Article published online: 2025-04-08

A Narrative Review on the Application of Large Language Models to Support Cancer Care and Research

Ryzen Benson[1, 2], Marianna Elia[1, 2], Benjamin Hyams[1, 2, 3], Ji Hyun Chang[1, 2, 4], Julian C. Hong[1, 2, 5]

1 Department of Radiation Oncology, University of California, San Francisco, San Francisco, California
2 Bakar Computational Health Sciences Institute, University of California, San Francisco, San Francisco, California
3 School of Medicine, University of California, San Francisco, San Francisco, California
4 Department of Radiation Oncology, Seoul National University Hospital, Seoul National University College of Medicine, Seoul, Korea
5 UCSF UC Berkeley Joint Program in Computational Precision Health (CPH), San Francisco, CA

Summary

Objectives: The emergence of large language models has resulted in a significant shift in informatics research and carries promise in clinical cancer care. Here we provide a narrative review of the recent use of large language models (LLMs) to support cancer care, prevention, and research.

Methods: We performed a search of the Scopus database for studies on the application of bidirectional encoder representations from transformers (BERT) and generative-pretrained transformer (GPT) LLMs in cancer care published between the start of 2021 and the end of 2023. We present salient and impactful papers related to each of these themes.

Results: Studies identified focused on aspects of clinical decision support (CDS), cancer education, and support for research activities. The use of LLMs for CDS primarily focused on aspects of treatment and screening planning, treatment response, and the management of adverse events. Studies using LLMs for cancer education typically focused on question-answering, assessing cancer myths and misconceptions, and text summarization and simplification. Finally, studies using LLMs to support research activities focused on scientific writing and idea generation, cohort identification and extraction, clinical data processing, and NLP-centric tasks.

Conclusions: The application of LLMs in cancer care has shown promise across a variety of diverse use cases. Future research should utilize quantitative metrics, qualitative insights, and user insights in the development and evaluation of LLM-based cancer care tools. The development of open-source LLMs for use in cancer care research and activities should also be a priority.

Keywords

large language models; clinical decision support; artificial intelligence; natural language processing; generative AI

Yearb Med Inform 2024:
http://dx.doi.org/10.1055/s-0044-1800726

1. Introduction

Cancer remains the second leading cause of death worldwide, with nearly 600,000 deaths and approximately two million new diagnoses estimated in 2024 [1]. Cancer phenotypes and treatment strategies vary across patient subpopulations, clinical settings, and socioeconomic backgrounds [2,3]. Therefore, the role of data analysis and utilization in this domain holds critical importance. However, the complex and multifaceted nature of cancer is further complicated by the time and costs frequently associated with cancer care, and research data collection and analysis approaches [4,5].

A well-established body of literature has investigated the potential utility of natural language processing (NLP) within cancer research [6–8]. NLP has historically relied on rule-based, statistical (*i.e.*, machine learning), and hybrid (*i.e.*, rule-based and statistical) approaches to carry out various tasks on textual data. Over the past several years, NLP has undergone a major increase in profile. In 2018, Google released their seminal paper on bi-directional encoder representations from transformer (BERT) models, a large language model (LLM) based on transformer architectures from deep learning capable of incorporating bi-directional contextual relationships between words in text for NLP tasks [9,10]. These models gave rise to generative pretrained transformer (GPT) models, an encoder-decoder model architecture capable of taking input text and producing novel output text [11] . GPT models first gained wider notoriety in the scientific community following the release of GPT-3, an iteration of the GPT model that was shown to demonstrate state-of-the-art performance in various NLP tasks and capable of generating coherent and accurate human-like text [12]. Consequently, an increasing body of research is using LLMs to support aspects of cancer treatment and education.

This narrative review aims to describe advances over the past three years in the application and development of LLMs within cancer care, prevention, and education. We focus on two key NLP architectures with BERT- and GPT-based LLMs. Both have attracted significant attention: BERT is open source with widespread use and adoption, and GPT represents the current state of the art in NLP tasks. Furthermore, these model architectures are easily accessible for researchers, clinical providers, and organizational support staff, promoting equitable access to their use and accompanied by extensive support and documentation. This review summarizes key use cases, recent work, and future directions.

2. Methods

We performed a search of Scopus journal articles published in the English language, excluding reviews and articles published on preprint servers. We limited our review to studies published over the past 3 years (*i.e.*, the start of 2021 through the end of 2023) to better reflect the contemporary state of LLM-based cancer research. Search keywords included ((“pretrained language model” OR “large language model” OR “llm”) OR (“generative pretrained transformer” OR gpt) OR (“bidirectional encoder representations from transformers” OR “BERT”)) AND ((oncology OR cancer)). Any additional articles of interest were retrospectively included. The resulting article titles and abstracts were screened for relevancy and kept where the primary focus was the use of LLMs within cancer care, education, and prevention activities. Four authors read the resulting articles while iteratively and inductively categorizing the articles into broad themes for qualitative analysis. Studies focusing on bioinformatics, genetics, and genomics captured in our study that were not directly related to cancer care were also excluded, along with studies in which the full text of the article could not be obtained.

3. Results

Our article selection approach yielded 51 peer-reviewed articles to be included in our analysis, along with an additional two studies of interest identified by the authors. LLM-based cancer research is in its infancy; therefore, cancer-focused LLM studies have demonstrated varying levels of study rigor and quality. When assessing the quality of LLM-based cancer research, readers should ensure that studies contain complete descriptions of the model, version, and available parameters used in the analysis, along with details on the model instance used (*e.g.*, API-based vs UI/chat interface). Additionally, since LLMs are non-deterministic in the responses they generate, studies should assess the reproducibility of their findings by replicating experiments multiple times to ensure consistency in their output. Finally, higher-quality studies frequently incorporate domain experts in the experimental design and interpretation of model output. By incorporating clinical experts and end-users in the development and evaluation of these tools, studies can focus on ensuring that model-generated output is consistent with best clinical practices, accounts for unique and diverse clinical workflows, and minimize potential health and psychological risks posed to patients with cancer.

The quality of studies included in our analysis varied and were often exploratory. Due to the lack of standardized LLM evaluation approaches in cancer research and clinical research more broadly, we observed varying levels of rigor in study design and model evaluation across studies. Studies often lacked critical details in their methods including whether API- or chat-based LLMs were used and which iterations/versions of LLMs were used. We focused our discussion on studies with well-described methods and results. Three primary themes emerged from the analysis: 1) clinical decision support, 2) patient and provider cancer education, and 3) support of cancer research activities. Figure 1 shows the distribution of primary themes and their foci across these studies.

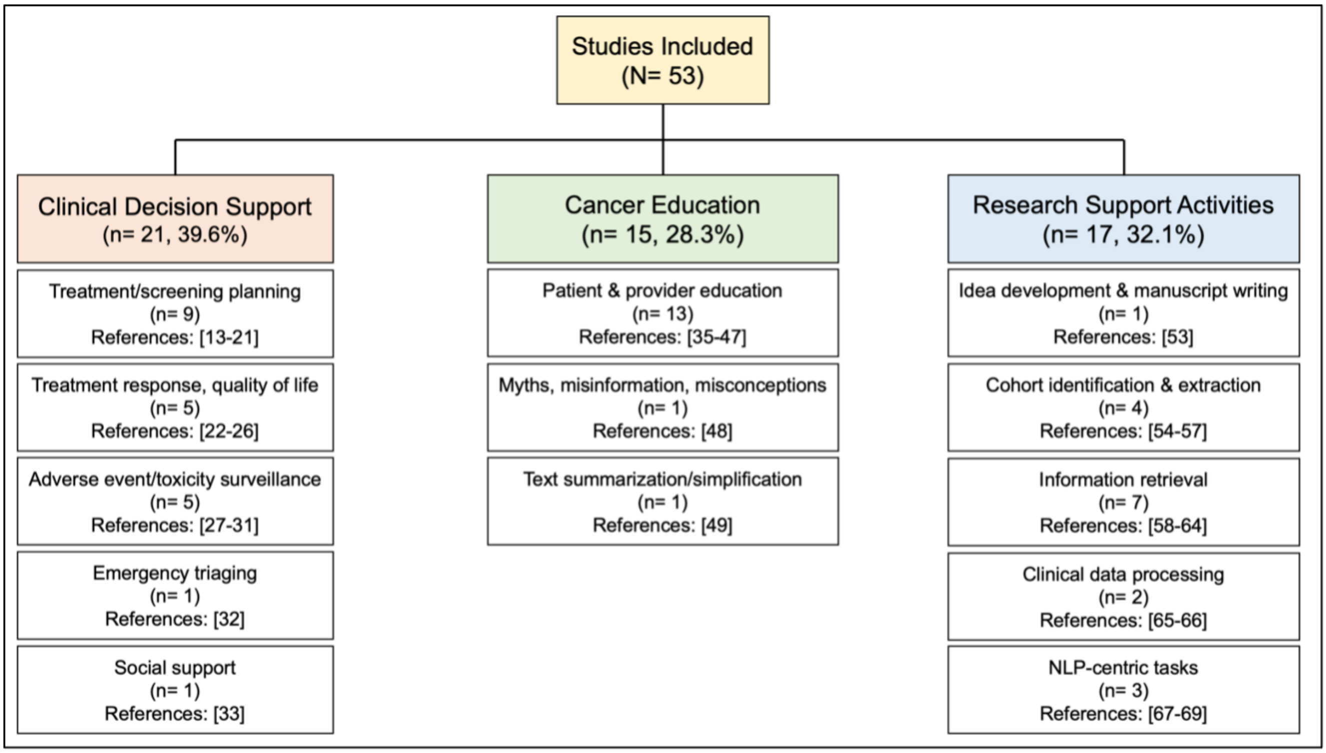

Figure 1. Broad themes and topics of articles in our narrative review.

4. Clinical Decision Support

The most frequently observed use of LLMs in cancer care focused on clinical decision support (CDS) activities. CDS can take shape in many forms but is typically characterized by providing clinical providers with information to assist in their clinical decision-making process. Selecting cancer treatment regimens that are effective and minimize adverse events is of paramount importance in cancer therapy. However, this process is nontrivial and highly dependent on numerous factors including patient cancer type, stage, genetics, and response to prior treatment.

4.1. Cancer Treatment and Screening Recommendations using LLMs

Research in cancer treatment planning has evaluated LLMs to assist with the recommendation of personalized cancer treatment regimens based on patient clinical histories, genetics, and current clinical guidelines [13–21], and has been largely limited to models using GPT architectures.

In a broad evaluation of cancer therapy recommendations, Bitterman and colleagues recently showed that when benchmarked against NCCN guidelines, GPT-3.5-turbo often provided incorrect treatment recommendations among other correct recommendations [14]. Lukac *et al.,* [13] observed similar deficiencies when they employed GPT-3.5 to make treatment recommendations for 10 breast cancer patients discussed at a multidisciplinary tumor board given various genetic components of the patient's cancer. They showed that GPT-3.5 demonstrated low coherence with tumor board recommendations for breast cancer therapies when provided with primary breast cancer cases in a standardized prompt format. Another study assessed breast cancer treatment regimen recommendations using LLMs, but augmented their prompting approach with relevant patient health data, and family history of cancer, along with patho- and immunopathological characteristics of the lesion [15]. Despite improved performance, the authors stated that their approach resulted in several instances of incorrect treatment regimen recommendations (*i.e.*, hallucinations).

However, the application of GPT-4 to recommend treatment for fictional case vignettes of various cancers resulted in has shown to improve performance while also minimizing hallucinations as compared to GPT-3.5 [16]. This difference in performance is consistent with the findings of Rao *et al.,* [17] who used GPT-4 and GPT-3.5 to recommend breast imaging based on clinical presentation. They showed that when provided with a list of potential imaging modalities GPT-4 models performed consistently higher than GPT-3.5 models in making imaging recommendations for those presenting with breast pain and those undergoing screening. Similar performance was observed in a study by Lim *et al.,* [18] where they showed that GPT-4 models provided with colorectal cancer (CRC) and polyp management screening guidelines, performed significantly better in recognizing high-risk polyp features, recommending CRC screening intervals, and minimizing hallucinations compared to the base GPT-4 model.

4.2. Response to and Surveillance of Adverse Events Associated with Cancer Treatment

Of equal importance is the growing body of research focused on the application of LLMs to better understand responses to cancer treatment [22–26], mitigate adverse events and toxicities [27–31], improve triaging efforts [32], and provide social support for patients [33]. BERT models are more frequently used in this research avenue as they allow for further training of the model on specialized prediction and classification tasks with the addition of a neural classification layer on top of the original BERT model output [34].

Elbatarny *et al.* [22] conducted a study to classify metastatic disease response to treatments from radiology reports. The authors demonstrated that when grounding metastatic cancer response predictions using a standardized oncologic response lexicon (OR-RADS), BERT-based models outperformed human classification performance. Another study evaluated the performance of a fine-tuned BERT model (DFCI-ImagingBERT) in identifying cancer response and progression, using imaging reports of patients with non-small cell lung cancer [23]. They found that although their BERT model (DFCI-ImagingBERT) demonstrated superior performance to other NLP approaches, this difference was marginally better than computationally simpler methods including bag-of-words and convolutional neural network approaches.

In addition to evaluating disease response to treatment, the surveillance and management of adverse events and drug interactions during cancer treatment is important for triaging patients to increase attention as needed to mitigate acute care [27,28]. Studies have explored adverse drug reactions using both clinical and patient-generated health data sources, with variable performance. Zitu *et al.* [30] constructed a corpus correlating drugs with adverse drug events among patients who received immune checkpoint inhibitors. They then used ClinicalBERT, to detect adverse drug events at the sentence level from clinical records, which they did with high performance (*i.e.*, F_1 score: 0.87) [30]. A similar study using BERT-base to extract sentence-level adverse events from breast cancer patient-generated data [31] demonstrated lower performance (*i.e.*, F_1 score: 0.58) in identifying these events, demonstrating variability in the performance of BERT-based models for some clinical decision support tasks dependent on the use case, approach, and training data.

5. Cancer Education

At this point, an emerging use case of LLMs in cancer care has focused on medical education tasks. This natural fit for question-answer interactions has been exclusive to GPT models because of their ability to produce comprehensible and coherent information, as well as the ability to be further prompted for clarification, simplification, and summarization. GPT models have been used as an educational and question-answering tool for patients and radiation oncologists

[35–47] for the detection and education of cancer myths, misconceptions, and misinformation [48], and to simplify and summarize radiology reports [49].

5.1. LLMs for Radiation Oncology Question Answering

We identified two studies using LLMs to answer questions related to radiation oncology to aid the oncologists themselves [35,36]. Holmes *et al.* demonstrated that when questioned on a set of 100 curated radiation oncology questions encompassing topics such as physics, brachytherapy, and imaging, GPT-4, on average, outperformed other LLMs and medical physicists. Huang *et al.* assessed the performance of GPT-4 in answering questions from the Radiation Oncologist in Training Exam (TXIT) [35]. They showed that GPT-4 was capable of achieving passing-level proficiency (*i.e.*, >60%) on questions related to topics including but not limited to treatment planning, treatment toxicity, and disease prognosis. Despite admirable performance in answering various radiation oncology questions, both studies acknowledged that GPT-4 still struggled to correctly answer math-intensive questions, such as calculating radiation decay [35].

Aside from answering domain expert-focused questions, prior work has shown that GPT models are also capable of proficiently displaying the soft skills necessary for the safe and empathetic practice of medicine [50]. These capabilities naturally position GPT models as a promising approach for patient question-answering related to cancer care.

5.2. LLMs to Answer Questions of Patients with Cancer

Multiple studies have demonstrated that GPT-3.5 can accurately answer patient-related questions across a variety of cancers, but often struggles to present this information in a comprehensible way for non-expert audiences. Hermann *et al.*, [37] assessed the ability of GPT-3.5 to answer patient questions related to cervical cancer prevention, diagnosis, treatment, and survivorship. They showed that performance varied according to the types of questions being asked, generating more accurate results for questions on cervical cancer prevention, survivorship, and quality of life, and less accurate responses to questions of treatment. They also noted that responses felt as though they lacked compassion and personability. Musheyev *et al.,* used GPT-3.5 to answer the commonly googled questions about genitourinary malignancies including prostate, bladder, testicular, and kidney cancers. They showed that although answers to these questions were generally correct, they also found that responses lacked a suitable comprehension level for non-experts and lacked actionable information [38]. In another study, Johnson *et al.,* demonstrated near-perfect performance of GPT models for answering questions on common cancer myths and misconceptions but observed similar difficulties in the readability of GPT-answers by non-expert audiences [48].

These issues appear to be better mitigated in studies using GPT-4 for patient cancer education. Rogasch *et al.,* evaluated GPT-4 in answering questions about lung cancer and Hodgkin's lymphoma PET/CT reports [39]. They demonstrated that GPT-4 produced appropriate and meaningful answers while frequently displaying human-like empathy. However, when asked to provide answers to these questions with supporting references, GPT-4 investigators found that only 21% of answers contained valid and current references. Szczesniewski *et al.,* observed similar issues when prompting GPT-4 with patient-focused questions related to urologic malignancies. However, their prompting approaches did not explicitly ask for GPT-4 to provide supporting references to question answers [40]. Another study by Floyd *et al.,* found that GPT-4 performed similarly to GPT-3.5 in answering patient-centric radiation oncology questions in terms of accuracy and comprehensiveness, but outperformed GPT-3.5 when summarizing literature supporting current standard-of-care practice in clinical radiation oncology and in providing existing references to support claims [41].

These shortcomings in producing layperson language for cancer-related tasks further emphasize the need to better consider data augmentation and prompting methods proposed in computer science literature during study design. Some approaches that have been shown to often improve the performance of LLM in text generation tasks include few-shot learning which involves providing sample text as expected output to guide LLM generation [12]. Another approach is the use of retrieval augmented generation (RAG), which involves providing a model with access to an external knowledge source to guide text generation [51]. These external knowledge bases may include resources such as clinical treatment guidelines and published scientific literature, resulting in text responses better guided by evidence-based medicine. However, studies on the effect of RAG approaches to enhance clinical text generation by LLMs are still lacking [52]. Furthermore, more experimental work on various prompting strategies is needed within cancer research. Many studies in our review only reported on a singular standardized prompting approach or did not assess the reproducibility of LLM-produced output. In addition to technical limitations, much of the discrepancies in the lack of patient-readability and compassion of GPT-generated responses may be described by a lack of patient-centered design in the development and evaluation of these approaches. Therefore, further research focused on developing patient-facing applications for patient cancer education should utilize a user-centered design approach, incorporating the feedback of patients and members of the patient's support network, along with oncologists and domain experts in health communications.

6. Research Support Activities

Finally, we identified numerous studies focused on the use of LLMs in various research-related activities in cancer including research idea development and manuscript writing [53], cohort identification and extraction [54–57], information retrieval of drug information and clinical literature [58–64], clinical data processing for research

[65,66], and various NLP-centric tasks (*e.g.*, named entity recognition or NER, clinical note section segmentation) to support cancer research [67–69].

6.1. Assistance with Scholarly Research Activities

There has been much discussion throughout the scientific community about how best to ethically utilize generative LLMs in scientific research, including debates on their use in manuscript writing, grant preparation, and peer-review activities [70,71]. Aside from concerns regarding the lack of scientific rigor [72], much of this debate is centered around 'hallucinations' frequently occurring in LLM-generated text and the lack of some critical information in their responses. We only found one instance of assessing the use of LLMs for scientific writing within cancer. In this study, radiation oncologists with varying levels of research experience were asked to perform research tasks (*e.g.*, literature review, hypothesis development, statistical analysis) with the assistance of GPT-3.5 [53]. Although inexperienced researchers found GPT-3.5 helpful for completing these tasks, the study also found that inexperienced researchers often believed in false GPT-generated information.

6.2. Information Retrieval and Extraction of Scientific Literature

Keeping current with scientific literature is becoming increasingly difficult and time-consuming due to the vast number of studies published each year. Aside from the sheer volume of papers, this process is time-consuming and requires comprehensive search strategies, abstract and full-text review, and interpretation of study relevancy. Zhang *et al.,* trained and employed a BERT model affixed with a text recurrent neural network (TextRNN) to identify cancer-focused literature and produce content-derived labels for each study. Their approach yielded very high performance, achieving an F_1-score of 0.93. However, at a more granular level, extracting meaningful individual elements from studies remains a difficult task. A study by Mutinda *et al.*, developed a BERT-based NER approach to automatically extract core elements of randomized controlled trials (RCT) including the study population, intervention, control, and outcomes from PubMed RCT abstracts [57]. Their approach demonstrated high performance when extracting these core elements, except for statistical analyses which the authors hypothesized was due to non-uniformity among included abstracts and the truncation of meaningful abstract information due to the BERT model's maximum input sequence length of 512 tokens.

These intricacies in cancer-related text are even more apparent in clinical text where typographical errors are incredibly common and can negatively impact the results of studies using this data. Lee *et al.* developed a BERT-based approach in which they used masked-language-modeling (a method in which the next word in a sentence is masked and predicted by the algorithm) in an attempt to identify and subsequently correct typographical issues present in clinical text [66]. They found that their approach yielded a higher performance in error correction compared to pre-existing methods, which is critical as information extraction studies remain very popular in clinical cancer NLP research.

6.3. General Information Extraction and Cancer Phenotyping.

Information extraction studies largely focused on extracting clinical features and phenotypes from clinical records for various cancers including breast, urologic, skin, and lung malignancies [55,56,62,63]. These types of studies are often a proof of concept, assessing the ability of NLP systems to create unique datasets for downstream tasks, whose findings are dependent on accurate extraction. LLM-facilitated information extraction not only enables robust data collection at scale but can also drastically reduce the time and costs typically associated with manual data extraction from clinical records [60]. The performance of LLM-based extraction systems, and most LLM-focused tasks in general, might be further enhanced through data augmentation and targeted prompt engineering approaches [12,51,52,73]. However, techniques in data augmentation and prompt engineering have still been understudied in the cancer literature and is an ongoing area of research.

7. Future Directions and Conclusions

Research and technological advances in LLMs continue to progress at a rapid pace, spurred by increasing computational resources and significant interest from the scientific community in their potential to enhance cancer care. Although this review focused on GPT- and BERT-based model architectures, some studies in our review compared GPT and BERT performances to currently available open-source LLMs such as GatorTron, Galactica, and Longformer [74–77]. However, the widespread development and use of open-source LLMs in cancer care and research is still lacking largely due to data privacy concerns, the immense computational requirements needed for their development, and ease of use for GPT- and BERT-based models by non-technical audiences. However, the LLM-developer community continues to innovate on open-source models, developing quantized LLMs (*i.e.*, models with reduced computation and memory costs) that can be more easily deployed locally with less intensive hardware [78,79]. Additionally, open-source LLM models are readily available on public repositories such as the Hugging Face Hub, allowing for easier access to source code, documentation, and additional information on the models, to promote transparency in their function. Although proprietary closed-source LLMs such as the GPT series and the Llama models have demonstrated state-of-the-art performance in numerous NLP tasks [80], these models lack transparency in their training processes and model parameters, and trail the customization capabilities afforded by open-source models. However, the overall adoption of open-source models in cancer research is still lacking and future research should focus on developing and deploying open-sourced models to better protect patient data, increase the adoption of LLM technologies in diverse clinical settings, and increase trust in LLM-based technologies and clinical decision support provided by LLM-based applications.

In addition to increasing attention towards open-source models, there is a lack of standardized evaluation approaches for generative LLMs within cancer care and clinical care in general. Despite providing comprehensible responses and state-of-the-art performance for many clinical NLP tasks, LLMs remain limited in their capabilities, and human oversight of these technologies is essential [81]. Much of the computer science literature evaluates the performance of models against benchmark datasets such as the GLUE (diverse set of natural language understanding tasks) [82], MMLU (questions on 57 diverse subjects) [83], and SQuAD (100,000 + human-generated question-answer pairs) [84] datasets, to name a few. Such approaches are not as feasible in medicine due to privacy concerns with patient data, insufficient training data, and the inherently complex nature of clinical care [85]. As a result, there remains a critical gap in evaluation metrics for medical applications of LLMs, and both quantitative and qualitative evaluation approaches are needed, along with standardized principles for LLM-text generation in medicine [86,87]. Qualitative metrics are incredibly important in the evaluation of clinically focused LLM applications, and we encourage future research to incorporate the perspectives of patients, caregivers, healthcare providers, and domain experts during the development of these tools to ensure their optimal fit, adoption, and effectiveness among end users. Further, human oversight of LLM-generated clinical information remains critical to ensure completeness and accuracy while minimizing potential health and psychological risks associated with their use in cancer.

The heterogeneity in study design, use cases, and evaluation approaches of studies included in our analysis underscore the need to regulate the development, deployment, and use of large language models within cancer. Throughout the scientific community, there has been a stated need for the regulation of LLMs in healthcare to promote data privacy and governance, ensure patient safety, and improve clinical processes. [88,89] In the clinical setting, LLMs are becoming increasingly exposed to and or trained on patient-specific health information including patient-generated data (*e.g.*, patient-authored messages in patient portals) and EHR data. Although these data sources are rich in clinical information, regulatory oversight of LLMs in cancer care can help ensure that patient data is not compromised or leaked through LLM-generated text and that clinical decision support and patient-facing applications using LLMs pose minimal risk to patients.

Finally, in addition to the need for regulation, rigorous model evaluation approaches, and more open-source LLMs, future research should look to bolster the reliability, clinical utility, interpretability, and accessibility of these models in cancer care. Current research opportunities include the need to improve the reliability and accuracy of LLM-generated output by training open-source LLMs on clinically meaningful text such as standardized ontologies, oncologic treatment guidelines, and deidentified EHR data, to name a few. Additionally, the interpretability and accessibility of LLM-generated content can be bolstered by increased development and deployment of open-source models in cancer research and by developing better approaches for visualizing LLM-generated output in clinically meaningful ways (*e.g.*, grounding LLMs in knowledge graphs, user-centered design of LLM user interfaces, further establishing platforms for hosting and using LLMs).

Our study has several limitations to consider. First, when screening for articles, we queried Scopus and did not include additional scholarly databases such as MEDBASE or Google Scholar. This was deliberate as much of the research focusing on the application of LLMs in cancer is not Medline-indexed and is published in computer science or NLP proceedings. Additionally, our search strategy was limited to studies using BERT- and GPT-based LLMs. We limited our analysis to BERT models due to their sustained use over the past 4 years and extensive documentation, and GPT models due to their current status as the state-of-the-art LLMs for NLP-related tasks, and the widespread integration of GPT-instances into the EHR and EHR software.

Second, we restricted our final articles to journal articles only, excluding reviews and articles posted on preprint servers. Although informative, we excluded review articles to provide primary accounts of the constituent research advances in the application of LLMs to cancer. Additionally, we excluded preprint articles as they are not formally peer-reviewed and published. By excluding studies posted on preprint servers such as arXiv and medRxiv, our review may exclude applications of open-source models in cancer and salient findings outlined in these studies. Finally, cancer-related terms included in our search strategy were limited to “cancer” and “oncology”. As a result, studies containing specific cancer terms or disease names (*e.g.*, “glioblastoma”) and not one of “oncology” or “cancer”, may not have been captured in our search strategy. Since the encompassing themes of our review were inductively identified based on our data collection, additional studies focused on LLMs for cancer care containing excluded keywords, may not have been captured by our search strategy.

In conclusion, the prospective impact of LLMs to support and enhance cancer care is highly apparent. These technologies hold the potential to transform clinical care by reducing burdensome clinical workloads, improving patient and caregiver understanding of cancer, and supporting scientific discovery and clinical advances. However, now more foundational work on the capabilities, limitations, and evaluation approaches of these technologies is needed prior to implementation within clinical care, along with greater utilization of open-source LLMs to promote model transparency, protect patient data privacy, and improve the interpretability of clinical decision support applications using LLMs.

Acknowledgments

This study was supported in part by the Conquer Cancer Foundation and the UCSF Computational Cancer Research Award. JCH is also supported by a career development grant from the American Society for Radiation Oncology and Prostate Cancer Foundation. RB served as a public health informatics consultant for Epi-Vant LLC in 2023.No other conflicts of interest have been declared by the author(s).

References

1. R.L. Siegel, A.N. Giaquinto, A. Jemal, Cancer statistics, 2024, CA: A Cancer Journal for Clinicians 74 (2024) 12–49. https://doi.org/10.3322/caac.21820.
2. N. Keshava, T.S. Toh, H. Yuan, B. Yang, M.P. Menden, D. Wang, Defining subpopulations of differential drug response to reveal novel target populations, Npj Syst Biol Appl 5 (2019) 1–11. https://doi.org/10.1038/s41540-019-0113-4.
3. L.F. Forrest, J. Adams, H. Wareham, G. Rubin, M. White, Socioeconomic Inequalities in Lung Cancer Treatment: Systematic Review and Meta-Analysis, PLOS Medicine 10 (2013) e1001376. https://doi.org/10.1371/journal.pmed.1001376.
4. K.R. Yabroff, J. Lund, D. Kepka, A. Mariotto, Economic Burden of Cancer in the US: Estimates, Projections, and Future Research, Cancer Epidemiol Biomarkers Prev 20 (2011) 2006–2014. https://doi.org/10.1158/1055-9965.EPI-11-0650.
5. K.R. Yabroff, A. Mariotto, F. Tangka, J. Zhao, F. Islami, H. Sung, R.L. Sherman, S.J. Henley, A. Jemal, E.M. Ward, Annual Report to the Nation on the Status of Cancer, Part 2: Patient Economic Burden Associated With Cancer Care, JNCI: Journal of the National Cancer Institute 113 (2021) 1670–1682. https://doi.org/10.1093/jnci/djab192.
6. M. Gholipour, R. Khajouei, P. Amiri, S. Hajesmaeel Gohari, L. Ahmadian, Extracting cancer concepts from clinical notes using natural language processing: a systematic review, BMC Bioinformatics 24 (2023) 405. https://doi.org/10.1186/s12859-023-05480-0.
7. W. Yim, M. Yetisgen, W.P. Harris, S.W. Kwan, Natural Language Processing in Oncology: A Review, JAMA Oncology 2 (2016) 797–804. https://doi.org/10.1001/jamaoncol.2016.0213.
8. L. Wang, S. Fu, A. Wen, X. Ruan, H. He, S. Liu, S. Moon, M. Mai, I.B. Riaz, N. Wang, P. Yang, H. Xu, J.L. Warner, H. Liu, Assessment of Electronic Health Record for Cancer Research and Patient Care Through a Scoping Review of Cancer Natural Language Processing, JCO Clin Cancer Inform (2022) e2200006. https://doi.org/10.1200/CCI.22.00006.
9. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, (2019). http://arxiv.org/abs/1810.04805 (accessed September 27, 2023).
10. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention Is All You Need, (2023). http://arxiv.org/abs/1706.03762 (accessed January 28, 2024).
11. A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving Language Understanding by Generative Pre-Training, (n.d.).
12. T.B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D.M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language Models are Few-Shot Learners, (2020). http://arxiv.org/abs/2005.14165 (accessed December 2, 2023).
13. S. Lukac, D. Dayan, V. Fink, E. Leinert, A. Hartkopf, K. Veselinovic, W. Janni, B. Rack, K. Pfister, B. Heitmeir, F. Ebner, Evaluating ChatGPT as an adjunct for the multidisciplinary tumor board decision-making in primary breast cancer cases, Arch Gynecol Obstet 308 (2023) 1831–1844. https://doi.org/10.1007/s00404-023-07130-5.
14. S. Chen, B.H. Kann, M.B. Foote, H.J.W.L. Aerts, G.K. Savova, R.H. Mak, D.S. Bitterman, Use of Artificial Intelligence Chatbots for Cancer Treatment Information, JAMA Oncol 9 (2023) 1459. https://doi.org/10.1001/jamaoncol.2023.2954.
15. S. Griewing, N. Gremke, U. Wagner, M. Lingenfelder, S. Kuhn, J. Boekhoff, Challenging ChatGPT 3.5 in Senology—An Assessment of Concordance with Breast Cancer Tumor Board Decision Making, Journal of Personalized Medicine 13 (2023). https://doi.org/10.3390/jpm13101502.
16. M. Benary, X.D. Wang, M. Schmidt, D. Soll, G. Hilfenhaus, M. Nassir, C. Sigler, M. Knödler, U. Keller, D. Beule, U. Keilholz, U. Leser, D.T. Rieke, Leveraging Large Language Models for Decision Support in Personalized Oncology, JAMA Netw Open 6 (2023) e2343689. https://doi.org/10.1001/jamanetworkopen.2023.43689.
17. A. Rao, J. Kim, M. Kamineni, M. Pang, W. Lie, K.J. Dreyer, M.D. Succi, Evaluating GPT as an Adjunct for Radiologic Decision Making: GPT-4 Versus GPT-3.5 in a Breast Imaging Pilot, Journal of the American College of Radiology 20 (2023) 990–997. https://doi.org/10.1016/j.jacr.2023.05.003.
18. D.Y.Z. Lim, Y.B. Tan, J.T.E. Koh, J.Y.M. Tung, G.G.R. Sng, D.M.Y. Tan, C.-K. Tan, ChatGPT on guidelines: Providing contextual knowledge to GPT allows it to provide advice on appropriate colonoscopy intervals, Journal of Gastroenterology and Hepatology (Australia) (2023). https://doi.org/10.1111/jgh.16375.
19. J. Zhou, T. Li, S.J. Fong, N. Dey, R.G. Crespo, Exploring ChatGPT's Potential for Consultation, Recommendations and Report Diagnosis: Gastric Cancer and Gastroscopy Reports' Case, International Journal of Interactive Multimedia and Artificial Intelligence 8 (2023) 7–13. https://doi.org/10.9781/ijimai.2023.04.007.
20. J.R. Lechien, C.-M. Chiesa-Estomba, R. Baudouin, S. Hans, Accuracy of ChatGPT in head and neck oncological board decisions: preliminary findings, European Archives of Oto-Rhino-Laryngology (2023). https://doi.org/10.1007/s00405-023-08326-w.
21. J.M. Choo, H.S. Ryu, J.S. Kim, J.Y. Cheong, S.-J. Baek, J.M. Kwak, J. Kim, Conversational artificial intelligence (chatGPTTM) in the management of complex colorectal cancer patients: early experience, ANZ Journal of Surgery (2023). https://doi.org/10.1111/ans.18749.
22. L. Elbatarny, R.K.G. Do, N. Gangai, F. Ahmed, S. Chhabra, A.L. Simpson, Applying Natural Language Processing to Single-Report Prediction of Metastatic Disease Response Using the OR-RADS Lexicon, Cancers 15 (2023). https://doi.org/10.3390/cancers15204909.
23. H.A. Elmarakeby, P.S. Trukhanov, V.M. Arroyo, I.B. Riaz, D. Schrag, E.M. Van Allen, K.L. Kehl, Empirical evaluation of language modeling to ascertain cancer outcomes from clinical text reports, BMC Bioinformatics 24 (2023). https://doi.org/10.1186/s12859-023-05439-1.
24. M.A. Fink, K. Kades, A. Bischoff, M. Moll, M. Schnell, M. Küchler, G. Köhler, J. Sellner, C.P. Heussel, H.-U. Kauczor, H.-P. Schlemmer, K. Maier-Hein, T.F. Weber, J. Kleesiek, Deep Learning–based Assessment of Oncologic Outcomes from Natural Language Processing of Structured Radiology Reports, Radiology: Artificial Intelligence 4 (2022). https://doi.org/10.1148/ryai.220055.
25. R. Lian, V. Hsiao, J. Hwang, Y. Ou, S.E. Robbins, N.P. Connor, C.L. Macdonald, R.S. Sippel, W.A. Sethares, D.F. Schneider, Predicting health-related quality of life change using natural language processing in thyroid cancer, Intelligence-Based Medicine 7 (2023). https://doi.org/10.1016/j.ibmed.2023.100097.
26. R.S.Y.C. Tan, Q. Lin, G.H. Low, R. Lin, T.C. Goh, C.C.E. Chang, F.F. Lee, W.Y. Chan, W.C. Tan, H.J. Tey, F.L. Leong, H.Q. Tan, W.L. Nei, W.Y. Chay, D.W.M. Tai, G.G.Y. Lai, L.T.-E. Cheng, F.Y. Wong, M.C.H. Chua, M.L.K. Chua, D.S.W. Tan, C.H. Thng, I.B.H. Tan, H.T. Ng, Inferring cancer disease response from radiology reports using large language models with data augmentation and prompting, Journal of the American Medical Informatics Association : JAMIA 30 (2023) 1657–1664. https://doi.org/10.1093/jamia/ocad133.
27. M. Ismail, S. Khan, F. Khan, S. Noor, H. Sajid, S. Yar, I. Rasheed, Prevalence and significance of potential drug-drug interactions among cancer patients receiving chemotherapy, BMC Cancer 20 (2020) 335. https://doi.org/10.1186/s12885-020-06855-9.
28. R. Du, X. Wang, L. Ma, L.M. Larcher, H. Tang, H. Zhou, C. Chen, T. Wang, Adverse reactions of targeted therapy in cancer patients: a retrospective study of hospital medical data in China, BMC Cancer 21 (2021) 206. https://doi.org/10.1186/s12885-021-07946-x.
29. S. Chen, M. Guevara, N. Ramirez, A. Murray, J.L. Warner, H.J.W.L. Aerts, T.A. Miller, G.K. Savova, R.H. Mak, D.S. Bitterman, Natural Language Processing to Automatically Extract the Presence and Severity of Esophagitis in Notes of Patients Undergoing Radiotherapy, JCO Clinical Cancer Informatics 7 (2023) e2300048. https://doi.org/10.1200/CCI.23.00048.
30. M.M. Zitu, S. Zhang, D.H. Owen, C. Chiang,

L. Li, Generalizability of machine learning methods in detecting adverse drug events from clinical narratives in electronic medical records, Frontiers in Pharmacology 14 (2023). https://doi.org/10.3389/fphar.2023.1218679.

31. S. Nishioka, M. Asano, S. Yada, E. Aramaki, H. Yajima, Y. Yanagisawa, K. Sayama, H. Kizaki, S. Hori, Adverse event signal extraction from cancer patients' narratives focusing on impact on their daily-life activities, Sci Rep 13 (2023) 15516. https://doi.org/10.1038/s41598-023-42496-1.
32. G. Gebrael, K.K. Sahu, B. Chigarira, N. Tripathi, V. Mathew Thomas, N. Sayegh, B.L. Maughan, N. Agarwal, U. Swami, H. Li, Enhancing Triage Efficiency and Accuracy in Emergency Rooms for Patients with Metastatic Prostate Cancer: A Retrospective Analysis of Artificial Intelligence-Assisted Triage Using ChatGPT 4.0, Cancers 15 (2023). https://doi.org/10.3390/cancers15143717.
33. T. Watanabe, S. Yada, E. Aramaki, H. Yajima, H. Kizaki, S. Hori, Extracting Multiple Worries from Breast Cancer Patient Blogs Using Multilabel Classification with the Natural Language Processing Model Bidirectional Encoder Representations from Transformers: Infodemiology Study of Blogs, JMIR Cancer 8 (2022). https://doi.org/10.2196/37840.
34. C. Sun, X. Qiu, Y. Xu, X. Huang, How to Fine-Tune BERT for Text Classification?, in: M. Sun, X. Huang, H. Ji, Z. Liu, Y. Liu (Eds.), Chinese Computational Linguistics, Springer International Publishing, Cham, 2019: pp. 194–206. https://doi.org/10.1007/978-3-030-32381-3_16.
35. Y. Huang, A. Gomaa, S. Semrau, M. Haderlein, S. Lettmaier, T. Weissmann, J. Grigo, H.B. Tkhayat, B. Frey, U. Gaipl, L. Distel, A. Maier, R. Fietkau, C. Bert, F. Putz, Benchmarking ChatGPT-4 on a radiation oncology in-training exam and Red Journal Gray Zone cases: potentials and challenges for ai-assisted medical education and decision making in radiation oncology, Frontiers in Oncology 13 (2023). https://doi.org/10.3389/fonc.2023.1265024.
36. J. Holmes, Z. Liu, L. Zhang, Y. Ding, T.T. Sio, L.A. McGee, J.B. Ashman, X. Li, T. Liu, J. Shen, W. Liu, Evaluating large language models on a highly-specialized topic, radiation oncology physics, Frontiers in Oncology 13 (2023). https://doi.org/10.3389/fonc.2023.1219326.
37. C.E. Hermann, J.M. Patel, L. Boyd, W.B. Growdon, E. Aviki, M. Stasenko, Let's chat about cervical cancer: Assessing the accuracy of ChatGPT responses to cervical cancer questions, Gynecologic Oncology 179 (2023) 164–168. https://doi.org/10.1016/j.ygyno.2023.11.008.
38. D. Musheyev, A. Pan, S. Loeb, A.E. Kabarriti, How Well Do Artificial Intelligence Chatbots Respond to the Top Search Queries About Urological Malignancies?, European Urology 85 (2024) 13–16. https://doi.org/10.1016/j.eururo.2023.07.004.
39. J.M.M. Rogasch, G. Metzger, M. Preisler, M. Galler, F. Thiele, W. Brenner, F. Feldhaus, C. Wetz, H. Amthauer, C. Furth, I. Schatka, ChatGPT: Can you prepare my patients for [18F] FDG PET/CT and explain my reports?, Journal of Nuclear Medicine 64 (2023) 1876–1879. https://doi.org/10.2967/jnumed.123.266114.
40. J.J. Szczesniewski, C. Tellez Fouz, A. Ramos Alba, F.J. Diaz Goizueta, A. García Tello, L. Llanes González, ChatGPT and most frequent urological diseases: analysing the quality of information and potential risks for patients, World Journal of Urology 41 (2023) 3149–3153. https://doi.org/10.1007/s00345-023-04563-0.
41. W. Floyd, T. Kleber, D.J. Carpenter, M. Pasli, J. Qazi, C. Huang, J. Leng, B.G. Ackerson, M. Pierpoint, J.K. Salama, M.J. Boyer, Current Strengths and Weaknesses of ChatGPT as a Resource for Radiation Oncology Patients and Providers, International Journal of Radiation Oncology*Biology*Physics (2023). https://doi.org/10.1016/j.ijrobp.2023.10.020.
42. A. Cocci, M. Pezzoli, M. Lo Re, G.I. Russo, M.G. Asmundo, M. Fode, G. Cacciamani, S. Cimino, A. Minervini, E. Durukan, Quality of information and appropriateness of ChatGPT outputs for urology patients, Prostate Cancer Prostatic Dis (2023) 1–6. https://doi.org/10.1038/s41391-023-00705-y.
43. B. Coskun, G. Ocakoglu, M. Yetemen, O. Kaygisiz, Can ChatGPT, an Artificial Intelligence Language Model, Provide Accurate and High-quality Patient Information on Prostate Cancer?, Urology 180 (2023) 35–58. https://doi.org/10.1016/j.urology.2023.05.040.
44. R.J. Davis, O. Ayo-Ajibola, M.E. Lin, M.S. Swanson, T.N. Chambers, D.I. Kwon, N.C. Kokot, Evaluation of Oropharyngeal Cancer Information from Revolutionary Artificial Intelligence Chatbot, Laryngoscope (2023). https://doi.org/10.1002/lary.31191.
45. E.-M. Braun, I. Juhasz-Böss, E.-F. Solomayer, D. Truhn, C. Keller, V. Heinrich, B.J. Braun, Will I soon be out of my job? Quality and guideline conformity of ChatGPT therapy suggestions to patient inquiries with gynecologic symptoms in a palliative setting, Archives of Gynecology and Obstetrics (2023). https://doi.org/10.1007/s00404-023-07272-6.
46. A.A. Rahsepar, N. Tavakoli, G.H.J. Kim, C. Hassani, F. Abtin, A. Bedayat, How AI Responds to Common Lung Cancer Questions: ChatGPT versus Google Bard, Radiology 307 (2023). https://doi.org/10.1148/radiol.230922.
47. A. Pan, D. Musheyev, D. Bockelman, S. Loeb, A.E. Kabarriti, Assessment of Artificial Intelligence Chatbot Responses to Top Searched Queries about Cancer, JAMA Oncology 9 (2023) 1437–1440. https://doi.org/10.1001/jamaoncol.2023.2947.
48. S.B. Johnson, A.J. King, E.L. Warner, S. Aneja, B.H. Kann, C.L. Bylund, Using ChatGPT to evaluate cancer myths and misconceptions: artificial intelligence and cancer information, JNCI Cancer Spectrum 7 (2023). https://doi.org/10.1093/jncics/pkad015.
49. Q. Lyu, J. Tan, M.E. Zapadka, J. Ponnatapura, C. Niu, K.J. Myers, G. Wang, C.T. Whitlow, Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential, Visual Computing for Industry, Biomedicine, and Art 6 (2023). https://doi.org/10.1186/s42492-023-00136-5.
50. D. Brin, V. Sorin, A. Vaid, A. Soroush, B.S. Glicksberg, A.W. Charney, G. Nadkarni, E. Klang, Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments, Sci Rep 13 (2023) 16492. https://doi.org/10.1038/s41598-023-43436-9.
51. D. Cai, Y. Wang, L. Liu, S. Shi, Recent Advances in Retrieval-Augmented Text Generation, in: 2022: pp. 3417–3419. https://doi.org/10.1145/3477495.3532682.
52. Y. Guo, W. Qiu, G. Leroy, S. Wang, T. Cohen, Retrieval augmentation of large language models for lay language generation, Journal of Biomedical Informatics 149 (2024). https://doi.org/10.1016/j.jbi.2023.104580.
53. M. Guckenberger, N. Andratschke, M. Ahmadsei, S.M. Christ, A.E. Heusel, S. Kamal, T.E. Kroese, E.L. Looman, S. Reichl, E. Vlaskou Badra, J. von der Grün, J. Willmann, S. Tanadini-Lang, M. Mayinger, Potential of ChatGPT in facilitating research in radiation oncology?, Radiotherapy and Oncology 188 (2023). https://doi.org/10.1016/j.radonc.2023.109894.
54. S. Mithun, A.K. Jha, U.B. Sherkhane, V. Jaiswar, N.C. Purandare, V. Rangarajan, A. Dekker, S. Puts, I. Bermejo, L. Wee, Development and validation of deep learning and BERT models for classification of lung cancer radiology reports, Informatics in Medicine Unlocked 40 (2023). https://doi.org/10.1016/j.imu.2023.101294.
55. M.A. Fink, A. Bischoff, C.A. Fink, M. Moll, J. Kroschke, L. Dulz, C.P. Heußel, H.-U. Kauczor, T.F. Weber, Potential of ChatGPT and GPT-4 for Data Mining of Free-Text CT Reports on Lung Cancer, Radiology 308 (2023). https://doi.org/10.1148/radiol.231362.
56. S. Zhou, N. Wang, L. Wang, H. Liu, R. Zhang, CancerBERT: A cancer domain-specific language model for extracting breast cancer phenotypes from electronic health records, Journal of the American Medical Informatics Association 29 (2022) 1208–1216. https://doi.org/10.1093/jamia/ocac040.
57. F.W. Mutinda, K. Liew, S. Yada, S. Wakamiya, E. Aramaki, Automatic data extraction to support meta-analysis statistical analysis: a case study on breast cancer, BMC Medical Informatics and Decision Making 22 (2022). https://doi.org/10.1186/s12911-022-01897-4.
58. A. Gendrin, L. Souliotis, J. Loudon-Griffiths, R. Aggarwal, D. Amoako, G. Desouza, S. Dimitrievska, P. Metcalfe, E. Louvet, H. Sahni, Identifying Patient Populations in Texts Describing Drug Approvals Through Deep Learning–Based Information Extraction: Development of a Natural Language Processing Algorithm, JMIR Formative Research 7 (2023). https://doi.org/10.2196/44876.
59. Y. Zhang, X. Li, Y. Liu, A. Li, X. Yang, X. Tang, A Multilabel Text Classifier of Cancer Literature at the Publication Level: Methods Study of Medical Text Classification, JMIR Medical Informatics 11 (2023). https://doi.org/10.2196/44892.
60. H.S. Choi, J.Y. Song, K.H. Shin, J.H. Chang, B.-S. Jang, Developing prompts from large language model for extracting clinical information from pathology and ultrasound reports in breast cancer, Radiation Oncology Journal

41 (2023) 209–216. https://doi.org/10.3857/roj.2023.00633.

61. G. Kuling, B. Curpen, A.L. Martel, BI-RADS BERT and Using Section Segmentation to Understand Radiology Reports, Journal of Imaging 8 (2022). https://doi.org/10.3390/jimaging8050131.
62. J.R. Mitchell, P. Szepietowski, R. Howard, P. Reisman, J.D. Jones, P. Lewis, B.L. Fridley, D.E. Rollison, A Question-and-Answer System to Extract Data From Free-Text Oncological Pathology Reports (CancerBERT Network): Development Study, Journal of Medical Internet Research 24 (2022). https://doi.org/10.2196/27210.
63. H. Huang, F.X.Y. Lim, G.T. Gu, M.J. Han, A.H.S. Fang, E.H.S. Chia, E.Y.T. Bei, S.Z. Tham, H.S.S. Ho, J.S.P. Yuen, A. Sun, J.K.S. Lim, Natural language processing in urology: Automated extraction of clinical information from histopathology reports of uro-oncology procedures, Heliyon 9 (2023). https://doi.org/10.1016/j.heliyon.2023.e14793.
64. P. Uskaner Hepsağ, S.A. Özel, K. Dalcı, A. Yazıcı, Using BERT models for breast cancer diagnosis from Turkish radiology reports, Language Resources and Evaluation (2023). https://doi.org/10.1007/s10579-023-09669-w.
65. C. Fang, N. Markuzon, N. Patel, J.-D. Rueda, Natural Language Processing for Automated Classification of Qualitative Data From Interviews of Patients With Cancer, Value in Health 25 (2022) 1995–2002. https://doi.org/10.1016/j.jval.2022.06.004.
66. E.B. Lee, G.E. Heo, C.M. Choi, M. Song, MLM-based typographical error correction of unstructured medical texts for named entity recognition, BMC Bioinformatics 23 (2022). https://doi.org/10.1186/s12859-022-05035-9.
67. M. Bali, A.S. Pichandi, NeRBERT- A Biomedical Named Entity Recognition Tagger, Revue d'Intelligence Artificielle 37 (2023) 239–247. https://doi.org/10.18280/ria.370130.
68. O.S. Pabón, O. Montenegro, M. Torrente, A.R. González, M. Provencio, E. Menasalvas, Negation and uncertainty detection in clinical texts written in Spanish: a deep learning-based approach, PeerJ Computer Science 8 (2022). https://doi.org/10.7717/PEERJ-CS.913.
69. O. Solarte-Pabón, O. Montenegro, A. García-Barragán, M. Torrente, M. Provencio, E. Menasalvas, V. Robles, Transformers for extracting breast cancer information from Spanish clinical narratives, Artificial Intelligence in Medicine 143 (2023). https://doi.org/10.1016/j.artmed.2023.102625.
70. M. Hosseini, S.P.J.M. Horbach, Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review, Research Integrity and Peer Review 8 (2023) 4. https://doi.org/10.1186/s41073-023-00133-5.
71. J.G. Meyer, R.J. Urbanowicz, P.C.N. Martin, K. O'Connor, R. Li, P.-C. Peng, T.J. Bright, N. Tatonetti, K.J. Won, G. Gonzalez-Hernandez, J.H. Moore, ChatGPT and large language models in academia: opportunities and challenges, BioData Mining 16 (2023) 20. https://doi.org/10.1186/s13040-023-00339-9.
72. B. Fecher, M. Hebing, M. Laufer, J. Pohle, F. Sofsky, Friend or foe? Exploring the implications of large language models on the science system, AI & Soc (2023). https://doi.org/10.1007/s00146-023-01791-1.
73. T. Kojima, S.S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large Language Models are Zero-Shot Reasoners, arXiv.Org (2022). https://arxiv.org/abs/2205.11916v4 (accessed January 14, 2024).
74. R. Taylor, M. Kardas, G. Cucurull, T. Scialom, A. Hartshorn, E. Saravia, A. Poulton, V. Kerkez, R. Stojnic, Galactica: A Large Language Model for Science, arXiv.Org (2022). https://arxiv.org/abs/2211.09085v1 (accessed May 3, 2024).
75. I. Beltagy, M.E. Peters, A. Cohan, Longformer: The Long-Document Transformer, arXiv.Org (2020). https://arxiv.org/abs/2004.05150v2 (accessed May 3, 2024).
76. X. Yang, A. Chen, N. PourNejatian, H.C. Shin, K.E. Smith, C. Parisien, C. Compas, C. Martin, M.G. Flores, Y. Zhang, T. Magoc, C.A. Harle, G. Lipori, D.A. Mitchell, W.R. Hogan, E.A. Shenkman, J. Bian, Y. Wu, GatorTron: A Large Clinical Language Model to Unlock Patient Information from Unstructured Electronic Health Records, arXiv.Org (2022). https://arxiv.org/abs/2203.03540v3 (accessed May 3, 2024).
77. C. Peng, X. Yang, A. Chen, K.E. Smith, N. PourNejatian, A.B. Costa, C. Martin, M.G. Flores, Y. Zhang, T. Magoc, G. Lipori, D.A. Mitchell, N.S. Ospina, M.M. Ahmed, W.R. Hogan, E.A. Shenkman, Y. Guo, J. Bian, Y. Wu, A study of generative large language model for medical research and healthcare, Npj Digit. Med. 6 (2023) 210. https://doi.org/10.1038/s41746-023-00958-w.
78. T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, QLoRA: Efficient Finetuning of Quantized LLMs, (2023). http://arxiv.org/abs/2305.14314 (accessed January 31, 2024).
79. G. Xiao, J. Lin, M. Seznec, H. Wu, J. Demouth, S. Han, SmoothQuant: Accurate and Efficient Post-Training Quantization for Large Language Models, in: Proceedings of the 40th International Conference on Machine Learning, PMLR, 2023: pp. 38087–38099. https://proceedings.mlr.press/v202/xiao23c.html (accessed January 31, 2024).
80. H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, LLaMA: Open and Efficient Foundation Language Models, (2023). https://doi.org/10.48550/arXiv.2302.13971.
81. M.R. Waters, S. Aneja, J.C. Hong, Unlocking the Power of ChatGPT, Artificial Intelligence, and Large Language Models: Practical Suggestions for Radiation Oncologists, Practical Radiation Oncology 13 (2023) e484–e490. https://doi.org/10.1016/j.prro.2023.06.011.
82. A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, S.R. Bowman, GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding, (2019). http://arxiv.org/abs/1804.07461 (accessed January 31, 2024).
83. D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, J. Steinhardt, Measuring Massive Multitask Language Understanding, (2021). http://arxiv.org/abs/2009.03300 (accessed January 31, 2024).
84. P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, SQuAD: 100,000+ Questions for Machine Comprehension of Text, arXiv.Org (2016). https://arxiv.org/abs/1606.05250v3 (accessed January 31, 2024).
85. J. Clusmann, F.R. Kolbinger, H.S. Muti, Z.I. Carrero, J.-N. Eckardt, N.G. Laleh, C.M.L. Löffler, S.-C. Schwarzkopf, M. Unger, G.P. Veldhuizen, S.J. Wagner, J.N. Kather, The future landscape of large language models in medicine, Commun Med 3 (2023) 141. https://doi.org/10.1038/s43856-023-00370-1.
86. K.E. Goodman, P.H. Yi, D.J. Morgan, AI-Generated Clinical Summaries Require More Than Accuracy, JAMA (2024). https://doi.org/10.1001/jama.2024.0555.
87. A.J. Thirunavukarasu, D.S.J. Ting, K. Elangovan, L. Gutierrez, T.F. Tan, D.S.W. Ting, Large language models in medicine, Nat Med 29 (2023) 1930–1940. https://doi.org/10.1038/s41591-023-02448-8.
88. B. Meskó, E.J. Topol, The imperative for regulatory oversight of large language models (or generative AI) in healthcare, Npj Digit. Med. 6 (2023) 1–6. https://doi.org/10.1038/s41746-023-00873-0.
89. J.C.L. Ong, S.Y.-H. Chang, W. William, A.J. Butte, N.H. Shah, L.S.T. Chew, N. Liu, F. Doshi-Velez, W. Lu, J. Savulescu, D.S.W. Ting, Ethical and regulatory challenges of large language models in medicine, The Lancet Digital Health 0 (2024). https://doi.org/10.1016/S2589-7500(24)00061-X.

Correspondence to:
Julian Hong, MD, MS
Address: 1825 Fourth St., Suite L1101
San Francisco, CA 94158
Email: julian.hong@ucsf.edu
Phone: 415.502.1645
Fax: 415.353.9883

Copyright