HHS Public Access

Author manuscript

NEJM AI. Author manuscript; available in PMC 2024 August 09.

Published in final edited form as:

NEJM AI. 2024 May ; 1(5): . doi:10.1056/aioa2300151.

Comparative Evaluation of LLMs in Clinical Oncology

Nicholas R. Rydzewski, M.D.[1], **Deepak Dinakaran, M.D., Ph.D.**[1,2,3], **Shuang G. Zhao, M.D.**[4,5], **Eytan Ruppin, M.D., Ph.D.**[6], **Baris Turkbey, M.D.**[7], **Deborah E. Citrin, M.D.**[1], **Krishnan R. Patel, M.D.**[1]

[1]Radiation Oncology Branch, National Cancer Institute, National Institutes of Health, Bethesda, MD

[2]Physical Sciences Platform, Sunnybrook Research Institute, Toronto, ON, Canada

[3]Department of Radiation Oncology, Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada

[4]Department of Human Oncology, University of Wisconsin, Madison, WI

[5]William S. Middleton Memorial Veterans Hospital, Madison, WI

[6]Cancer Data Science Laboratory, National Cancer Institute, National Institutes of Health, Bethesda, MD

[7]Molecular Imaging Branch, National Cancer Institute, National Institutes of Health, Bethesda, MD

Abstract

BACKGROUND—As artificial intelligence (AI) tools become widely accessible, more patients and medical professionals will turn to them for medical information. Large language models (LLMs), a subset of AI, excel in natural language processing tasks and hold considerable promise for clinical use. Fields such as oncology, in which clinical decisions are highly dependent on a continuous influx of new clinical trial data and evolving guidelines, stand to gain immensely from such advancements. It is therefore of critical importance to benchmark these models and describe their performance characteristics to guide their safe application to clinical oncology. Accordingly, the primary objectives of this work were to conduct comprehensive evaluations of LLMs in the field of oncology and to identify and characterize strategies that medical professionals can use to bolster their confidence in a model's response.

METHODS—This study tested five publicly available LLMs (LLaMA 1, PaLM 2, Claude-v1, generative pretrained transformer 3.5 [GPT-3.5], and GPT-4) on a comprehensive battery of 2044 oncology questions, including topics from medical oncology, surgical oncology, radiation oncology, medical statistics, medical physics, and cancer biology. Model prompts were presented independently of each other, and each prompt was repeated three times to assess output consistency. For each response, models were instructed to provide a self-appraised confidence score (from 1 to 4). Model performance was also evaluated against a novel validation set

Dr. Rydzewski can be contacted at nickryd@gmail.com or at National Cancer Institute, 10 Center Drive, Building 10 – Hatfield CRC, Bethesda, MD 20892, Dr. Patel can be contacted at Krishnan.patel@nih.gov or at National Cancer Institute, 10 Center Drive, Building 10 – Hatfield CRC, Bethesda, MD 20892.

Author disclosures and other supplementary materials are available at ai.nejm.org.

comprising 50 oncology questions curated to eliminate any risk of overlap with the data used to train the LLMs.

RESULTS—There was significant heterogeneity in performance between models (analysis of variance, P<0.001). Relative to a human benchmark (2013 and 2014 examination results), GPT-4 was the only model to perform above the 50th percentile. Overall, model performance varied as a function of subject area across all models, with worse performance observed in clinical oncology subcategories compared with foundational topics (medical statistics, medical physics, and cancer biology). Within the clinical oncology subdomain, worse performance was observed in female-predominant malignancies. A combination of model selection, prompt repetition, and confidence self-appraisal allowed for the identification of high-performing subgroups of questions with observed accuracies of 81.7 and 81.1% in the Claude-v1 and GPT-4 models, respectively. Evaluation of the novel validation question set produced similar trends in model performance while also highlighting improved performance in newer, centrally hosted models (GPT-4 Turbo and Gemini 1.0 Ultra) and local models (Mixtral 8×7B and LLaMA 2).

CONCLUSIONS—Of the models tested on a standardized set of oncology questions, GPT-4 was observed to have the highest performance. Although this performance is impressive, all LLMs continue to have clinically significant error rates, including examples of overconfidence and consistent inaccuracies. Given the enthusiasm to integrate these new implementations of AI into clinical practice, continued standardized evaluations of the strengths and limitations of these products will be critical to guide both patients and medical professionals. (Funded by the National Institutes of Health Clinical Center for Research and the Intramural Research Program of the National Institutes of Health; Z99 CA999999.)

Introduction

Recently, several artificial intelligence (AI)-based tools have become available to the general public that will have significant implications for health care and medical practice. Many medical professionals have casually interacted with large language models (LLMs) such as ChatGPT,[1] Bard/Gemini,[2] and Claude,[3] and some have begun to use these models as augmented search engines to serve as reference tools for complex medical information. Although models capable of generating convincing and coherent text have existed for years,[4] these prior models have fallen short of matching the depth of human cognition. With recent advances, LLMs have begun to exhibit emergent properties that appear to replicate human-level intelligence.[5] Remarkably, LLMs have displayed high performance on subspeciality medical examinations, including those in plastic surgery.[6] otolaryngology,[7] ophthalmology,[8] dermatology,[9] and neurosurgery.[10] They have also shown the ability to pass the bar examination[11] and the United States Medical Licensing Examination.[12,13]

Although LLMs promise to fulfill a growing need in medicine to index, incorporate, and synthesize the ever-growing volume of information, their utility for this application in clinical oncology remains unexplored. The aim of the present study, therefore, was to evaluate the current state-of-the-art LLMs (generative pretrained transformer 3.5 [GPT-3.5], GPT-4, PaLM 2, Claude-v1, and LLaMA 1) on more than 2000 oncology questions to assess the models' accuracy, self-appraised confidence, and consistency of response across independent replicates, representing the most comprehensive head-to-head comparison of

LLMs across any medical specialty to date. We also developed a novel benchmarking dataset comprising 50 oncology questions designed to evaluate model performance without the risk of data leakage (i.e., the risk that a specific question and answer set was in the LLM training set). Our intention was to provide foundational information to aid medical professionals and patients in understanding the utility and limitations of LLMs in oncology and to profile strategies to improve their reliability.

Methods

Our primary aim was to compare the performance of a set of state-of-the-art LLMs on a test set of multiple choice questions in clinical oncology and foundational oncology topics (medical statistics, medical physics, and cancer biology). Our secondary aim was to evaluate strategies available to end users to improve their confidence in model output. In the primary analysis, we studied a subset of commonly used LLMs, including four centrally hosted models with nonpublic model weights (PaLM 2[2] [Google Bard], Claude-v1[3] [Anthropic], GPT-3.5[1] [OpenAI/Microsoft], and GPT-4[14] [OpenAI/Microsoft]) and a local model with public model weights (LLaMA 1[15] [Meta]). Because of the large number of prompts required for this analysis (more than 6000 per model), the user-friendly chatbot interface was not an ideal prompting environment. All prompting was conducted from April 2023 to May 2023 and performed independently and programmatically using the application programming interface, which allowed for an automated and efficient workflow. Meta's LLaMA 1 model[16] includes four versions of increasing complexity — 7 billion (B), 13B, 33B, and 65B — with the model name indicating how many parameters are in the network. LLaMA models were run locally on the computational resources of the National Institutes of Health's high-performance computing Biowulf cluster (http://hpc.nih.gov). For the secondary analysis on prompting strategies, the LLaMA model evaluation focused on the best-performing LLaMA model, LLaMA 65B.

The LLMs evaluated in this study fell into two categories: base models (LLaMA models) and models fine-tuned with reinforcement learning from human feedback (RLHF), including PaLM 2, Claude-v1, GPT-3.5, and GPT-4. Base models require examples within the prompt (few-shot learning) to achieve proper output formatting, whereas RLHF-tuned models can achieve the appropriate formatting through prompting instructions alone (zero-shot learning). For each prompt (Prompts 1 and 2 in the Supplemental Methods in Supplementary Appendix 1), the question and four answer choices were provided in addition to instructions to provide an answer (A, B, C, or D), a confidence score (1, 2, 3, or 4), and an explanation of the response. Models were instructed to deliver a confidence score, with 1 indicating a random guess and 4 indicating maximal confidence.

The testing set was composed of standardized questions in clinical oncology as well as other foundational topics in oncology, and it was sourced from the American College of Radiology in-training radiation oncology examinations[17] (2013, 2014, 2015, 2016, 2017, 2020, and 2021). Although these questions test radiation oncology trainees, they include questions that cover the breadth of clinical practice for all oncologists (medical, surgical, and radiation). Questions with multiple answers or those that included images in the question were removed. Answers were determined to be correct on the basis of a published answer

key created by expert oncologists, and unanswered queries were scored as incorrect. Each model was evaluated on 2044 unique questions, repeating every question across three independent replicates (6132 independent prompts per model). Accuracy was assessed independently for each set of 2044 questions, scoring each model replicate on the basis of the number of 2044 total queries that were answered correctly. Model performance was compared with a strategy of random guessing across 100 replicates and was contextualized using the subset of questions for which human performance was available (2013 and 2014 examinations).[18] We benchmarked the performance of the models against the performance of radiation oncology trainees using the means and standard deviation (SD) of the human performance on these examinations, allowing for the identification of a percentile performance for each LLM (reported as an average percentile performance across three replicates).

A subset of 1168 questions with available subject labels was used to describe the variation of model performance according to subject, including foundational topics (cancer biology, medical physics, and medical statistics) and clinical oncology (sarcoma; breast; central nervous system; gastrointestinal; genitourinary; gynecology; head, neck, and skin; lung; lymphoma/leukemia; and pediatrics). As part of the secondary aim of this study (i.e., evaluating strategies available to end users to improve their confidence in model output), LLM self-assessed confidence, consistent responses across question replicates, and the combination of these strategies were used to identify a subset of model responses in which accuracy was highest.

Although this question bank is not widely available online in its entirety, there remains a tangible risk of data leakage into the LLM training sets. Furthermore, this question bank is not open source, and thus we are unable to make these questions, prompts, and output available for future analysis and benchmarking. To address these limitations, a novel clinical oncology validation set was constructed that consisted of 50 questions; this set was created and reviewed by three oncologists and benchmarked on an expanded set of up-to-date models available at the time of this subsequent validation effort (February 2024). Although PaLM 2 was no longer available for medical benchmarking, this expanded list included all other models in the primary analysis in addition to Google’s most state-of-the-art models (Gemini 1.0 Pro and Gemini 1.0 Ultra), GPT-4 Turbo, Claude-v2, Mistral[19,20] models (both centrally hosted and local), and LLaMA 2.[21] Each model was evaluated on the validation question set over 10 independent replicates.

All 50 questions, answer choices, answer keys, and model prompts (zero-shot learning and few-shot learning prompts) are provided (Supplementary Appendix 2) along with model output corresponding to the zero-shot learning prompts (Supplementary Appendix 3). Further details on statistical methods used in this study, local computational requirements (Table S1), prompt engineering (Fig. S1A), and impacts of fine-tuning (including medical fine-tuning; Fig. S1B)[22,23] are reported in the Supplemental Materials.

Results

OVERALL MODEL PERFORMANCE ON STANDARDIZED ONCOLOGY QUESTIONS

The accuracy of each model was assessed for three independent replicates (Fig. 1) across 2044 questions. For each model across the three independent replicates, we observed a mean accuracy (minimum to maximum) of 25.6% (25.4 to 25.7%) for LLaMA 7B, 27.8% (26.5 to 28.6%) for LLaMA 13B, 34.3% (33.4 to 35.0%) for LLaMA 33B, 38.5% (38.0 to 38.8%) for LLaMA 65B, 45.1% (43.9 to 45.9%) for PaLM 2, 51.8% (51.4 to 52.5%) for GPT-3.5, 55.3% (54.5 to 55.7%) for Claude-v1, and 68.7% (68.6 to 68.8%) for GPT-4. For comparison, the random guess strategy had a mean accuracy of 25.2% across 100 replicates. Heterogeneity in performance between models was observed (analysis of variance, $P<0.001$). Pairwise comparisons revealed a difference between all models (Tukey honestly significant difference, $P_{adjusted}<0.01$), with the exception of LLaMA 7B compared with LLaMA 13B ($P_{adjusted}=0.08$) and LLaMA 7B compared with random guesses ($P_{adjusted}=1.00$).

Within model class, accuracy improved monotonically as the number of parameters increased (as observed in the LLaMA and GPT model classes); however, this monotonic relationship was violated in comparisons between models, as PaLM 2 has 340B parameters, GPT-3.5 has 135B, and Claude-v1 has 52B.

MODEL PERFORMANCE RELATIVE TO TRAINEE PERFORMANCE

In a prior publication,[18] the distribution of the human performance on the 2013 (mean, 61.9%; SD, 8.2%) and 2014 (mean, 57.2%; SD, 7.6%) examinations was described. Using these data, LLaMA 65B, PaLM 2, GPT-3.5, Claude-v1, and GPT-4 were found to perform at the less than first, third, fifth, 16th, and 69th percentiles for the 2013 examination and the less than first, sixth, 14th, 23rd, and 89th percentiles for the 2014 examination, respectively.

VARIATION OF MODEL PERFORMANCE ACCORDING TO SUBJECT

Model performance varied as a function of subject area across all models (analysis of variance, all $P<0.001$) (Fig. 2), and there was a positive correlation between a model's overall performance and subject area–specific performance (Pearson's $r=0.630$; $P<0.001$). On the basis of this observation, exploratory analyses were conducted to evaluate model performance according to question subject. Except for LLaMA 65B (Student's t-test, $P=0.27$), all other LLMs exhibited higher performance on the group of foundational topics (medical statistics, medical physics, and cancer biology) compared with clinical oncology subcategories (all, $P\leq0.02$). The worst-performing clinical subjects were female-predominant malignancies (breast and gynecologic) compared with other malignancies ($P=0.11$ for LLaMA 65B, $P<0.01$ for all other models).

SELF-ASSESSED CONFIDENCE

On the basis of the observation that the error rate of each LLM ranged from 31.3% (GPT-4) to 61.4% (LLaMA 65B), a set of analyses was constructed to determine strategies that could help end users identify subgroups of queries with lower error rates, beginning with LLM self-assessed confidence. To this end, accuracy as a function of self-appraised confidence

was evaluated (Fig. 3), and all models were observed to have self-appraised confidence scores with discriminatory capability (chi-square test, P<0.001) except for LLaMA 65B (P=0.99). Furthermore, all models with discriminatory power had an improved accuracy when reporting maximal self-assessed confidence (a score of 4 of 4) compared with overall accuracy (test of proportions, all P≤0.004). Finally, all models trended toward high self-appraised confidence, with more than 94% of responses returning a confidence score of 3 or 4.

PROMPT REPETITION

The technique of prompt repetition (i.e., independently repeating the same query multiple times) was investigated as another means by which end users could appraise their confidence in the accuracy of a model's response (Fig. 4A). A significant proportion of cases failed to produce the same response for all three replicates: 75.5% for LLaMA 65B, 53.2% for PaLM 2, 26.4% for GPT-3.5, 38.0% for Claude-v1, and 16.3% for GPT-4. In general, as model performance increased, the proportion of queries for which a model provided a consistent correct answer (i.e., three of three correct) increased (Pearson's r=0.988; P=0.002). Compared with the overall model accuracy, a higher accuracy (test of proportions, all P<0.001) was noted for the subset of responses in which there was consistency across all replicates (i.e., triplicate agreement), with the largest increase observed in LLaMA 65B (38.5 to 64.4%). Notably, a significant proportion of queries had triplicate agreement, and they were consistently incorrect (i.e., three of three of the same incorrect answer): 9.0% for LLaMA 65B, 18.2% for PaLM 2, 30.7% for GPT-3.5, 19.6% for Claude-v1, and 21.0% for GPT-4, representing a "fixed, false belief." For context, random guesses on this task (across 100 simulations) would result in an average rate of 4.7% (minimum, 3.8%; maximum, 5.8%) of consistently incorrect responses.

COMBINING STRATEGIES FOR IDENTIFYING MORE RELIABLE RESPONSES

We next evaluated the utility of the combination of these factors (model selection, self-assessed confidence, and output consistency) for identifying cases in which model output was relatively more reliable, defined here as higher subgroup accuracy (Fig. 4B). The combination of these factors allowed for the identification of high-performing subgroups of queries in Claude-v1 and GPT-4, with observed accuracies of 81.7% (n=383) and 81.1% (n=1306), respectively. A significant interaction between the self-assessed confidence score and triplicate agreement was observed in all models (all interaction coefficients positive, P≤0.03) with the exception of LLaMA 65B ($P_{interaction}$=0.956), characterizing the increase in accuracy at the intersection of high confidence and triplicate agreement.

VALIDATION OF PERFORMANCE WITH A NOVEL QUESTION SET

With our newly developed oncology question set, we evaluated an updated range of models across 10 replicates (Fig. 5), focusing on models that could be prompted with zero-shot learning. Heterogeneity was again observed in model responses, including in the model explanations, which have been provided (outputs from the first replicate for each model are in Supplementary Appendix 3). Altogether, these validation results confirm the previously discussed trends of performance across models. The newest centrally hosted models evaluated in this effort, GPT-4 Turbo and Gemini 1.0 Ultra, showed the highest

performance, with median accuracies of 80 and 79%, respectively. We also observed considerable improvement in local models, with Mixtral 8×7B Instruct exhibiting better performance than many centrally hosted models. The updated LLaMA 2 models also provided an opportunity to compare the differential impacts of prompt structure and fine-tuning on performance (Fig. S1).

Discussion

OVERALL FINDINGS

The current investigation offers the most comprehensive description and head-to-head comparison yet of modern AI-based LLMs in oncology. The performances of Meta's LLaMA 1, Google's Bard/PaLM 2, Anthropic's Claude-v1, and OpenAI's GPT-3.5 and GPT-4 were evaluated on a standardized, comprehensive battery of oncology questions to benchmark these models for medical professionals and patients. A wide range of performance was observed on the tested metrics of accuracy, self-assessed confidence, and consistency of response across models. GPT-4 consistently outperformed other models, achieving the highest overall accuracy of 69%. GPT-4 was also the only model to perform above the 50th percentile of oncology trainees on the subset of questions for which human performance was available.[18] On validation with a novel question set, we confirm these general trends, highlighting the state-of-the-art performance with GPT-4 Turbo and Gemini 1.0 Ultra. This analysis also evaluated local models, which can operate on local network systems, including those in hospitals, thereby improving security and facilitating more effective management of protected health information. Although no local model was competitive with the top-performing models, Mixtral 8×7B did exhibit an impressive performance compared with many centrally hosted models, highlighting the rapid evolution of local models that can handle protected health information. LLMs are augmented neural networks,[24] primarily designed to predict the next word in a string of text.[4,25] The output of these models may be further refined through RLHF, as was true for four of the five models under examination in the primary analysis (PaLM 2, Claude-v1, GPT-3.5, and GPT-4). Effective prompt engineering also plays a crucial role in guiding these models to generate more targeted and relevant outputs. Surprisingly, from this narrow focus, models can produce output that appears to replicate aspects of human intelligence[5] and medical expertise[6–10,12,13] as illustrated by the performance of the current state-of-the-art LLMs on oncology questions. However, our observation of persistently high error rates, even with end-user strategies meant to isolate high-performing subgroups, reveals the inherent limitations of LLMs in their current form to guide clinical practice.

TRAINED BIAS IN LLMs

Substantial differences in performance were observed according to subject category. The LLMs were trained through self-supervised learning on a database primarily sourced from the Internet. Except for LLaMA 65B, all models in the primary analysis exhibited superior accuracy on foundational topics (cancer biology, medical physics, and medical statistics) compared with clinical oncology. Within clinical oncology, these models showed inferior performance on female-predominant malignancies compared with all other malignancies. The consistency of subgroup performance across models suggests a common origin for this

inaccuracy, and one possible explanation could relate to medically inaccurate information being part of the training set. This is compatible with prior studies characterizing substantial rates of misinformation in these subdomains on the Internet,[26,27] including a recent study showing that one third of Internet-based oncology articles contain misinformation.[28]

STRATEGIES TO IMPROVE RELIABILITY OF LLMs

Because even the best-performing models continued to exhibit a significant error rate (31.3% for GPT-4 in the primary analysis), our final objective was to describe methods by which medical professionals and patients might optimize their confidence in these AI systems. We attempted to explore two strategies available to end users to evaluate the accuracy of an AI's response: LLM self-assessed confidence and consistency of response across replicates. Although either of these strategies can help identify higher-accuracy subgroups, the combination of maximal self-appraised confidence and triplicate agreement can identify subgroups in which accuracy exceeds 80% for some models.

During this exploration, two notable subsets of inaccurate responses were observed. The first consisted of incorrect responses delivered with high confidence, which we observed in 27% (GPT-4) to 52% (PaLM 2) of responses among the models with meaningful confidence self-appraisal. The second comprised queries to which models consistently delivered the same incorrect answer in triplicate, representing a "fixed, false belief," which was observed in 9% (LLaMA 65B) to 21% (GPT-4) of responses.

The existence of these phenomena identifies two possible pitfalls in the application of LLMs in clinical practice. First, these models almost always (>94%) exhibit high confidence disproportionate to their accuracy, consistent with the tendency for these models to present confabulated information with high confidence. Second, the existence of these "fixed, false beliefs" likely represents trained bias encoded into these models, further highlighted by the consistency across models in their poor performance on female-predominant malignancies. Although these two pitfalls are not mutually exclusive, they represent independent limitations related to internal model behavior (overconfidence) and training bias (consistent inaccuracies). These may require independent solutions, including the implementation of retrieval-augmented generation systems, which use outside databases of ground truths that can be supplied to models through prompt engineering.

CLINICAL UTILITY OF LLMs

Validation against multiple choice examinations[6–11] replicates one aspect of how clinical competency is evaluated in medical practice, but the scope of such assessment is narrow in determining whether these models are safe for clinical implementation. Clinicians are assessed in clinical settings extensively as part of their certification process, and, similarly, rigorous validation of LLMs in such settings will be essential to guarantee safety. We have thus far highlighted studies and data pertinent to this first aspect of clinical competency (examination knowledge), but evaluation of performance in clinical settings is equally imperative. There is a growing body of literature documenting the clinical utility of these models for tasks of diagnosis,[29] management,[30] and patient counseling[31] in medicine. Benary et al.[32] evaluated the use of LLMs as tools in oncology to personalize treatment

decisions and noted that although these models in their current form fail to achieve a level of performance shown by human experts, they still provided valuable recommendations that could complement established care. Although this growing body of literature provides promising examples for clinical utilization, widespread implementation will first require further validation of these tools in clinical settings, with appropriate safety monitoring.

Limitations

Although this study aimed to benchmark the current state-of-the-art AI models in oncology, the pace of innovation in the field will inevitably limit the generalizability of these results to future models. Our primary benchmark task used standardized questions, designed by experts on a broad range of oncology topics; however, it may still fail to represent the true complexity and ambiguity of clinical practice or capture the changes to practice over time. Although generalizing these findings to all of oncology may be limited because this benchmarking used examinations given to radiation oncologists, the subject material does include questions relevant to all oncologists. Furthermore, although this question bank is an imperfect evaluation of proficiency relevant to clinical practice, its use in this task attempts to replicate efforts to determine the proficiency of human oncology trainees. Therefore, any critique of the utility of this dataset as a benchmark for assessing LLM competency can be equally used to highlight the current limitations of such examinations to evaluate the competency of oncology trainees. These methodologic limitations are primarily driven by the lack of available, well-designed benchmarking tasks for this purpose, underscoring the critical need for collaboration with the oncology and medical communities to design standardized, consensus-based benchmarks to evaluate the proficiency of LLMs for clinical use.

In addition, although our methodology was designed to mitigate any bias in performance as a result of prompt engineering (Supplementary Appendix 1), recent work has shown that even small variations in a prompt can have a large impact on an LLM response.[33] Base models (LLaMA 1, LLaMA 2 Base, and Mistral non-Instruct models) require a different prompting style (few-shot learning) than the comparators in this study (zero-shot learning), which can affect head-to-head comparisons of models. Utilization of the validation set on both base models and chat/instruct models for LLaMA 2 and Mistral allowed for a better understanding of how these differences can affect model performance (Fig. S1A). Future work will be required to identify the best way to prompt LLMs to optimize accuracy.

Finally, intrinsic to the nature of this research is a concern regarding the possible influence of data leakage, and to address this limitation, we conducted a validation study on a set of previously unseen questions. This concern was mitigated by the consistency in the trend in model performance observed on our novel validation set and in the primary analysis, reinforcing the robustness of our main conclusions.

Conclusions

LLMs currently available to medical professionals and patients exhibit a wide range of performance on clinical oncology questions. Some models appear to perform no better than random chance, whereas others may achieve a level of accuracy competitive with resident

physicians.[18] Strategies such as confidence self-appraisal and prompt repetition show some promise in identifying subgroups of responses more likely to be correct, but their utility is currently limited.

Furthermore, we report observations consistent with overconfidence and "fixed, false beliefs" across models, which may limit their clinical utility as trustworthy tools. In this regard, one particularly noteworthy finding was the poor performance observed on prompts relating to female-predominant malignancies across models. This likely represents a trained bias that cannot be easily mitigated with prompt repetition or LLM self-assessed confidence. Such training biases emphasize the need for partnerships between developers and medical professionals to curate reliable training data. Given the enthusiasm to integrate LLMs into clinical practice, continued standardized evaluations of the strengths and limitations of these models will be critical to guide both patients and medical professionals in identifying appropriate use cases and building appropriate expectations for model performance in this clinical application.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funded by the National Institutes of Health Clinical Center for Research and the Intramural Research Program of the National Institutes of Health (Z99 CA999999).

References

1. OpenAI. Introducing ChatGPT. November 30, 2022 (https://openai.com/blog/chatgpt).
2. Anil R, Dai AM, Firat O, et al. PaLM 2 technical report. September 13, 2023 (https://arxiv.org/abs/2305.10403). Preprint.
3. Bai Y, Jones A, Ndousse K, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. April 12, 2022 (https://arxiv.org/abs/2204.05862). Preprint.
4. Zhao WX, Zhou K, Li J, et al. A survey of large language models. November 24, 2023 (https://arxiv.org/abs/2303.18223). Preprint.
5. Bubeck S, Chandrasekaran V, Eldan R, et al. Sparks of artificial general intelligence: early experiments with GPT-4. April 13, 2023 (https://arxiv.org/abs/2303.12712). Preprint.
6. Gupta R, Herzog I, Park JB, et al. Performance of ChatGPT on the plastic surgery inservice training examination. Aesthet Surg J 2023;43:NP1078–NP1082. DOI: 10.1093/asj/sjad128. [PubMed: 37128784]
7. 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;280:4271–4278. DOI: 10.1007/s00405-023-08051-4. [PubMed: 37285018]
8. Mihalache A, Popovic MM, Muni RH. Performance of an artificial intelligence chatbot in ophthalmic knowledge assessment. JAMA Ophthalmol 2023;141:589–597. DOI: 10.1001/jamaophthalmol.2023.1144. [PubMed: 37103928]
9. Ravipati A, Pradeep T, Elman SA. The role of artificial intelligence in dermatology: the promising but limited accuracy of ChatGPT in diagnosing clinical scenarios. Int J Dermatol 2023;62:e547–e548. DOI: 10.1111/ijd.16746. [PubMed: 37306147]

10. Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT, GPT-4, and Google Bard on a neurosurgery oral boards preparation question bank. Neurosurgery 2023;93:1090–1098. DOI: 10.1227/neu.0000000000002551. [PubMed: 37306460]
11. Katz DM, Bommarito MJ, Gao S, Arredondo P. GPT-4 passes the bar exam. March 15, 2023 (10.2139/ssrn.4389233). Preprint.
12. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. April 12, 2023 (https://arxiv.org/abs/2303.13375). Preprint.
13. 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. DOI: 10.1371/journal.pdig.0000198. [PubMed: 36812645]
14. OpenAI. GPT-4 technical report. March 4, 2024 (https://arxiv.org/abs/2303.08774). Preprint.
15. Touvron H, Lavril T, Izacard G, et al. LLaMA: open and efficient foundation language models. February 27, 2023 (https://arxiv.org/abs/2302.13971). Preprint.
16. Meta. Introducing LLaMA: a foundational, 65-billion-parameter large language model. February 24, 2023 (https://ai.facebook.com/blog/large-language-model-llama-meta-ai/).
17. American College of Radiology. Radiation oncology in-training exam (TXIT). 2023 (https://www.acr.org/Lifelong-Learning-and-CME/Learning-Activities/In-Training-Exams/Radiation-Oncology-In-Training-Exam).
18. Hatch SS, Vapiwala N, Rosenthal SA, et al. Radiation oncology resident in-training examination. Int J Radiat Oncol Biol Phys 2015;92:532–535. DOI: 10.1016/j.ijrobp.2015.02.038. [PubMed: 26068487]
19. Jiang AQ, Sablayrolles A, Mensch A, et al. Mistral 7B. October 10, 2023 (https://arxiv.org/abs/2310.06825). Preprint.
20. Jiang AQ, Sablayrolles A, Roux A, et al. Mixtral of experts. January 8, 2023 (https://arxiv.org/abs/2401.04088). Preprint.
21. Touvron H, Martin L, Stone K, et al. LLaMA 2: open foundation and fine-tuned chat models. July 19, 2023 (https://arxiv.org/abs/2307.09288). Preprint.
22. Chen Z, Cano AH, Romanou A, et al. MEDITRON-70b: scaling medical pretraining for large language models. November 27, 2023 (https://arxiv.org/abs/2311.16079). Preprint.
23. Christophe C, Gupta A, Hayat N, et al. Med42 — a clinical large language model. HuggingFace. October 9, 2023 (https://huggingface.co/m42-health/med42-70b).
24. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Adv Neural Inf Process Syst 2017;30:6000–6010.
25. Wolfram S What is ChatGPT doing … and why does it work? Champaign, IL: Wolfram Media, 2023.
26. Chen L, Wang X, Peng TQ. Nature and diffusion of gynecologic cancer-related misinformation on social media: analysis of tweets. J Med Internet Res 2018;20:e11515. DOI: 10.2196/11515. [PubMed: 30327289]
27. Wilner T, Holton A. Breast cancer prevention and treatment: misinformation on Pinterest, 2018. Am J Public Health 2020;110(Suppl 3):S300–S304. DOI: 10.2105/AJPH.2020.305812. [PubMed: 33001732]
28. Johnson SB, Parsons M, Dorff T, et al. Cancer misinformation and harmful information on facebook and other social media: a brief report. J Natl Cancer Inst 2022;114:1036–1039. DOI: 10.1093/jnci/djab141. [PubMed: 34291289]
29. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. JAMA 2023;330:78–80. DOI: 10.1001/jama.2023.8288. [PubMed: 37318797]
30. Truhn D, Weber CD, Braun BJ, et al. A pilot study on the efficacy of GPT-4 in providing orthopedic treatment recommendations from MRI reports [published correction appears in Sci Rep 2024;14:5431]. Sci Rep 2023;13:20159. DOI: 10.1038/s41598-023-47500-2. [PubMed: 37978240]
31. Adams LC, Truhn D, Busch F, et al. Leveraging GPT-4 for post hoc transformation of free-text radiology reports into structured reporting: a multilingual feasibility study. Radiology 2023;307:e230725. DOI: 10.1148/radiol.230725. [PubMed: 37014240]

32. Benary M, Wang XD, Schmidt M, et al. Leveraging large language models for decision support in personalized oncology. JAMA Netw Open 2023;6:e2343689. DOI: 10.1001/jamanetworkopen.2023.43689. [PubMed: 37976064]

33. Ullman TD. Large language models fail on trivial alterations to theory-of-mind tasks. March 14, 2023 (https://arxiv.org/abs/2302.08399). Preprint.

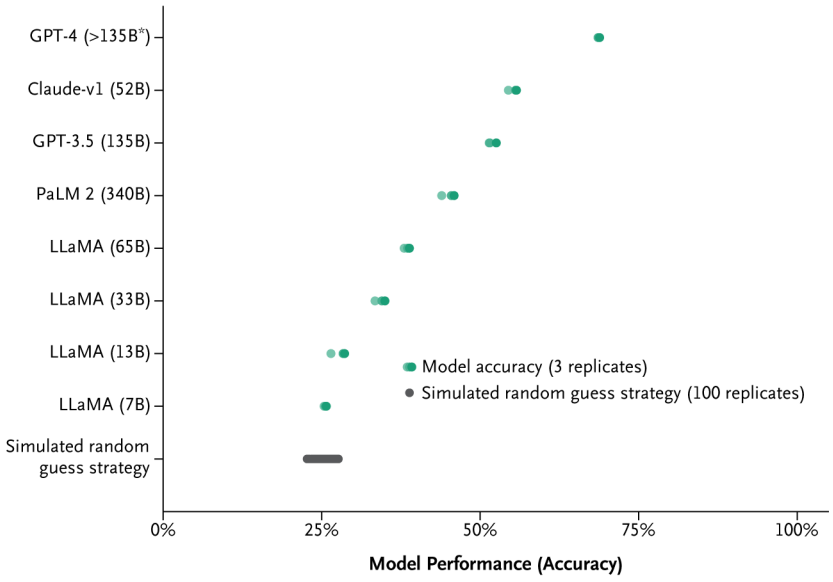

Figure 1. Overall Model Performance on Standardized Oncology Questions.

All models were evaluated on 2044 oncology questions, with each point representing one of three independent replicates on the full question set. Models were benchmarked against 100 replicates of random guesses. The number of parameters in billions (B) for each model is listed in parentheses. *Of note, the number of parameters for generative pretrained transformer 4 (GPT-4) is not published, but it is likely more than 135B (GPT-3.5) and estimated by some to be on the order of 1000B.

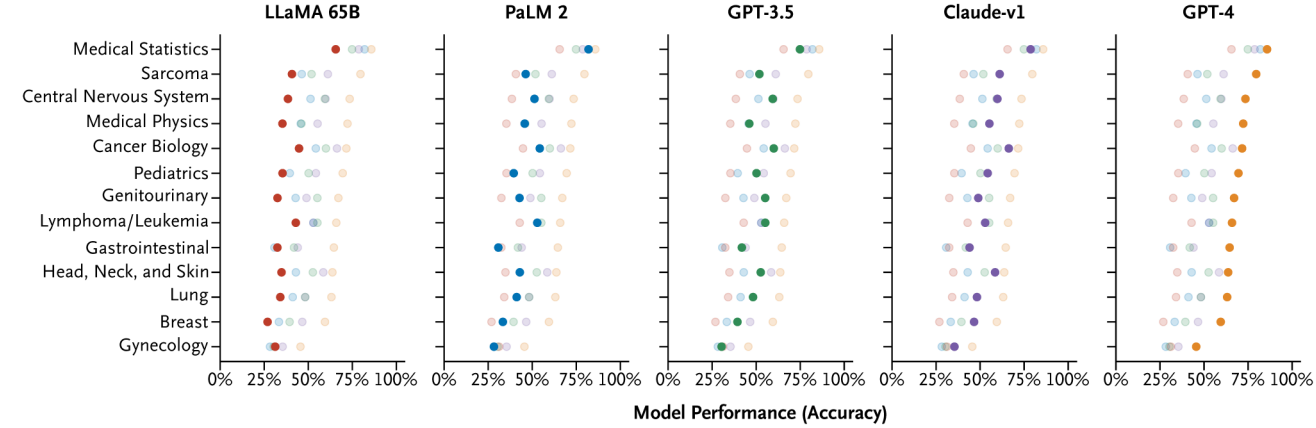

Figure 2. Variation of Model Performance According to Subject.

Performance stratified according to subject was evaluated for each model in the subset of questions with subject domain labels (n=1168). Overall, there is a correlation between global model performance and performance according to subject domain. Models were found to have better performance on queries in foundational concepts in oncology (cancer biology, medical physics, and medical statistics) than those pertaining to clinical oncology. Of clinical oncology queries, models exhibited worse performance in the domain of female-predominant malignancies (breast and gynecologic origin) than the remainder of other clinical oncology inquiries. B denotes billion; and GPT, generative pretrained transformer.

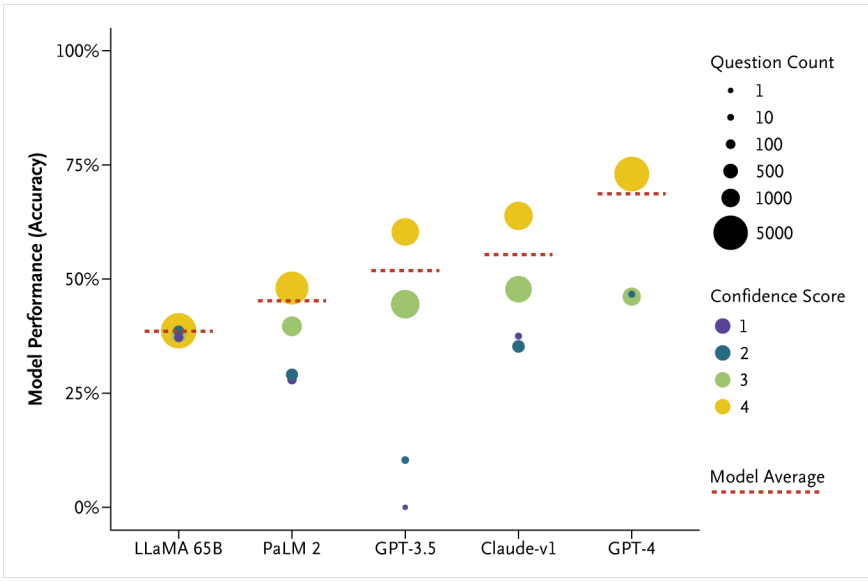

Figure 3. Self-Assessed Confidence Has Discriminatory Power in High-Performing Models.
In each question prompt, the models were asked to evaluate their confidence (from 1 to 4) in the response, in which 1 represented minimal confidence (i.e., a random guess) and 4 represented maximal confidence. The self-assessed confidence score had discriminatory power for PaLM 2, generative pretrained transformer 3.5 (GPT-3.5), Claude-v1, and GPT-4. B denotes billion.

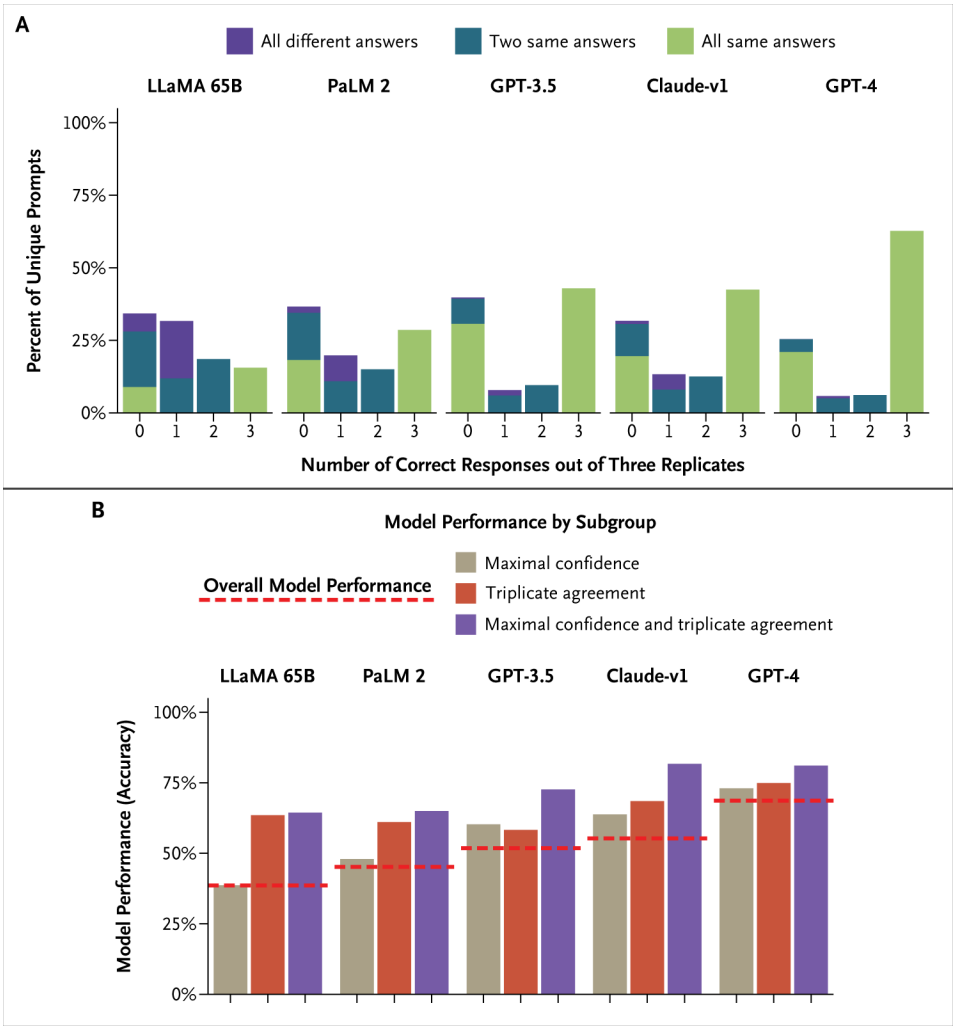

Figure 4. Artificial Intelligence Consistency and Self-Appraised Confidence Reporting Identifies Output with Higher Reliability.

For each question, the model was prompted to identify a correct answer and return a self-assessed confidence score over three separate replicates. The distribution of consistency in response is shown, which demonstrates a notable subset of queries to which models respond with the same incorrect answer (Panel A). The performance of subgroups of prompts for each model based on whether the model output had maximal confidence (i.e., the large language model returned a confidence score of 4 of 4 for each of the three responses), triplicate agreement (i.e., the large language model returned the same response for each of the three replicates), or both is shown (Panel B). B denotes billion; and GPT, generative pretrained transformer.

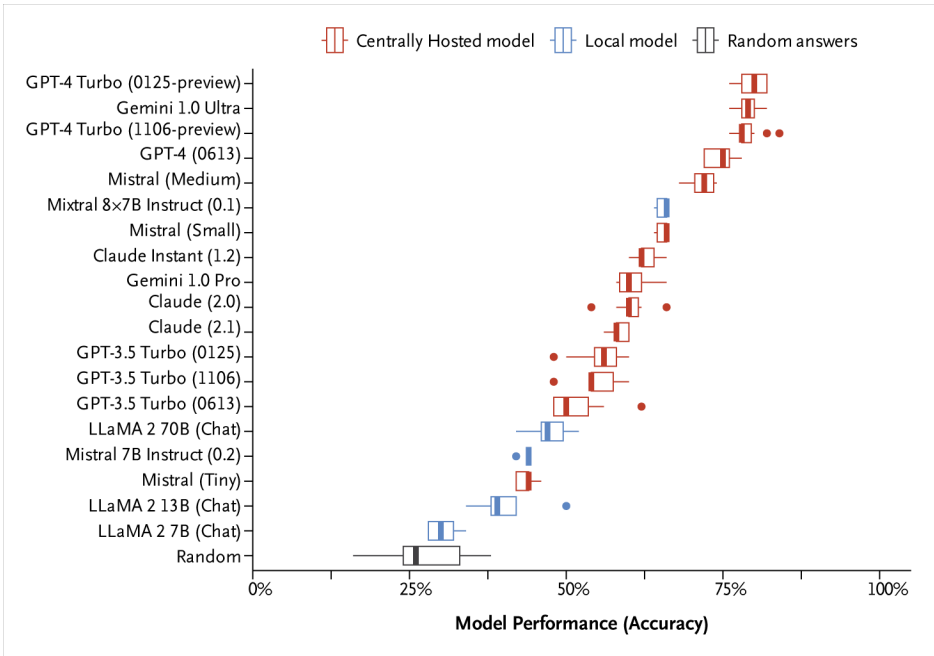

Figure 5. Model Performance on a Novel Validation Question Set.

A novel set of 50 oncology questions was developed to evaluate model performance independent of any risk of data leakage. Performance trends are consistent with the more comprehensive question set but further highlight improved performance with state-of-the-art models. Generative pretrained transformer 4 (GPT-4) Turbo and Gemini 1.0 Ultra exhibit the top performance for any centrally hosted model, whereas Mixtral 8×7B was the highest-performing locally run model. All accuracy results come from models prompted with the zero-shot learning technique (no examples in the prompts). Every box plot represents the accuracy distribution across 10 independent model prompts. B denotes billion.