The potential of large language models to advance precision oncology

Shufan Liang,[a,e] Jiangjiang Zhang,[b,e] Xingting Liu,[a,e] Yinkui Huang,[b] Jun Shao,[a] Xiaohong Liu,[b,c] Weimin Li,[a,d,***] Guangyu Wang,[b,**] and Chengdi Wang[a,d,*]

[a]Department of Pulmonary and Critical Care Medicine, State Key Laboratory of Respiratory Health and Multimorbidity, Targeted Tracer Research and Development Laboratory, Frontiers Science Center for Disease-Related Molecular Network, West China Hospital, West China School of Medicine, Sichuan University, Chengdu, China
[b]State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, China
[c]UCL Cancer Institute, University College London, London WC1E 6BT, UK
[d]Frontiers Medical Center, Tianfu Jincheng Laboratory, Chengdu, China

eBioMedicine
2025;115: 105695

Published Online xxx
https://doi.org/10.1016/j.ebiom.2025.105695

Summary

With the rapid development of artificial intelligence (AI) within medicine, the emergence of large language models (LLMs) has gradually reached the forefront of clinical research. In oncology, by mining the underlying connection between a text or image input and the desired output, LLMs demonstrate great potential for managing tumours. In this review, we provide a brief description of the development of LLMs, followed by model construction strategies and general medical functions. We then elaborate on the role of LLMs in cancer screening and diagnosis, metastasis identification, tumour staging, treatment recommendation, and documentation processing tasks by decoding various types of clinical data. Moreover, the current barriers faced by LLMs, such as hallucinations, ethical problems, limited application, and so on, are outlined along with corresponding solutions, where the further purpose is to inspire improvement and innovation in this field with respect to harnessing LLMs for advancing precision oncology.

Keywords: Oncology; Large language model; Artificial intelligence; Diagnosis; Treatment

Introduction

In recent years, the drastic growth exhibited by artificial intelligence (AI) has gradually driven the evolution of medicine, showing excellent performance in multiple tasks.[1–3] Deep learning, which is a major branch of AI that involves the basic construction of deep neural networks (DNNs), such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs),[4,5] has attracted much attention from researchers and has achieved impressive progress in management of various diseases, including respiratory infections, ophthalmopathy, and neurologic disorders.[6–10] Cancer, as one of the leading causes of mortality worldwide, poses a tremendous challenge to public health.[11,12] The gradual expansion of the use of AI techniques has led to a paradigm shift in the fight against oncology, resulting in advancements in various areas, from diagnosis, treatment, and prognosis to molecular characterization.[13]

Although the release of the Chat Generative Pre-trained Transformer (ChatGPT) by OpenAI has attracted much attention, the more recently created open model from China, DeepSeek, which was developed at lower costs, has caused another extraordinary stir worldwide since AI has been developing for decades from the 1950s (Fig. 1).[14,15] The foundation technology of these models is large language model (LLM), which is essentially an implemented form of deep learning that contains numerous parameters and facilitates natural language processing (NLP) tasks.[16] In general, language modelling (LM) aims to model the generative likelihood of word sequences and predict the probabilities of posterior tokens. By guiding a neural network to accomplish the LM task on a training corpus, a pre-trained language model (PLM) can be created. Additionally, researchers have reported that scaling a PLM often enhances the capacity of the model to address downstream tasks, thereby transforming the PLM into an LLM.[17] Compared with conventional language

*Corresponding author. Department of Pulmonary and Critical Care Medicine, State Key Laboratory of Respiratory Health and Multimorbidity, Targeted Tracer Research and Development Laboratory, Frontiers Science Center for Disease-Related Molecular Network, West China Hospital, West China School of Medicine, Sichuan University, Chengdu 610041, China.
**Corresponding author. State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100088, China.
***Corresponding author. Department of Pulmonary and Critical Care Medicine, State Key Laboratory of Respiratory Health and Multimorbidity, Targeted Tracer Research and Development Laboratory, Frontiers Science Center for Disease-Related Molecular Network, West China Hospital, West China School of Medicine, Sichuan University, Chengdu 610041, China.
E-mail addresses: chengdi_wang@scu.edu.cn (C. Wang), guangyu.wang24@gmail.com (G. Wang), weimi003@scu.edu.cn (W. Li).
[e]These authors contributed equally to this study.

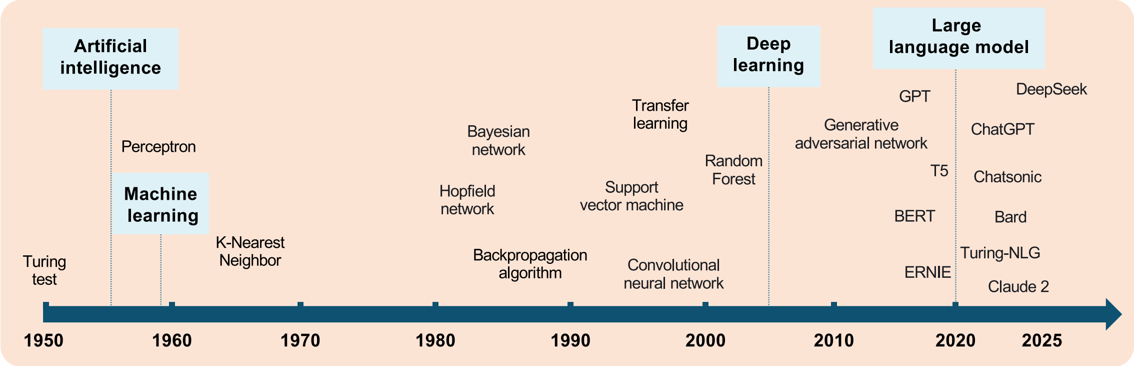

Fig. 1: **A graph of the development of artificial intelligence (AI), machine learning, deep learning, and large language models (LLM) with time.** The development of AI underwent rapid progress from machine learning, deep learning, to the emerging LLMs. The important milestones were coloured in light blue and the examples of models were listed in time order. BERT, Bidirectional Encoder Representation from Transformers; ChatGPT, Chat Generative Pre-trained Transformer; ERNIE, Enhanced Representation through Knowledge Integration; GPT, Generative Pre-trained Transformer; Turing-NLG, Turing Natural Language Generation.

processing methods, LLMs are particularly adept at handling large and complex datasets due to their expansive parameter space and extended context windows, which enable them to capture intricate relationships within data. Along with further explorations of inputting diverse data more than texts into the LLMs, multimodal LLMs (MLLMs), capable of handling data in diverse modes, have been developed, representing another big step forward comprehensive document processing.[18,19]

In this review, we initially introduce a set of model construction strategies and general use of LLMs in medicine. Inspired by the progression of this technology in oncology, we emphasize the applications of LLMs in screening, diagnosis, metastasis prediction, tumour staging, treatment recommendation, and documentation processing tasks related to malignancies, such as breast cancer, lung cancer, and prostate cancer. Finally, we highlight the core challenges that must be overcome for LLMs to deliver sufficient clinical value, along with corresponding strategies that can accelerate this new generation of models to move closer to precision oncology.

Model construction strategies and medical application of LLMs

Since GPT-1 was released in 2018 which marked a key milestone of the LLM evolution process, the development of LLMs has exhibited overall progress in terms of increasing their numbers of parameters and expanding knowledge, thereby improving the performance of the updated models, such as GPT-4.[20,21] Here, we divide the LLM implementation process into three key components. First, to enable a model to acquire domain-specific knowledge and make personalized decisions for individual patients, it is essential for the model to leverage three levels of information: internal knowledge, external knowledge, and patient-specific data. Then, the model effectively leverages the information and subsequently generates the task-specific outputs based on appropriate prompts. In terms of applications in medicine, empowered by the previously mentioned approaches, a wide range of tasks that are commonly encountered in the clinic can potentially be solved by LLMs. This structured framework outlines a potential pathway for integrating LLMs into cancer research, supporting both knowledge acquisition and strong task-specific performance across various applications (Fig. 2).

Task-agnostic property with internal knowledge

An LLM is capable of being a generalist, which means that it can address diverse clinical issues, whereas task-specific models are adept in a certain domain, such as a single type of cancer. Possessing a wealth of medical knowledge, LLMs have the potential to serve as medical domain-specific knowledge bases (KBs), allowing themselves to aid users across a range of clinical issues.[22] Equipped with zero-shot or few-shot skills from their pretraining stages, LLMs can make inferences concerning unseen downstream tasks without conducting task-specific training. A notable example is that models have demonstrated the ability to accurately answer complex questions on medical board exams even without fine-tuning, a crucial technique that enhances the adaptability of LLMs for use in domain-specific tasks. In contrast, models that rely on supervised learning, such as bidirectional encoder representations from transformers (BERT), require a significant amount of labelled data.

Fine-tuning enhances the task-specific performance of LLMs

While the models are pre-trained on extensive, general medical datasets, fine-tuning refines their performance by adjusting their parameters based on a smaller, domain-specific dataset.[23] This process enables such a model to not only retain the foundational knowledge

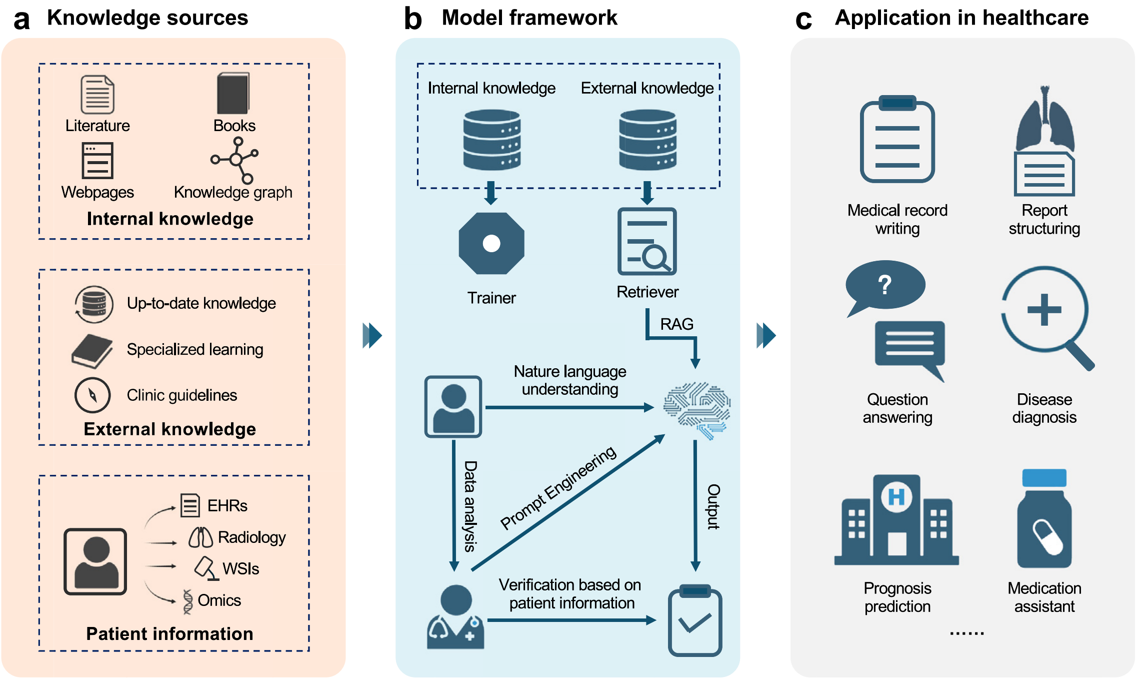

Fig. 2: **Overall model construction strategies of large language models (LLMs) and application in healthcare. a,** The pre-training procedure of LLMs could be conducted on a comprehensive corpus of general medical knowledge to address clinical reasoning challenges. The foundational knowledge encompasses a wide array of sources, including literature, publications, webpages, and structured knowledge graphs. Auxiliary external knowledge will be integrated to enhance the knowledge retrieval capacity of the model, followed by comprehensive analysis of diverse patient data, such as electronic health records (EHRs) and whole slide images (WSIs), to formulate clinical decisions. **b,** The diagram delineates the core components in the operational framework of LLMs targeting clinic tasks. Through the amalgamation of embedded parametric knowledge and supplemental external information, the model is equipped with a comprehensive understanding of various diseases, enabling LLMs to deliver personalized care based on the patient's information by capitalizing on its advanced inference mechanisms. Moreover, physicians are then afforded the opportunity to evaluate the output derived from LLMs and adjust if deemed necessary. **c,** LLMs have been utilized from text processing to disease management in the field of medicine. In textual information handling, LLMs could be utilized in automatic medical record writing to improve the efficiency of clinicians and radiology report structuring to extract valuable results for disease evaluation. As for disease management, LLMs could provide recommendations from diagnosis to prognosis. RAG, retrieval-augmented generation.

acquired during pretraining but also acquire specialized skills and insights that are relevant to certain fields, such as oncology.[24] For instance, the initial pretraining phase provides the model with a broad understanding of medical concepts, terminology, and general practices across various specialities. However, in the context of oncology, fine-tuning allows for the infusion of more specialized knowledge regarding cancer diagnoses, treatment protocols, and the complexities of tumour biology. By utilizing curated datasets that include recent research findings, clinical guidelines, and detailed case studies specific to oncology, the model becomes adept at recognizing and interpreting clinical information that is pertinent to cancer patients.[25]

Dynamic customization via prompting

Compared with traditional joint or pipelined NLP-based information extraction schemes, LLMs can respond rapidly to users and align with human preferences via prompt engineering or in-context learning without modifying their model parameters. This is done because pretraining or fine-tuning a model is expensive and time-consuming.[26] With instruction tuning or reinforcement learning from human feedback (RLHF), LLMs are capable of aligning with human intentions, whereas pre-trained smaller models or traditional machine learning pipelines are confined to supervised learning datasets rather than catering to broad demands.[23,27]

Normalizing clinical data via natural language understanding

Medical information extraction is a significant data standardization technique that aims to identify named entities and extract the relations between them.[28] Given an electronic health record (EHR) or clinical report, information extraction is dedicated to extracting triples (entity$_1$, relation, entity$_2$), which are composed of entities and their particular relations. The entities can be symptoms or lesion locations, and the relationships can be descriptive (e.g., relating a biomarker to its results).[29] A variety of downstream tasks can be accomplished with these triples. Moreover, through information extraction, unstructured medical data can be normalized to a uniform format, thereby preventing the processed data from being synonymous or ambiguous content.

Improving the inference capacity via chain of thought

Through chain of thought (CoT), which is an emerging model scale property that allows LLMs to conduct reasoning tasks that would otherwise have flat scaling curves, an LLM can better simulate human thought processes, thereby exhibiting higher levels of intelligence across various application scenarios.[30] The construction of such thought chains enables a model to interconnect different concepts and pieces of information, forming a comprehensive cognitive framework whose coherence not only enhances the ability of the model to comprehend problems but also increases its efficiency and accuracy with respect to generating text, making inferences, and tackling complex tasks.

Domain knowledge enhancement via external knowledge

External knowledge enables LLMs to generate more precise content after enriching their domain knowledge. Retrieval-augmented generation (RAG), which is a process that is composed of a retriever, a repository, and an LLM generator, is a novel approach for enhancing the capabilities of LLMs by accessing an external KB.[31] The repository can be structured, unstructured, or semi-structured, encompassing a vast quantity of medical knowledge derived from sources such as PubMed, MIMIC III, StatPearls, medical textbooks, knowledge graphs, or encyclopaedias. Although the training corpus used for the pretraining or fine-tuning stages may include the abovementioned repositories, it might encounter an obstacle, as it would need to undergo continuous training to update the knowledge of the model, which is both time-consuming and expensive. However, adopting RAG in the inference stage prevents LLMs from undergoing training while still utilizing external knowledge. Notably, an adversary can exploit the instruction-following capabilities of models to easily extract text data verbatim from the data of RAG systems built with instruction-tuned LMs via prompt injection.[32] Thus, a strict safety evaluation system is crucial for keeping private patient data confidential from an ethical or legitimate perspective.

Role of LLMs in medicine

In addition to the abovementioned capacities, LLMs play a useful role in the medical field. During clinical practice, a wealth of textual information is generated from patients along with their disease trajectories, and the potential of LLMs to process these data has been explored.[33,34] Previous evidence has provided examples of LLMs being used in specific scenarios, such as medical record writing, discharge summary generation, and imaging report structuring.[35,36] These studies have confirmed that the application of LLMs can alleviate the workloads imposed on physicians. In disease management scenarios, LLMs enable novel clinical decision-making support schemes ranging from diagnosis to prognosis prediction by learning a large amount of patient data and medical knowledge, providing valuable healthcare insights when guided by prompts.[37–39] In addition, the medication direction copilot (MEDIC), which was proposed to decrease the number of medication errors, can assist in improving the accuracy of online pharmacies.[40] In a certain application scenario, a site-specific prompt engineering chatbot (SSPEC) was developed to aid outpatient navigation, providing inspiration for future research on the practicability of LLMs in clinical settings.[41]

Application of LLMs in oncology

Regarding specialized oncology tasks, previous evidence has provided certain use cases involving various malignancies, such as lung cancer, breast cancer, and colorectal cancer (CRC). Here, we introduce the specific application of LLMs related to tumours from five perspectives, namely, cancer screening and diagnosis, metastasis identification, tumour staging, treatment suggestion, and documentation processing; related studies are listed in Table 1.[42–61] Through these workflows, LLMs can potentially drive precise malignancy management (Fig. 3).

Cancer screening and diagnosis

Early detection of cancer is the key for optimal intervention planning. With the aim of identifying the pulmonary nodules at high lung cancer risk, a multicenter study utilized GPT-4o to follow the radiological changes on longitudinal CT images that were converted into videos, and the model prediction agreed well with the radiologists.[42] As for breast cancer screening and prevention tasks, a team designed 25 questions according to the Breast Imaging Reporting and Data System (BI-RADS) and clinical experience to test the ability of ChatGPT to manage this malignancy; the accuracy of the model was 88%.[43] A similar study designed 22 key questions for prostate cancer (PCa) based on guidelines and practical experience. Five LLMs, ChatGPT, YouChat, NeevaAI, Perplexity, and Chatsonic, were included for the model comparison. Three of them, ChatGPT, YouChat, and Perplexity, achieved average accuracies over 90% in terms of answering all questions about PCa, whereas ChatGPT performed best, with 100% correctness. In particular, the interpretations of different prostate-specific antigen (PSA) levels generated by LLMs are of potential significance for screening efficiency

Year	Cancer types	Tasks	Reference standard	Materials	Data size	Models	Performance	Ref
2025	Lung cancer	Screening	Pathological results	CT images	647 patients	GPT-4o	Accuracy of 0.88	42
2023	Breast cancer	Screening and prevention	Three radiologists	Designed questions	25 questions	ChatGPT	Appropriate answering of 88% questions	43
2023	Prostate cancer	Screening, prevention, and treatment recommendation	Three urologists	Designed questions	22 questions	ChatGPT, YouChat, NeevaAI, Perplexity, and Chatsonic	Accuracy of 75–100%, comprehensiveness of 30–95.45%, readability of 84.21–100%, stability of 63.64–100%	44
2025	Colorectal cancer	Screening	Twenty experts and twenty non-experts	Designed questions	15 questions	ChatGPT	Mean accuracy scores of 4.8 and 5.6 in expert and non-expert assessments	45
2024	Various cancers	Diagnosis and tumour staging	Three pathologists	Pathology reports	1134 reports	Bard and GPT-4	Correctness of 91.77% and 92.75% for Bard and GPT-4 in malignancy diagnosis, correctness of 91.75% and 95.88% for Bard and GPT-4 in pathologic staging	46
2024	Colorectal cancer and breast cancer	Diagnosis and metastasis assessment	NA	Histopathology images	43,100 image patches	GPT-4 with Vision capabilities	Accuracies of 90% and 88.3% for CRC diagnosis and breast cancer metastasis identification	47
2023	Lung cancer	Metastasis assessment	Four radiologists	CT reports	449 reports	ChatGPT and GPT-4	Overall accuracies of 90.3% and 98.1% for metastasis identification of ChatGPT and GPT-4	48
2024	Lung cancer	Tumour staging and histological classification	AJCC guidelines	Pathology reports	852 reports	ChatGPT	Average accuracy of 0.89	49
2025	Thyroid cancer	Tumour staging and ATA risk prediction	AJCC guidelines, ATA risk stratification system, and two endocrine surgeons	Pathology reports	374 patients	Mistral-7B-Instruct, Gemma-2-9B-Instruct, Llama 3.1-8B-Instruct, and Qwen2.5-7B-Instruct	F1-scores up to 98.1% and 95.5% for staging and risk prediction in the validation set by ensemble strategies	50
2024	Pancreatic cancer	Treatment recommendation and report simplification	Three radiologists	CT reports	180 reports	GPT-3.5 and GPT-4	Overall accuracies of 99.9% and 99% for GPT-4 and GPT-3.5 in report simplification; accuracies of 92% and 75% for GPT-4 and GPT-3.5 in tumour resectability categorization	51
2024	Eleven types of cancers	Treatment recommendation	Three radiation oncologists and three radiation physicists	Questions acquired from websites	115 questions	ChatGPT	Correctness of 100%, 91%, and 92% for general questions, treatment modality- and site-specific questions	52
2025	Head-and-neck cancer	Treatment recommendation	NA	CT images and clinical data	2985 patients	GPT-4	Dice similarity coefficient of 0.76	53
2024	Various cancers	Drug sensitivity evaluation	NA	Single-cell RNA sequencing data	NA	Graph-augmented LLM	AUC of 0.975 for treatment response prediction	54
2024	Various cancers	Evaluation of irAEs	Physicians and medical students	Clinical records	8825 records	Mistral OpenOrca	Average sensitivity of 98.1% and specificity of 95.7%	55
2023	Breast cancer	Treatment recommendation and text summarization	Tumour board and two radiologists	Clinical records	10 patients	ChatGPT	Coincidence rate of 70% between the model and tumour board in treatment recommendation, mean scores of 4.3 and 4.6 in clinical record summarization	56
2024	Colorectal cancer	Treatment and examination recommendation	NCCN guidelines	Designed questions	150 questions	Eight chatbots	Accuracies of 61.78%–82.67%	57
2023	Lung cancer, breast cancer, and prostate cancer	Treatment recommendation	NCCN guidelines	Designed prompts	104 prompts	GPT-3.5	Coincidence rate of 61.9% between the model and guidelines	58
2024	Breast cancer	BI-RADS categorization	Three radiologists	MRI, mammography, and ultrasound reports	2400 reports	GPT-3.5, GPT-4, and Bard	Gwet agreement coefficients of 0.42–0.52 between the models and radiologists	59
2024	Lung cancer	Information extraction	A clinician	CT reports	847 reports	ChatGPT	Average accuracy of 0.896–0.940	60
2024	Glioblastoma	Text summarization	Clinical data and tumour board	MRI reports	375 reports	GPT-4	Agreement of 91% between the model and experts	61

AJCC, American Joint Committee on Cancer; ATA, American Thyroid Association; BI-RADS, Breast Imaging Reporting and Data System; ChatGPT, Chat Generative Pre-trained Transformer; CT, computed tomography; GPT, Generative Pre-trained Transformer; irAEs, immune-related adverse events; LLM, large language model; MRI, magnetic resonance imaging; NA, not available; NCCN, National Comprehensive Cancer Network.

***Table 1:* Application of large language models in screening, diagnosis, and treatment of oncology.**

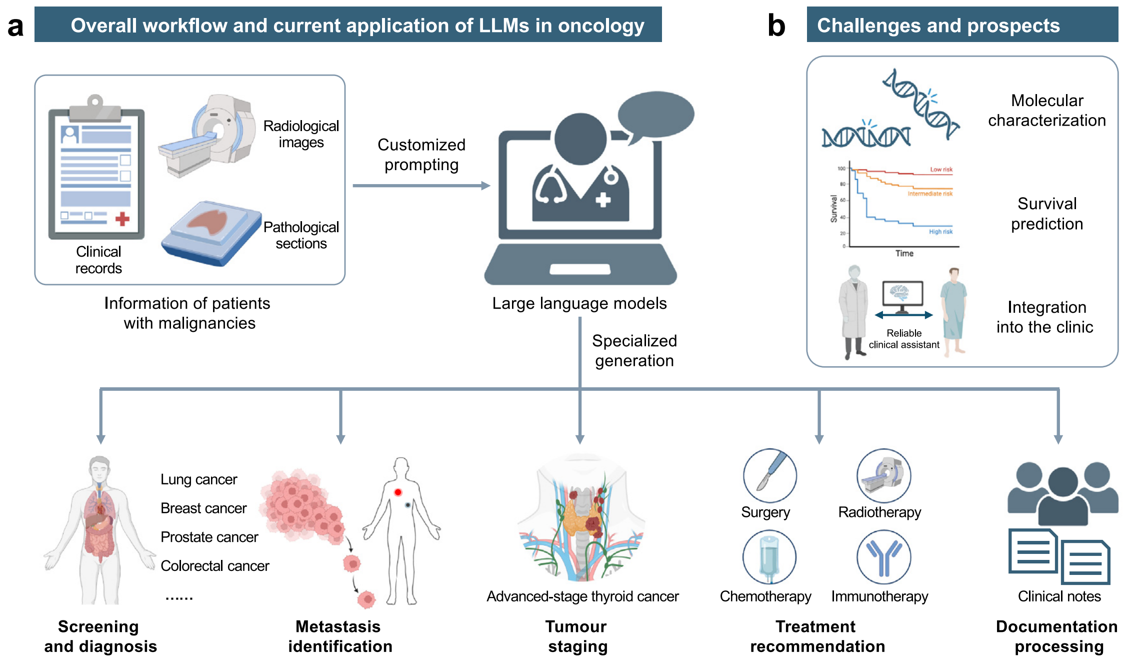

Fig. 3: **Application and potential of large language models (LLMs) in oncology. a,** Through decoding medical data, including clinical text, radiology, and pathology, LLMs have been applied in cancer screening and diagnosis, metastasis prediction, tumour staging, treatment suggestion, and documentation processing. **b,** Certain implementations of LLMs in oncology, such as identification of gene mutation and survival prediction, have not been thoroughly explored. Moreover, validation and generalization of the LLMs need to be carried out before the models can be reliable clinical assistants.

improvement.[44] To improve the screening awareness of CRC, 15 questions, including general screening and endoscopy examination-related queries, were specially designed. After being evaluated by human experts, the ChatGPT achieved a mean accuracy score of 4.8.[45]

Furthermore, not limited to one type of cancer, a team assessed the capabilities of two chatbots, Bard and GPT-4, in diagnosing malignancies from 1134 pathology reports involving different organs, including the breast, lungs, prostate, and colon. These two models both reached diagnosis accuracies surpassing 90%, along with report simplification.[46] In addition, GPT-4 with Vision capabilities (GPT-4V) was adopted in pathology classification tasks conducted on a histopathology dataset of colorectal samples, finally reaching an accuracy of 90% with respect to discriminating between colorectal tumour and nontumorous normal tissues in a ten-shot setting.[47]

Metastasis prediction

Tumour metastasis is a vital cancer progression pathway that may affect the overall prognosis of patients. LLMs are not only capable of assisting physicians in diagnosis tasks but also for predicting cancer metastasis and target organs.[62] A study applied ChatGPT and GPT-4 to extract lesion diameters, identify metastasis, and assess tumour progression via free-text CT reports concerning different types of lung cancer, including adenocarcinoma, squamous cell carcinoma (SCC), and small cell lung cancer (SCLC). The results showed that GPT-4 outperformed ChatGPT in almost all tasks, achieving a respective accuracy of 98.1% and 90.3% in recognition of metastases to different sites, such as pleura, liver, renal, and bone.[48] Moreover, the GPT-4V model mentioned above for sessile-serrated adenoma diagnosis was also trained on histologic sections for lymph-node breast cancer metastasis detection, demonstrating an accuracy of 88.3%.[47] Since CNNs have been deployed to predict the origin of tumours on whole slide images (WSIs),[63] how to convert this realization into LLMs is promising and deserves further research.

Tumour staging

A precise therapeutic schedule for treating a malignancy usually requires a comprehensive and individualized analysis of the condition of the patient and disease. The tumour stages and pathological subtypes of malignancies can significantly affect management regimens. Thus, a study evaluated the performance of ChatGPT-3.5 in the pathological assessment of lung cancer in 774 pathology reports after prompt optimization was conducted within 78 other reports. As a result, the model presented an average accuracy of 0.89 in evaluating the pathological primary tumour (pT), lymph node involvement (pN), the overall tumour stage, and the histology type.[49] Then, for thyroid cancer, another study investigated the power of four LLMs in complete Tumour-Node-Metastasis (TNM) staging and American Thyroid Association (ATA) risk prediction, and the ensemble classifiers achieved F1-scores up to 98.1% and 95.5% in the two tasks, respectively.[50] Moreover, the study that compared the performance of the Bard and GPT-4 models in diagnosing malignancies further investigated the ability of the models to stage various tumours, and the accuracies of the two models were 91.75% and 95.88%, respectively.[46]

Treatment recommendation

In terms of recommending specific treatments, to explore the accuracy of surgical resectability determination for pancreatic ductal adenocarcinoma (PDAC) through radiological information, 180 original reports were utilized to evaluate the capabilities of GPT-4 and GPT-3.5. Ultimately, GPT-4 surpassed GPT-3.5 with a higher accuracy (92% vs. 75%) by CoT prompting, which was superior to other methods that utilized default or in-context knowledge.[51] For radiation oncology, a visual language model named Radformer was developed based on GPT-4 and CNN to automatically delineate the target tumour volume in patients with head-and-neck cancer, reaching a mean Dice similarity coefficient of 0.76.[53] Another study utilized 115 questions derived from oncological websites belonging three thematic categories, consisting of general issues, modalities, and radiotherapy sites, to measure the domain-specific and -agnostic response qualities of ChatGPT-3.5. The results indicated that in the above three tasks, compared with the experts, the model achieved factual correctness rates of 100%, 91%, and 92%. Moreover, in a domain-agnostic metric assessment, ChatGPT-3.5 demonstrated a college-level reading capability.[52] Regarding medication, considering the challenge of drug resistance of anti-tumour treatment, DrugFormer, a graph-enhanced LLM, was developed to predict the drug sensitivity on cellular-level by the drug resistance data extracted from public database, achieving an AUC of 0.975 and providing a reference for treatment planning.[54] In addition, for immunotherapy, another team specifically utilized a RAG-assisted LLM to predict immune-related adverse events (irAEs), including colitis, hepatitis, myocarditis, and pneumonitis, in a variety of malignancies. In the validation set, the model reached an average sensitivity of 98.1% and specificity of 95.7%, surpassing ICD codes and potentially providing indications for immunotherapy monitoring.[55]

To address unspecified treatment scheduling, a team specifically appraised the auxiliary role of ChatGPT-3.5 in clinical decision-making tasks by integrating clinical notes and medical reports from ten patients with breast cancer. The final concordance rate between the machine and human specialists reached a moderate level of 70% in the treatment recommendation. Nevertheless, the data size was small, potentially causing this conclusion to have limited generality.[56] For CRC, a study utilized the National Comprehensive Cancer Network (NCCN) guidelines as a reference standard to compare the response accuracies of eight chatbots, including Claude 2.1, Doctor GPT, ChatGPT-4, and so on. Finally, Claude 2.1 achieved the highest accuracy of 82.67% in the task of responding to questions comprising various domains, such as treatments and imaging or pathology examinations.[57] Another study more comprehensively estimated the treatment suggestion ability of GPT-3.5 for patients with different malignancies including breast cancer, lung cancer, and PCa based on the NCCN guidelines. The model achieved an agreement rate of 61.9% with three reviewers and generated hallucinations in 12.5% of all cases, implying that room for improvement remains in enabling LLMs to provide reliable assistance.[58]

Documentation processing

Since LLMs have achieved comparable performance to that of human experts in clinical textual data summarization tasks, these advanced approaches possess potential for medical documentation processing.[64] In the abovementioned breast cancer-specific investigation, in addition to decision-making, the ability of ChatGPT-3.5 to summarize medical vignettes has been studied. According to two reviewers, ChatGPT made a respective score at 4.3 and 4.6.[56] Another study specifically investigated the ability of LLMs, including GPT-3.5, GPT-4, and Bard, of BI-RADS categorizing through magnetic resonance imaging (MRI), mammography, and ultrasound reports. The final Gwet agreement coefficients between the LLMs and humans reached 0.42–0.52, and these scores were substantially lower than the human–human consensus rate of 0.91. Among different languages, the agreement in English was greater than that in Italian and Dutch.[59] For lung cancer, a study utilized ChatGPT to extract crucial radiological features, including the site, diameter, density, lymphadenectasis, and so on, in a zero-shot setting, and the model achieved an average accuracy of 0.937. Moreover, adding prior medical knowledge to prompts may increase the accuracy of answers to certain questions while decreasing that of other questions.[60] Another study validated the capabilities of GPT-4 and GPT-3.5 of report simplifying from patients with PDAC, reaching respective accuracy of 99.9% and 99% in radiological feature extraction.[51] For patients with glioblastoma, GPT-4 was utilized to perform MRI report summarization by learning information from 375 reports. GPT-4 demonstrated an agreement level of 91% with neuro-oncological experts in disease course representation, showing potential in the tumour monitoring.[61] Therefore, appropriate prompt and task design allows LLMs to assist in medical record management to facilitate clinical workflow.

Challenges

Although certain achievements have been made with respect to incorporating LLMs into oncology, obstacles remain to be overcome in the path of applying LLMs dependably. The current challenges facing LLMs along with possible solutions are summarized as follows.

As the majority of LLMs leverage uninterpretable approaches to make their final predictions, which is called the "black box" effect, it may be difficult to understand how the model makes the final decision. An LLM that imitates realistic reasoning via CoT prompting has been developed, providing examples of the diagnostic rationale.[65] Nevertheless, how to depict flowcharts that are more concise and understandable for visualizing the logics of LLMs is undetermined. Moreover, fact fabrications, also termed hallucinations, are commonplace across LLMs.[66] Because most LLMs undergo pretraining on public databases that have not been authoritatively annotated with reference labels and do not conduct tailored learning, misjudgements or textual mistakes may occasionally occur; this may be a critical factor hindering LLMs from achieving a professional level of accuracy.[67] However, when medical knowledge was added to prompts, performance enhancements were not always observed for all the tasks.[50] This phenomenon potentially implies the difficulty of complex learning and the unavoidable mechanical characteristics of LLMs. Thus, it is necessary to ensure that human specialists hold the final word in decision-making until LLMs reach true human-level thinking and become reliable oncology assistants. Another obstacle that remains to be solved is how to realize real human–computer interactions in the field of malignancies. An emerging study developed the interactive SkinGPT-4 model based on Llama-2-13b-chat for dermatological diagnosis; through this system, users can obtain medical advice in a manner similar to face-to-face communication after uploading their skin pictures.[68] This effort inspires more exploration of integrating chatbots to develop automated response generators that can interact deeply with oncologists or patients.

And LLMs exhibit potential societal biases. For example, demographic diversity in a population could not be captured by GPT-4, and the known associations between diseases and demographics were exaggerated, potentially leading to the recognition of stereotypes in disease management.[69] GPT-4 has been utilized to mine the critical problems of transgender patients facing during the suffering from breast and gynaecological cancers, revealing the challenges of awareness lack and access issues, which provides a novel insight into LLM-aided disease management in more diversified circumstances.[70] Other factors that may cause prejudice are language and culture biases. In the previously mentioned study, the performance of LLMs was significantly better in English than in other languages that are not in public use.[59] And for tasks related to specific cultures, such as Traditional Chinese Medicine (TCM), the models developed by Chinese companies substantially surpassed the models constructed by Western companies, highlighting the cultural barriers of LLMs.[71]

Then, ethical issues have also been widely discussed across various AI applications, such as that accountability remains unclear when diagnostic or therapeutic mishaps are produced.[72] Moreover, the leakage of user data is of great concern.[73] Hence, establishing agencies that possess supervision authority for legally acquiring, utilizing, and storing data, as well as approving medical devices, is necessary before clinically applying any LLM on the market.[74] Notably, for scientific investigators, exploring the potential of LLMs in literature retrieval, information mining, and data analysis tasks may help accelerate the progress of medical knowledge acquisition. However, humankind is still responsible for the overall research, and LLMs should serve only as auxiliary tools rather than real authors.[75–77] This point emphasizes the notion that the appropriate role of LLMs is augmenting rather than replacing the cognitive processes of human beings, and the harms of over-relying upon AI surely outweigh the benefits.[78,79]

Finally, in terms of the overall healthcare of patients with malignancies, there is still a long pathway before LLMs can play an all-around role. Recently, a chatbot for cancer proteomics analysis has been proposed, which implies that molecular study based on LLMs is of potential benefit to accelerate carcinogenesis mechanism exploration.[80] Moreover, another study investigated the capability of GPT-4o mini in recognizing the causal relationships between malignancies and genes; however, how to directly identify the gene mutation in patients remains to be studied.[81] In addition, survival and prognosis prediction which serve as critical components of individualized care in oncology, have not yet been thoroughly studied. Another consideration is that most of the studies described herein were more akin to explorations than real clinical applications; thus, more comprehensive research and large-scale multicenter prospective validations of these LLMs, as well as the implementation of a proper resource allocation scheme especially in backwards regions, to make the chatbots become real clinical assistants, are required. In addition, as medical knowledge and research develop with each passing day, how to keep the intelligence of LLMs at the frontier of

Search strategy and selection criteria.

We conducted a comprehensive literature search in PubMed and the Web of Science via the key search terms "large language model" or "GPT" and "oncology" or "cancer" or "tumour" or "malignancy". The preliminarily selected studies underwent careful full-text checks to ensure that the final studies described herein were qualified and relevant. Most of the included studies were published in the last five years.

oncological progress is a problem worthy of deep reflection.[82] Ensuring the dynamics and accuracy of the information acquisition of LLMs is of vital importance to improve the credibility of the model output.[83]

Conclusion

In summary, the innovation of LLMs has remodelled the conventional paradigm of language processing and is swiftly blossoming in many realms. In oncology, LLMs provide opportunities to promote precision oncology and alleviate the health burden. However, the technical barriers and application limitations of LLMs impact their feasibility in real-world clinical settings. Therefore, to address these bottlenecks, future in-depth research and strict evaluations of LLMs could be conducive to realizing personalized health care guidance for malignancy.

Outstanding questions

The emergence of LLMs has inspired the exploitation of research and applications in various areas. In oncology, LLM-related studies have achieved moderate to profound performance. Going forward, it is essential to consider issues such as accuracy improvement and scope expansion to develop an automated system that is applicable to the whole process of precise cancer management. In addition, researchers must deeply reflect on the clinical translations and realistic implementations of the new technology.

Contributors

Chengdi Wang, Guangyu Wang, and Weimin Li conceptualized this study. Shufan Liang, Jiangjiang Zhang, Xingting Liu, Yinkui Huang, Jun Shao, and Xiaohong Liu searched the literature and drafted the original version of the manuscript. Shufan Liang, Jiangjiang Zhang, and Xingting Liu designed the figures and table. Chengdi Wang, Guangyu Wang, and Weimin Li revised the final submission. All authors read and approved the final manuscript.

Declaration of interests

The authors declare no competing interests.

Acknowledgements

The study was supported by the Noncommunicable Chronic Diseases-National Science and Technology Major Project of China (2024ZD0529504/2024ZD0529500), National Natural Science Foundation of China (82470109, 62272055, 92159302), Science and Technology Project of Sichuan (2022ZDZX0018), New Cornerstone Science Foundation through the XPLORER PRIZE, 1.3.5 Project for Disciplines of Excellence, West China Hospital, Sichuan University (ZYYC23027), 1·3·5 projects for Artificial Intelligence, West China Hospital, Sichuan University (ZYAI24016), and 1·3·5 Project of State Key Laboratory of Respiratory Health and Multimorbidity, West China Hospital, Sichuan University (RHM24208). We acknowledged the BioRender.com for the support of illustration design. The funders were not involved in paper design, data collection, data analysis, interpretation, writing of the paper.

References

1 Haug CJ, Drazen JM. Artificial intelligence and machine learning in clinical medicine, 2023. *N Engl J Med*. 2023;388(13):1201–1208.
2 Xu H, Usuyama N, Bagga J, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*. 2024;630(8015):181–188.
3 Wang J, Wang K, Yu Y, et al. Self-improving generative foundation model for synthetic medical image generation and clinical applications. *Nat Med*. 2024;31(2):609–617.
4 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436–444.
5 Wang C, Ma J, Zhang S, et al. Development and validation of an abnormality-derived deep-learning diagnostic system for major respiratory diseases. *NPJ Digit Med*. 2022;5(1):124.
6 Zhou HY, Yu Y, Wang C, et al. A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. *Nat Biomed Eng*. 2023;7(6):743–755.
7 Shao J, Ma J, Yu Y, et al. A multimodal integration pipeline for accurate diagnosis, pathogen identification, and prognosis prediction of pulmonary infections. *Innovation (Camb)*. 2024;5(4):100648.
8 Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316(22):2402–2410.
9 Hollon TC, Pandian B, Adapa AR, et al. Near real-time intraoperative brain tumor diagnosis using stimulated Raman histology and deep neural networks. *Nat Med*. 2020;26(1):52–58.
10 Zhou L, Wang J. An investigation into the applicability of rapid artificial intelligence-assisted compressed sensing in brain magnetic resonance imaging performed at 5 Tesla field strength. *iRADIOLOGY*. 2024;2(5):1–10.
11 Siegel RL, Giaquinto AN, Jemal A. Cancer statistics, 2024. *CA Cancer J Clin*. 2024;74(1):12–49.
12 Liang W, He J, Zhong N. Towards zero lung cancer. *Chin Med J Pulm Crit Care Med*. 2023;1(4):195–197.
13 Bhinder B, Gilvary C, Madhukar NS, Elemento O. Artificial intelligence in cancer research and precision medicine. *Cancer Discov*. 2021;11(4):900–915.
14 Stokel-Walker C, Van Noorden R. What ChatGPT and generative AI mean for science. *Nature*. 2023;614(7947):214–216.
15 Conroy G, Mallapaty S. How China created AI model DeepSeek and shocked the world. *Nature*. 2025;638(8050):300–301.
16 Betzler BK, Chen H, Cheng CY, et al. Large language models and their impact in ophthalmology. *Lancet Digit Health*. 2023;5(12):e917–e924.
17 Kaplan J, McCandlish S, Henighan T, et al. Scaling laws for neural language models. *ArXiv*; 2020. https://arxiv.org/abs/2001.08361 (preprint).
18 Nishino M, Ballard DH. Multimodal large language models to solve image-based diagnostic challenges: the next big wave is already here. *Radiology*. 2024;312(1):e241379.
19 Lu MY, Chen B, Williamson DFK, et al. A multimodal generative AI copilot for human pathology. *Nature*. 2024;634(8033):466–473.
20 Achiam OJ, Adler S, Agarwal S, et al. GPT-4 technical report. *ArXiv*; 2023. https://arxiv.org/abs/2303.08774 (preprint).
21 Radford A, Narasimhan K. *Improving language understanding by generative pre-training*. 2018.
22 Sung M, Lee J, Yi SS, et al. Can language models be biomedical knowledge bases?. In: *Proceedings of the 2021 conference on empirical methods in natural language processing*. 2021.
23 Ouyang L, Wu J, Jiang X, et al. *Training language models to follow instructions with human feedback*. 2022. Proceedings of the 36th International Conference on Neural Information Processing Systems.
24 Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence. *Nature*. 2023;616(7956):259–265.
25 Han T, Adams LC, Papaioannou J-M, et al. MedAlpaca - an open-source collection of medical conversational AI models and training data. *ArXiv*; 2023. https://arxiv.org/abs/2304.08247 (preprint).
26 Brown TB, Mann B, Ryder N, et al. *Language models are few-shot learners*. 2020. Proceedings of the 34th International Conference on Neural Information Processing Systems.
27 Wei J, Bosma M, Zhao V, et al. Finetuned language models are zero-shot learners. *ArXiv*; 2021. https://arxiv.org/abs/2109.01652 (preprint).
28 Juric D, Stoilos G, Melo A, Moore J, Khodadadi M. A system for medical information extraction and verification from unstructured text. In: *Proceedings of the AAAI conference on artificial intelligence*. 2020.
29 Sushil M, Kennedy VE, Mandair D, et al. CORAL: expert-curated oncology reports to advance language model inference. *NEJM AI*. 2024;1(4):AIdbp2300110.
30 Wei J, Wang X, Schuurmans D, et al. *Chain-of-thought prompting elicits reasoning in large language models*. 2022. Proceedings of the 36th International Conference on Neural Information Processing Systems.
31 Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: *Proceedings of the 34th international conference on neural information processing systems*. 2020.

32 Qi Z, Zhang H, Xing E, Kakade SM, Lakkaraju H. Follow my instruction and spill the beans: scalable data extraction from retrieval-augmented generation systems. *ArXiv*; 2024. https://arxiv.org/abs/2402.17840 (preprint).
33 Bhayana R, Alwahbi O, Ladak AM, et al. Leveraging large language models to generate clinical histories for oncologic imaging requisitions. *Radiology*. 2025;314(2):e242134.
34 Ding H, Xia W, Zhou Y, et al. Evaluation and practical application of prompt-driven ChatGPTs for EMR generation. *NPJ Digit Med*. 2025;8(1):77.
35 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(4):e230725.
36 Ali SR, Dobbs TD, Hutchings HA, Whitaker IS. Using ChatGPT to write patient clinic letters. *Lancet Digit Health*. 2023;5(4):e179–e181.
37 Liu X, Liu H, Yang G, et al. A generalist medical language model for disease diagnosis assistance. *Nat Med*. 2025;31(3):932–942.
38 Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172–180.
39 Jiang LY, Liu XC, Nejatian NP, et al. Health system-scale language models are all-purpose prediction engines. *Nature*. 2023;619(7969):357–362.
40 Pais C, Liu J, Voigt R, et al. Large language models for preventing medication direction errors in online pharmacies. *Nat Med*. 2024;30(6):1574–1582.
41 Wan P, Huang Z, Tang W, et al. Outpatient reception via collaboration between nurses and a large language model: a randomized controlled trial. *Nat Med*. 2024;30(10):2878–2885.
42 Mao Y, Xu N, Wu Y, et al. Assessments of lung nodules by an artificial intelligence chatbot using longitudinal CT images. *Cell Rep Med*. 2025;101988.
43 Haver HL, Ambinder EB, Bahl M, et al. Appropriateness of breast cancer prevention and screening recommendations provided by ChatGPT. *Radiology*. 2023;307(4):e230424.
44 Zhu L, Mou W, Chen R. Can the ChatGPT and other large language models with internet-connected database solve the questions and concerns of patient with prostate cancer and help democratize medical knowledge? *J Transl Med*. 2023;21(1):269.
45 Maida M, Ramai D, Mori Y, et al. The role of generative language systems in increasing patient awareness of colon cancer screening. *Endoscopy*. 2025;57(3):262–268.
46 Steimetz E, Minkowitz J, Gabutan EC, et al. Use of artificial intelligence Chatbots in interpretation of pathology reports. *JAMA Netw Open*. 2024;7(5):e2412767.
47 Ferber D, Wölflein G, Wiest IC, et al. In-context learning enables multimodal large language models to classify cancer pathology images. *Nat Commun*. 2024;15(1):10104.
48 Fink MA, Bischoff A, Fink CA, et al. Potential of ChatGPT and GPT-4 for data mining of free-text CT reports on lung cancer. *Radiology*. 2023;308(3):e231362.
49 Huang J, Yang DM, Rong R, et al. A critical assessment of using ChatGPT for extracting structured data from clinical notes. *NPJ Digit Med*. 2024;7(1):106.
50 Fung MMH, Tang EHM, Wu T, et al. Developing a named entity framework for thyroid cancer staging and risk level classification using large language models. *NPJ Digit Med*. 2025;8(1):134.
51 Bhayana R, Nanda B, Dehkharghanian T, et al. Large language models for automated synoptic reports and resectability categorization in pancreatic cancer. *Radiology*. 2024;311(3):e233117.
52 Yalamanchili A, Sengupta B, Song J, et al. Quality of large language model responses to radiation oncology patient care questions. *JAMA Netw Open*. 2024;7(4):e244630.
53 Rajendran P, Yang Y, Niedermayr TR, et al. Large language model-augmented learning for auto-delineation of treatment targets in head-and-neck cancer radiotherapy. *Radiother Oncol*. 2025;205:110740.
54 Liu X, Wang Q, Zhou M, et al. DrugFormer: graph-enhanced language model to predict drug sensitivity. *Adv Sci (Weinh)*. 2024;11(40):e2405861.
55 Sun VH, Heemelaar JC, Hadzic I, et al. Enhancing precision in detecting severe immune-related adverse events: comparative analysis of large language models and International Classification of Disease codes in patient records. *J Clin Oncol*. 2024;42(35):4134–4144.
56 Sorin V, Klang E, Sklair-Levy M, et al. Large language model (ChatGPT) as a support tool for breast tumor board. *NPJ Breast Cancer*. 2023;9(1):44.
57 Zhou S, Luo X, Chen C, et al. The performance of large language model powered chatbots compared to oncology physicians on colorectal cancer queries. *Int J Surg*. 2024;110(10):6509–6517.
58 Chen S, Kann BH, Foote MB, et al. Use of artificial intelligence chatbots for cancer treatment information. *JAMA Oncol*. 2023;9(10):1459–1462.
59 Cozzi A, Pinker K, Hidber A, et al. BI-RADS category assignments by GPT-3.5, GPT-4, and Google Bard: a multilanguage study. *Radiology*. 2024;311(1):e232133.
60 Hu D, Liu B, Zhu X, Lu X, Wu N. Zero-shot information extraction from radiological reports using ChatGPT. *Int J Med Inform*. 2024;183:105321.
61 Laukamp KR, Terzis RA, Werner JM, et al. Monitoring patients with glioblastoma by using a large language model: accurate summarization of radiology reports with GPT-4. *Radiology*. 2024;312(1):e232640.
62 Webster P. Six ways large language models are changing healthcare. *Nat Med*. 2023;29(12):2969–2971.
63 Lu MY, Chen TY, Williamson DFK, et al. AI-based pathology predicts origins for cancers of unknown primary. *Nature*. 2021;594(7861):106–110.
64 Van Veen D, Van Uden C, Blankemeier L, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nat Med*. 2024;30(4):1134–1142.
65 Savage T, Nayak A, Gallo R, Rangan E, Chen JH. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *npj Digit Med*. 2024;7(1):20.
66 Menz BD, Modi ND, Abuhelwa AY, et al. Generative AI chatbots for reliable cancer information: evaluating web-search, multilingual, and reference capabilities of emerging large language models. *Eur J Cancer*. 2025;218:115274.
67 Bhayana R. Chatbots and large language models in radiology: a practical primer for clinical and research applications. *Radiology*. 2024;310(1):e232756.
68 Zhou J, He X, Sun L, et al. Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4. *Nat Commun*. 2024;15(1):5649.
69 Zack T, Lehman E, Suzgun M, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *Lancet Digit Health*. 2024;6(1):e12–e22.
70 Annan A, Li Y, Du J, et al. Using AI and social media to understand health disparities for transgender cancer care. *JAMA Netw Open*. 2024;7(8):e2429792.
71 Zhu L, Mou W, Lai Y, Lin J, Luo P. Language and cultural bias in AI: comparing the performance of large language models developed in different countries on Traditional Chinese Medicine highlights the need for localized models. *J Transl Med*. 2024;22(1):319.
72 Rengers TA, Thiels CA, Salehinejad H. Academic surgery in the era of large language models: a review. *JAMA Surg*. 2024;159(4):445–450.
73 Ong JCL, Chang SY, William W, et al. Ethical and regulatory challenges of large language models in medicine. *Lancet Digit Health*. 2024;6(6):e428–e432.
74 Gilbert S, Harvey H, Melvin T, Vollebregt E, Wicks P. Large language model AI chatbots require approval as medical devices. *Nat Med*. 2023;29(10):2396–2398.
75 Thorp HH. ChatGPT is fun, but not an author. *Science*. 2023;379(6630):313.
76 Stokel-Walker C. ChatGPT listed as author on research papers: many scientists disapprove. *Nature*. 2023;613(7945):620–621.
77 Moy L. Guidelines for use of large language models by authors, reviewers, and editors: considerations for imaging journals. *Radiology*. 2023;309(1):e239024.
78 Harrer S. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. *eBioMedicine*. 2023;90:104512.
79 Zhu L, Lai Y, Mou W, et al. ChatGPT's ability to generate realistic experimental images poses a new challenge to academic integrity. *J Hematol Oncol*. 2024;17(1):27.
80 Liu W, Li J, Tang Y, et al. DrBioRight 2.0: an LLM-powered bioinformatics chatbot for large-scale cancer functional proteomics analysis. *Nat Commun*. 2025;16(1):2256.
81 Zeng H, Yin C, Chai C, et al. Cancer gene identification through integrating causal prompting large language model with omics data-driven causal inference. *Brief Bioinform*. 2025;26(2).
82 Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. *Nat Med*. 2023;29(8):1930–1940.
83 Omar M, Ullanat V, Loda M, Marchionni L, Umeton R. ChatGPT for digital pathology research. *Lancet Digit Health*. 2024;6(8):e595–e600.