irinotecan (FOLFIRI), plus bevacizumab

Proprietary LLM 2 Response

The patient in this scenario is experiencing neuropathy, a common side effect of oxaliplatin, a component of the FOLFOX regimen. Oxaliplatin-induced peripheral neuropathy is typically dose-dependent and cumulative, often leading to symptoms such as numbness, tingling, and discomfort in the hands and feet. These side effects can significantly affect the patient's quality of life and can be an indication to change chemotherapy regimens.

Capecitabine could be an alternative, as it is another form of chemotherapy that does not have neuropathy as a common side effect. However, the presence of the *BRAF* V600E mutation is important to note, as it implies a less favorable prognosis but also opens the opportunity for targeted therapy.

The *BRAF* V600E mutation is found in approximately 10% of colorectal cancers and is associated with a poorer prognosis. Fortunately, targeted therapies are available. A combination of encorafenib (a *BRAF* inhibitor) and cetuximab (an *EGFR* inhibitor) has been shown in clinical trials to improve survival in patients with metastatic colorectal cancer harboring the BRAF V600E mutation.

Therefore, given the patient's adverse effects with the FOLFOX regimen and the presence of the *BRAF* V600E mutation, the most appropriate next step would be to switch to encorafenib and cetuximab (Choice C).

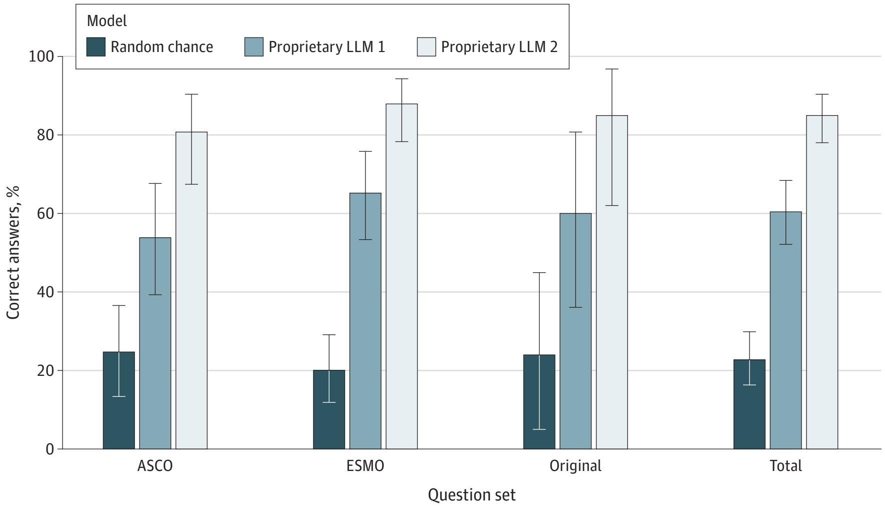
Abbreviations: ASCO, American Society of Clinical Oncology; CT, computed tomography; VTE, venous thromboembolism.

answered both questions (2 of 2 [100%]; 95% CI, 15.8%-100.0%) that required knowledge after that cut-off.

The medical oncologists identified 1 answer from proprietary LLM 2 classified as incorrect that reflects rapidly changing literature and regional practice variations around rectal cancer. For this question on treatment recommendations of a cT3N1 high rectal tumor at 15 cm from the anal verge, proprietary LLM 2 recommended neoadjuvant chemoradiotherapy, while ESMO recommended upfront surgery followed by adjuvant therapy.

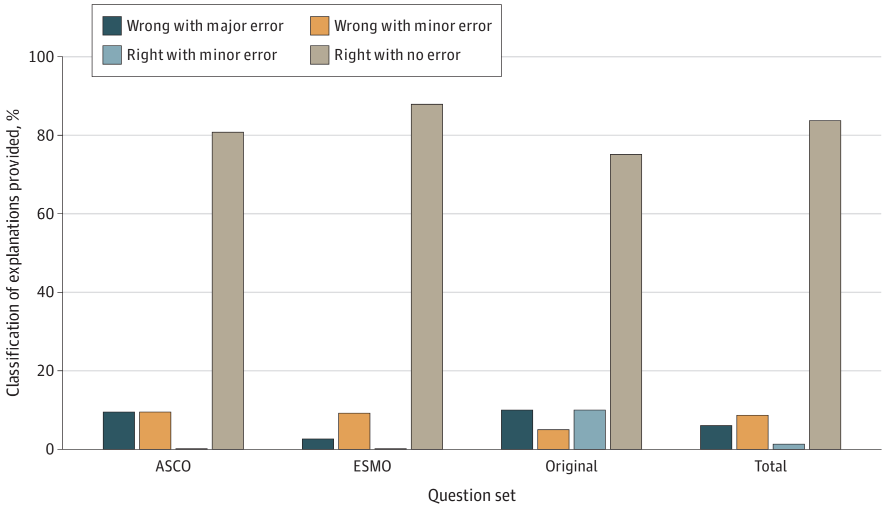
The prose accompanying incorrect answers contained the following errors in knowledge recall, reasoning, and reading comprehension (eTable 2 in Supplement 2): 14 of 22 (63.6%) contained errors in knowledge recall alone, while 3 of 22 (13.6%) were due to mistakes in reasoning, and 3 of 22 (13.6%) were due to errors in both recall and reasoning. Only 2 of 22 incorrect answers (9.0%) were associated with erroneous reading comprehension, 1 of which also exhibited errors in knowledge recall. As an example, 1 question asked about a treatment recommendation for a patient diagnosed

Figure 1. Percentage of Correct Answers Expected by Random Chance and Achieved by 2 Proprietary Large Language Models (LLMs)



Error bars indicate 95% CIs. ASCO indicates American Society of Clinical Oncology; ESMO, European Society for Medical Oncology.

Figure 2. Classifications of Explanations for Answers Provided by Proprietary Large Language Model 2 in Relation to the Scientific and Clinical Consensus



ASCO indicates American Society of Clinical Oncology; ESMO, European Society for Medical Oncology.

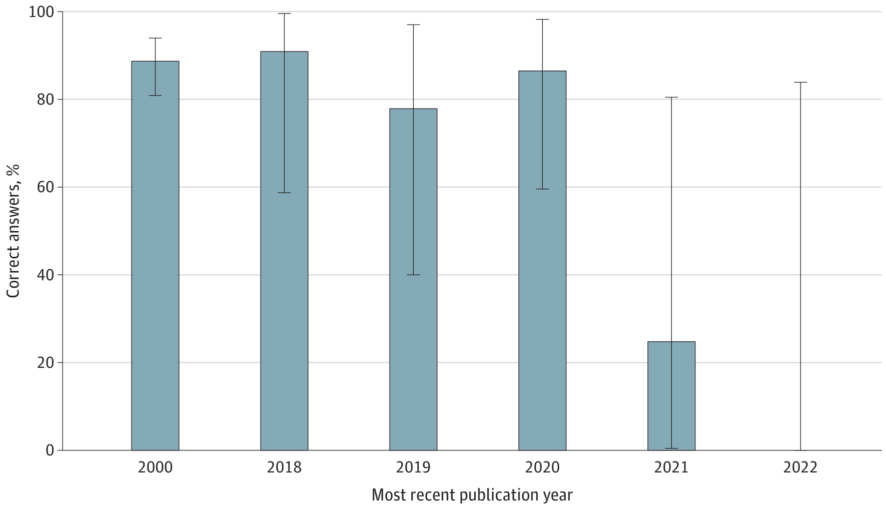
with metastatic right-sided colon cancer with a *BRAF* V600E variant who achieved a partial response to 5-fluorouracil, oxaliplatin, and bevacizumab but also developed peripheral sensory neuropathy. proprietary LLM 2 recommended switching to targeted therapy instead of maintaining the same chemotherapy regimen without oxaliplatin, which we classified as an error in reasoning.

Among incorrect answers, clinicians considered the likelihood of the error causing patient harm in practice to be medium in 14 of 22 questions (63.6%; 95% CI, 43.0%-85.4%) and high in 4 of 22 questions (18.2%; 95% CI, 5.2%-40.3%). The explanations provided by proprietary LLM 2 contained no or minor errors for 138 of 147 questions (93.9%; 95% CI, 88.7%-97.2%). Incorrect responses were most commonly associated with errors in information retrieval, particularly with recent publications, followed by erroneous reasoning and reading comprehension. If acted upon in clinical practice, 18 of 22 incorrect answers (81.8%; 95% CI, 59.7%-94.8%) would have a medium or high likelihood of moderate to severe harm. Clinicians rated the extent of possible harm as mild in 4 of 22 questions (18.2%; 95% CI, 5.2%-40.3%), moderate in 14 of 22 questions (63.6%; 95% CI, 43.0%-85.4%), and likely to be severe or cause death in 4 of 22 questions (18.2%; 95% CI, 5.2%-40.3%). Whenever there was a high likelihood of harm, the extent was severe, and the same for low likelihood and mild extent. Returning to the previous incorrect answer as an example (Box), we deemed proprietary LLM 2's recommendation to abandon effective therapy to have a medium likelihood and moderate possible extent for harm, given the recommended course of action could lead to suboptimal treatment of the cancer but does not have immediate life-threatening implications. Four answers were rated as having a high likelihood of harm with potential for severe or life-threatening harm. For example, 1 ESMO question concerned a patient with muscle-invasive bladder cancer who presented with acute kidney injury, whose creatinine clearance only recovered to 38 mL/min with nephrostomy tubes. ChatGPT's answer recommended neoadjuvant cisplatin-based chemotherapy, which is the standard approach but would be dangerous in a patient with substantially reduced kidney function.

Discussion

In this study, LLMs performed remarkably well on medical oncology examination-style questions designed for the final stages of training before entering clinical practice. The most recent version, proprietary LLM 2, correctly answered 85.0% of multiple-choice questions and provided accurate written explanations supporting the answers. These results demonstrate that LLMs contain substantial medical oncology knowledge and can provide written responses demonstrating

Figure 3. Correct Answers to Medical Oncology Examination-Style Questions by Proprietary Large Language Model 2, by Year of the Most Recent Publication Required to Answer the Question Correctly



Error bars indicate 95% CIs.

comprehension, retrieval, and reasoning. However, the incorrect answers raise critical safety concerns particularly relevant to medical oncology, where dynamic evidence informs decisions with high-stakes consequences. The explanations accompanying incorrect answers from proprietary LLM 2 showed errors in information retrieval, especially with recent publications. If acted upon in clinical practice, most incorrect answers from proprietary LLM 2 displayed a potential for harm.

Proprietary LLM 2 answered 85.0% of medical oncology questions correctly, the highest proportion reported on medical examination questions to date. Proprietary LLM 2 substantially outperformed its earlier version, proprietary LLM 1. This improvement may explain some of the variation in performance observed across examinations from other medical subspecialties, where proprietary LLM 1 and other LLMs answered 45% to 77% of questions correctly,[6-9,11,35] while proprietary LLM 2 answered 57% to 84% correctly.[10,12] Our results extend beyond a recent study[36] where proprietary LLM 1 correctly answered 56.1% (583 of 1040) of questions on the ASCO Oncology Self-Assessment Series. We included ESMO examination questions, as well as original questions that could not be in the training set. We also characterized the errors and estimated their potential for harm. Finally, we included proprietary LLM 2 and open-source LLMs. Extrapolating these results, the capabilities of LLMs will likely continue to improve rapidly with larger training datasets and models[37] and other technical advances.

The performance of LLMs was comparable with medical oncology trainees. According to the annual Examination Report by ESMO, in 2022, the passing threshold was 53.2% correct answers, with the average candidate achieving 60.7% (standard deviation of 11.8%). ASCO does not release summary scores for the ASCO Oncology Self-Assessment Series on ASCO Connection. However, as a comparison, one author participated in the 2019 ASCO In-Training exam as a trainee and answered 35% to 88% of questions correctly across different topics, ranking at the 91st percentile.

The incorrect answers from ChatGPT were most commonly associated with explanations that contained errors in information retrieval, particularly with more recent publications. Errors in retrieving recent information may be intrinsic to the current LLMs because the pretraining procedures weight all text in the training data equally, without specific emphasis on more recent or high-quality information.[27] We could not refocus the model on the most recent evidence through prompt engineering.[38] Furthermore, ChatGPT training data only included information up until September 2021. New training procedures and frequent updates may be required for LLMs to provide up-to-date medical oncology information.

Open-source LLMs offer potential advantages over proprietary ones, including lower costs and opportunities for customization. On these examination questions, the best open-source LLM, Mixtral-8x7B-v0.1, was comparable with the proprietary LLM 1 but inferior to proprietary LLM 2. Open-source LLMs can be customized through pretraining, such as with personal health information or fine-tuning for specific tasks, such as mortality prediction.[39] Surprisingly, BioMistral-7B DARE,[30] an LLM that pretrained the Mistral foundation model on PubMed Central, underperformed. Together, these results demonstrate a current gap between open-source and proprietary LLMs and highlight the need for research on how to pretrain or fine-tune models for specific health care tasks.

We evaluated LLMs on examination questions as a proxy for their potential when applied to medical oncology more broadly. Proposed use cases include drafting responses to patient inbox messages,[40] generating clinical reports from patient encounters,[41] as well as other research and administrative activities.[42] How our results extend to specific clinical use cases requires further research. Of note, we rated most incorrect answers from proprietary LLM 1 and 2 as having at least a medium likelihood of causing harm, with the possible harm caused as being at least moderate in severity. These results contrast with previous research in general medicine, where clinicians mostly rated responses from LLMs as having a low likelihood of harm.[7] These findings suggest that the safe use of LLMs in medical oncology may be restricted to low-risk settings or require intensive human oversight. Future research should evaluate techniques to improve safety, such as retrieval augmented generation[43] and fine-tuning with human feedback.[44,45] Guidelines should also be developed to promote the safe deployment of LLMs in medical oncology.[46,47]

Limitations

Our study has several limitations. First, classifications such as the severity and cause of the errors and the likelihood and extent of potential harm are subjective. While our ratings showed high interrater concordance, we acknowledge that others could reasonably disagree with specific ratings. Second, the training data used to create proprietary LLM 1 and 2 is private, so we cannot rule out the possibility that training included the ASCO and ESMO examinations. However, we observed similar results on a set of original questions created for the study, making the performance observed in this study unlikely to reflect memorization during LLM training. Third, our analysis of incorrect answers was based on a small sample size because proprietary LLM 2 made few errors. Fourth, we considered popular and performant LLMs available when the manuscript was drafted, so whether results will generalize to future LLMs is unknown, considering that LLMs are rapidly improving and results from LLMs change over time.[48]

Conclusions

In this cross-sectional study of the performance of LLMs or examination-style questions in medical oncology, the remarkable performance of proprietary LLM 1 and 2 suggest an urgent opportunity to improve the practice of medical oncology. Potential applications of LLMs span the clinician experience, direct patient care, education, and research.[49] Our results should encourage oncologists and researchers to develop LLM-based tools within specific workflows and evaluate their impact on measurable outcomes, paying particular attention to safety.[50]

ARTICLE INFORMATION

Accepted for Publication: April 18, 2024.

Published: June 18, 2024. doi:10.1001/jamanetworkopen.2024.17641

Corresponding Author: Robert C. Grant, MD, PhD, Princess Margaret Cancer Centre, 7-811 700 University Ave, Toronto, ON M5G 1X6, Canada (robert.grant@uhn.ca).

Author Affiliations: Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada (Longwell, Hirsch, Binder, Gonzalez Conchas, Mau, Jang, Grant); Department of Medicine, University of Toronto, Toronto, Ontario, Canada (Hirsch, Binder, Jang, Grant); Institute of Medical Science, University of Toronto, Toronto, Ontario, Canada (Mau, Grant); Department of Computer Science, University of Toronto, Toronto, Ontario, Canada (Krishnan); Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario, Canada (Krishnan); Vector Institute, Toronto, Ontario, Canada (Krishnan); ICES, Toronto, Ontario, Canada (Grant).

Author Contributions: Dr Grant had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. Mr Longwell and Dr Hirsch were co–first authors.

Concept and design: Longwell, Hirsch, Jang, Krishnan, Grant.

Acquisition, analysis, or interpretation of data: Longwell, Hirsch, Binder, Gonzalez Conchas, Mau, Grant.

Drafting of the manuscript: Longwell, Hirsch, Binder, Gonzalez Conchas, Jang, Grant.

Critical review of the manuscript for important intellectual content: Longwell, Hirsch, Gonzalez Conchas, Mau, Krishnan, Grant.

Statistical analysis: Longwell, Grant.

Obtained funding: Grant.

Administrative, technical, or material support: Longwell, Binder, Gonzalez Conchas, Mau, Grant.

Supervision: Hirsch, Krishnan, Grant.

Conflict of Interest Disclosures: Dr Krishnan reported holding options in Iterative Scopes and ViyaMD outside the submitted work. Dr Grant reported receiving personal fees from Astra Zeneca, Eisai, Incyte, Knight Therapeutics, Guardant Health, and Ipsen outside the submitted work. No other disclosures were reported.

Funding/Support: Mr Longwell received financial support and compensation for this study from the TD Ready Challenge Grant and the Princess Margaret Cancer Foundation. Drs Hirsch, Binder, Gonzalez Conchas, and Mr Mau received financial support and compensation for this study from the Princess Margaret Cancer Foundation. Dr Grant received financial support for this study from the TD Ready Challenge Grant, the AMS Fellowship in Compassion and Artificial Intelligence, the Princess Margaret Cancer Foundation, and the Ontario Institute for Cancer Research.

Role of the Funder/Sponsor: The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Data Sharing Statement: See Supplement 3.

REFERENCES

1. Drazen JM, Kohane IS, Leong TY. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med*. 2023;388(13):93. doi:10.1056/NEJMsr2214184

2. Korngiebel DM, Mooney SD. Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery. *NPJ Digit Med*. 2021;4(1):93. doi:10.1038/s41746-021-00464-x

3. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. *Adv Neural Inf Process Syst*. 2020;33:1877-1901. doi:10.5555/3495724.3495883

4. Iannantuono GM, Bracken-Clarke D, Floudas CS, Roselli M, Gulley JL, Karzai F. Applications of large language models in cancer care: current evidence and future perspectives. *Front Oncol*. 2023;13:1268915. doi:10.3389/fonc.2023.1268915

5. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare (Basel)*. 2023;11(6):887. doi:10.3390/healthcare11060887

6. 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(2):e0000198. doi:10.1371/journal.pdig.0000198

7. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172-180. 2023;620(7972):172-180. doi:10.1038/s41586-023-06291-2

8. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. *Radiology*. 2023;307(5):e230582. doi:10.1148/radiol.230582

9. Beam K, Sharma P, Kumar B, et al. Performance of a large language model on practice questions for the neonatal board examination. *JAMA Pediatr*. 2023;177(9):977-979. doi:10.1001/jamapediatrics.2023.2373

10. Mihalache A, Huang RS, Popovic MM, Muni RH. Performance of an upgraded artificial intelligence chatbot for ophthalmic knowledge assessment. *JAMA Ophthalmol*. 2023;141(8):798-800. doi:10.1001/jamaophthalmol.2023.2754

11. Deebel NA, Terlecki R. ChatGPT performance on the American Urological Association self-assessment study program and the potential influence of artificial intelligence in urologic training. *Urology*. 2023;177:29-33. doi:10.1016/j.urology.2023.05.010

12. Hoch CC, Wollenberg B, Lüers JC, et al. ChatGPT's quiz skills in different otolaryngology subspecialties: an analysis of 2576 single-choice and multiple-choice board certification preparation questions. *Eur Arch Otorhinolaryngol*. 2023;280(9):4271-4278. doi:10.1007/s00405-023-08051-4

13. Scott EC, Baines AC, Gong Y, et al. Trends in the approval of cancer therapies by the FDA in the twenty-first century. *Nat Rev Drug Discov*. 2023;22(8):625-640. doi:10.1038/s41573-023-00723-4

14. Reyes-Aldasoro CC. The proportion of cancer-related entries in PubMed has increased considerably; is cancer truly "the emperor of all maladies"? *PLoS One*. 2017;12(3):e0173671. doi:10.1371/journal.pone.0173671

15. Samaan JS, Yeo YH, Rajeev N, et al. Assessing the accuracy of responses by the language model ChatGPT to questions regarding bariatric surgery. *Obes Surg*. 2023;33(6):1790-1796. doi:10.1007/s11695-023-06603-5

16. Haver HL, Ambinder EB, Bahl M, Oluyemi ET, Jeudy J, Yi PH. Appropriateness of breast cancer prevention and screening recommendations provided by ChatGPT. *Radiology*. 2023;307(4):e230424. doi:10.1148/radiol.230424

17. Moazzam Z, Cloyd J, Lima HA, Pawlik TM. Quality of ChatGPT responses to questions related to pancreatic cancer and its surgical care. *Ann Surg Oncol*. 2023;30(11):6284-6286. doi:10.1245/s10434-023-13777-w

18. Coskun B, Ocakoglu G, Yetemen M, Kaygisiz O. Can ChatGPT, an artificial intelligence language model, provide accurate and high-quality patient information on prostate cancer? *Urology*. 2023;180:35-58. doi:10.1016/j.urology.2023.05.040

19. Chen S, Kann BH, Foote MB, et al. Use of artificial intelligence chatbots for cancer treatment information. *JAMA Oncol*. 2023;9(10):1459-1462. doi:10.1001/jamaoncol.2023.2954

20. American Society of Clinical Oncology. Oncology Self-Assessment. Accessed June 5, 2023. https://connection.asco.org/tec/Oncology%20Self-Assessment

21. ESMO. ESMO examination trial questions. Accessed June 21, 2023. https://tryexam.esmo.org/type-a

22. ESMO. ESMO academy. Accessed August 1, 2023. https://www.esmo.org/about-esmo-meetings/esmo-academy

23. ChatGPT. Accessed August 1, 2023. https://chat.openai.com/

24. Chiang WL, Zheng L, Sheng Y, et al. Chatbot arena: an open platform for evaluating LLMs by human preference. *arXiv*. Preprint posted online March 7, 2024. doi:10.48550/arXiv.2403.04132

25. Jiang AQ, Sablayrolles A, Mensch A, et al. Mistral 7B. *arXiv*. Preprint posted online October 10, 2023. doi:10.48550/ARXIV.2310.06825

26. Jiang AQ, Sablayrolles A, Roux A, et al. Mixtral of experts. *arXiv*. Preprint posted online January 8, 2024. doi:10.48550/arXiv.2401.04088

27. Touvron H, Martin L, Stone K, et al. Llama 2: open foundation and fine-tuned chat models. *arXiv*. Preprint posted online July 18, 2023. doi:10.48550/arXiv.2307.09288

28. Hugging Face. Nous-Hermes-Llama2-70b. Accessed April 14, 2024. https://huggingface.co/NousResearch/Nous-Hermes-Llama2-70b

29. Wang G, Cheng S, Zhan X, Li X, Song S, Liu Y. OpenChat: advancing open-source language models with mixed-quality data. *arXiv*. Preprint posted online September 20, 2023. doi:10.48550/arXiv.2309.11235

30. Labrak Y, Bazoge A, Morin E, Gourraud PA, Rouvier M, Dufour R. BioMistral: a collection of open-source pretrained large language models for medical domains. *arXiv*. Preprint posted online February 15, 2024. doi:10.48550/arXiv.2402.10373

31. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas*. 1960;20(1):37-46. doi:10.1177/001316446002000104

32. Wilcoxon F. Individual comparisons by ranking methods. In: Kotz S, Johnson NL, eds. *Breakthroughs in Statistics: Methodology and Distribution*. Springer New York; 1992:196-202.

33. Ripley BD. The R project in statistical computing. *MSOR Connect.* 2001;1(1):23-25. doi:10.11120/msor.2001.01010023

34. Li J, Cheng X, Zhao WX, Nie JY, Wen JR. HaluEval: a large-scale hallucination evaluation benchmark for large language models. *arXiv*. Preprint posted online May 19, 2023. doi:10.18653/v1/2023.emnlp-main.397

35. Yalamanchili A, Sengupta B, Song J, et al. Quality of large language model responses to radiation oncology patient care questions. *JAMA Netw Open*. 2024;7(4):e244630. doi:10.1001/jamanetworkopen.2024.4630

36. Odabashian R, Bastin D, Jones G, et al. Assessment of ChatGPT-3.5's knowledge in oncology: comparative study with ASCO-SEP benchmarks. *JMIR AI*. 2024;3(1):e50442. doi:10.2196/50442

37. Hoffmann J, Borgeaud S, Mensch A, et al. Training compute-optimal large language models. *arXiv*. Preprint posted online March 29, 2022. doi:10.48550/arXiv.2203.15556

38. Liu J, Liu A, Lu X, et al. Generated knowledge prompting for commonsense reasoning. *arXiv*. Preprint posted online October 15, 2021. doi:10.18653/v1/2022.acl-long.225

39. Jiang LY, Liu XC, Nejatian NP, et al. Health system-scale language models are all-purpose prediction engines. *Nature*. 2023;619(7969):357-362. doi:10.1038/s41586-023-06160-y

40. Garcia P, Ma SP, Shah S, et al. Artificial intelligence-generated draft replies to patient inbox messages. *JAMA Netw Open*. 2024;7(3):e243201. doi:10.1001/jamanetworkopen.2024.3201

41. Giorgi J, Toma A, Xie R, et al. WangLab at MEDIQA-Chat 2023: clinical note generation from doctor-patient conversations using large language models. *arXiv*. Preprint posted online May 3, 2023. doi:10.48550/arXiv.2305.02220

42. Umeton R, Kwok A, Maurya R, et al GPT-4 in a cancer center—institute-wide deployment challenges and lessons learned. *NEJM AI*. 2024;1(4). doi:10.1056/AIcs2300191

43. Zakka C, Shad R, Chaurasia A, et al. Almanac—retrieval-augmented language models for clinical medicine. *NEJM AI*. 2024;1(2). doi:10.1056/AIoa2300068

44. Bai Y, Kadavath S, Kundu S, et al. Constitutional AI: harmlessness from AI feedback. *arXiv*. Preprint posted online December 15, 2022. doi:10.48550/arXiv.2212.08073

45. Han T, Kumar A, Agarwal C, Lakkaraju H. Towards safe and aligned large language models for medicine. *arXiv*. Preprint posted online March 6, 2024. doi:10.48550/arXiv.2403.03744

46. Goldberg CB, Adams L, Blumenthal D, et al. To do no harm—and the most good—with AI in health care. *NEJM AI*. 2024;1(3):623-627. doi:10.1038/s41591-024-02853-7

47. World Health Organization. WHO calls for safe and ethical AI for health. May 16, 2023. Accessed April 14, 2024. https://www.who.int/news/item/16-05-2023-who-calls-for-safe-and-ethical-ai-for-health

48. Chen L, Zaharia M, Zou J. How is ChatGPT's behavior changing over time? *arXiv*. Preprint posted online July 18, 2023. doi:10.48550/arXiv.2307.09009

49. Haupt CE, Marks M. AI-generated medical advice-GPT and beyond. *JAMA*. 2023;329(16):1349-1350. doi:10.1001/jama.2023.5321

50. Shah NH, Entwistle D, Pfeffer MA. Creation and adoption of large language models in medicine. *JAMA*. 2023; 330(9):866-869. doi:10.1001/jama.2023.14217

SUPPLEMENT 1.
eMethods.
eReferences.
eTable 1. Classification of Errors in Explanations Accompanying Incorrect Multiple-Choice Answers by Large Language Model
eFigure. Percentage of Correct Multiple-Choice Answers on Medical Oncology Examinations Across Large Language Models

SUPPLEMENT 2.
eTable 2.

SUPPLEMENT 3.
Data Sharing Statement