Artificial Intelligence Review (2024) 57:299
https://doi.org/10.1007/s10462-024-10921-0

Large language models in medical and healthcare fields: applications, advances, and challenges

Dandan Wang[1] · **Shiqing Zhang**[1]

Accepted: 22 August 2024 / Published online: 20 September 2024

Abstract

Large language models (LLMs) are increasingly recognized for their advanced language capabilities, offering significant assistance in diverse areas like medical communication, patient data optimization, and surgical planning. Our survey meticulously searched for papers with keywords such as “medical,” “clinical,” “healthcare,” and “LLMs” across various databases, including ACM and Google Scholar. It sought to delve into the latest trends and applications of LLMs in healthcare, analyzing 175 relevant publications to support both practitioners and researchers in the field. We have compiled 56 experimental datasets, various evaluation methods and reviewed cutting-edge LLMs across tasks. Our comprehensive analysis of LLMs in healthcare applications, including medical question-answering, dialogue summarization, electronic health record generation, scientific research, medical education, medical product safety monitoring, clinical health reasoning, and clinical decision support. Furthermore, we have identified the challenges, including data security, inaccurate information, fairness and bias, plagiarism, copyrights, and accountability, and the potential solutions, namely de-identification framework, references,counterfactually fair prompting,opening and ending control codes, and establishing normative standards,to address these open issues,respectively. The findings of this survey exert a profound impact on spurring innovation in practical applications and addressing inherent challenges within the academic and medical communities.

Keywords Large language models · Healthcare applications · Medical question-answering · Electronic health record generation · Clinical health reasoning · Fairness and bias

✉ Shiqing Zhang
tzczsq@163.com

Dandan Wang
dandanw0707@tzc.edu.cn

[1] Department of Computer Science, Taizhou University, Taizhou 318000, Zhejiang, China

1 Introduction

Recent advancements in large language models (LLMs) have garnered significant interest in both academic and industrial domains due to their impressive success in language understanding and text generation (Chang et al. 2023). The extraordinary capabilities of contemporary LLMs hold great promise for applications in the medical and healthcare domain, surpassing the performance of smaller models with limited data (Sallam 2023a). Large language models in the medical and healthcare domain (LLMMs) have the potential to facilitate communication among healthcare professionals, patients, and their families, streamline the collection and analysis of patient health data, and assist in the development of surgical plans (Zhao et al. 2023). Furthermore, LLMMs can acquire real-time surgical navigation information and physiological parameters, offer postoperative rehabilitation guidance to patients, and provide intraoperative support to surgeons. LLMMs can also be trained to recognize and analyze medical images (e.g., X-rays, magnetic resonance imaging, and ultrasound), video, audio and remote photoplethysmograph signals to identify features and structures(Fan et al. 2024), aiding doctors in accurately and rapidly detecting anomalies and diagnosing diseases or injuries, thereby alleviating the workload of radiologists (Waisberg et al. 2023).

The recent swift advancement of LLMMs has opened up a wide range of application prospects across various scientific research fields (Archana and Jeevaraj 2024). These include aiding in the composition of influential articles through literature review synthesis (Chen and Li 2023), facilitating the retrieval and discovery of the latest scientific developments, supporting grammar correction and text translation (Fatani 2023), offering novel perspectives and research directions (Liebrenz et al. 2023), and providing feedback and improvement suggestions for draft or manuscript inputs (Castellanos-Gomez 2023). Additionally, LLMMs are demonstrating their prowess in data analysis and interpretation. In the realm of medical education, LLMMs have exhibited remarkable performance, for instance, in assisting human learners and educators with United States Medical Licensing Examination (USMLE), Japanese Medical Licensing Examination (JMLE), and other medical licensing examinations, as well as in generating and evaluating multiple-choice tests (Sallam 2023b). Although scholars and medical practitioners have increasingly expressed interest in the application of LLMMs, the practical utility of LLMMs in clinical and research settings is fraught with distinct challenges. These challenges include data security, privacy preservation, the risk of inaccurate information, fairness and bias issues, plagiarism concerns, copyright considerations, and accountability.

This paper offers a comprehensive guide for medical researchers and enthusiasts in the LLMM field, providing a swift introduction to the applications of LLMMs in various medical domains, accessible experimental databases, the performance of different models across tasks, current challenges confronting LLMMs, and potential solutions. It holds significant importance in accelerating the enhancement of LLMMs in artificial intelligence technology for healthcare and their capacity to address real-world problems. Furthermore, it plays a crucial role in promoting the swift deployment of LLMs in practical medical settings and in boosting the efficiency and effectiveness of doctors, patients, and other healthcare professionals in clinical, educational, and research activities. To delineate our work more clearly, we juxtapose the existing reviews of the most advanced LLMMs, outline the process of curating relevant publications, and highlight the contributions of this survey.

1.1 Distinctions of this survey from prior research

The remarkable performance and widespread adoption of LLMs have spurred an extensive body of research in this field. For instance, (Wang et al. 2023a) examined the clinical language understanding capabilities of LLMs in healthcare, Tian et al. (Tian et al. 2023) explored the potential and challenges of ChatGPT in the biomedical and health sectors, Liu et al. (Liu et al. 2023a) provided an examination of ChatGPT and GPT-4 across diverse domains, and Sallam et al. (Sallam 2023c) assessed the applicability of ChatGPT in healthcare. Nonetheless, these studies often concentrate on a limited number of LLMs or fail to offer a thorough and expansive analysis of LLM applications and the associated potential issues, such as medical dialog summarization, scientific research, medical product safety monitoring, disease diagnosis, clinical decision support, administrative tasks assistance, and ethical concerns regarding data security and privacy preservation. In contrast, our survey delves deeply into LLMs within the medical and healthcare realm, encompassing research scenarios, accessible medical datasets, evaluation methodologies, and the challenges LLMs encounter in the medical field. Table 1 delineates the disparities between our survey and previous studies.

1.2 Methodology for collecting relevant publications

This study was conducted by searching for relevant papers on the renowned digital library Google Scholar, which encompasses literature from ACM, Springer, Elsevier, arXiv, medRxiv, and other multi-source databases. The focus of this paper is on research work from January 2022 up until the submission of this paper in January 2024, with a particular emphasis on models introduced after the launch of ChatGPT in November 2022, as well as those with parameters exceeding 10^9. For the search terms used on Google Scholar, we adopted a combination of application domains and large models, such as ("medical" or "clinical" or "healthcare") and ("large language model"). The "or" operator allows for the inclusion of papers that meet any of the connected search terms, while the "and" operator requires that all connected terms be satisfied. We found that this approach yielded a significant number of survey and overview articles, with only a few papers containing relevant models. Consequently, we replaced the keyword "large language model" with specific names of large language models (e.g., ChatGPT, LLaMA, PaLM, etc.) and combined them with "medical", "clinical", and "healthcare". The specific names of the large language models were referenced from (Zhao et al. 2023), (Hadi et al. 2023), and others. The rationale for this approach is that, as specialized models in the medical and healthcare fields, they are generally pre-trained from general large models or generated for fine-tuning, making it reasonable and efficient to refer to existing large models. All the papers retrieved from the above searches were included in the candidate corpus. We then reviewed the titles, abstracts, and keywords of the candidate papers; those that did not include specific large models in the medical or healthcare fields were excluded. Additionally, to ensure a more comprehensive coverage of research content, we also included articles discussing LLMs in the healthcare domain from multiple review papers in the candidate set, ultimately arriving at 175 papers closely related to our research.

To our understanding, our survey is the first to focus on LLMMs in relation to real-world application scenarios, available datasets, evaluation methods, and ethical and safety considerations. By delineating the applications of various LLMMs across different fields,

Table 1 The disparities between our survey and previous studies

Main concerns	(Wang et al. 2023a)	(Tian et al. 2023)	.(Liu et al. 2023a)	(Sallam 2023c)	Our survey	Corresponding chapter
Medical question-answering	Few	×	Few	×	√	Section 2.1
Medical dialog sumarization	×	×	×	×	√	Section 2.2
Electronic health records, clinical letters, medical note generation	×	×	×	Few	√	Section 2.3
Scientific research	×	×	×	×	√	Section 2.4
Medical education, language translation	×	√	√	√	√	Section 2.5
Medical imaging recognition, analysis	×	Few	×	×	√	Section 2.6
Clinical health reasoning, diagnostic reasoning	×	Few	×	×		Section 2.7
Medical product safety monitoring, disease diagnosis	×	×	×	×	√	Section 2.8
Clinical decision support, administrative tasks assistance	×	×	×	×	√	Section 2.9
Experimental datasets for LLMMs	Few	Few	Few	Few	√	Section 3
Evaluation methods for LLMMs	Few	Few	×	×	√	Section 4
Data security and privacy-preserving	×	×	×	×	√	Section 6.1
Incorrectness, risk of inaccurate information	×	×	×	Few	√	Section 6.2
Fairness/bias	×	Few	×	×	√	Section 6.3
Transparency, explainability, and trustworthiness	Few	×	×	×	√	Section 6.4
Plagiarism, copyrights, accountability	×	×	Few	Few	√	Section 6.5

we illustrate how LLMMs assist medical professionals, patients, and other healthcare stakeholders in decision-making, answering related questions, and generating electronic health records, as well as in medical education and scientific research. We analyze the latest algorithms and the most appropriate model frameworks in each domain and summarize the challenges, along with potential medical and healthcare solutions.

1.3 Contributions of this survey

This survey systematically delves into the applications of LLMMs, examining their usage in various scenarios, the availability of medical datasets, evaluation methodologies, performance across various tasks, and the challenges they face in the medical field. Our goal is to provide dynamic and constructive guidelines for scientific researchers, practitioners, and developers interested in LLMMs. The primary contributions of this work are as follows:

(1) We comprehensively summarize and provide an overview of the state-of-the-art LLMs across diverse application scenarios within the medical and healthcare fields.
(2) We categorize and analyze the works of publications, integrating various tasks and evaluation metrics to assess the performance of LLMMs.
(3) We thoroughly summarize and categorize the current challenges in the medical and healthcare domains and envision potential solutions to address these open issues.

In the remainder of this paper, Sect. 2 provides an overview of ten common application scenarios of LLMMs. Section 3 introduces several experimental datasets that are most frequently utilized by researchers in the medical and healthcare domains. Section 4 discusses the commonly employed metrics for assessing the performance of LLMMs. Section 5 analyzes the capabilities of state-of-the-art LLMMs across a range of tasks. Section 6 identifies the challenges encountered by LLMMs and offers potential solutions. Finally, Sect. 7 concludes the survey with a summary of the entire work.

2 Application scenarios of state-of-the-art LLMMs

The advanced language comprehension and text generation capabilities of LLMs have significantly impacted various aspects of medical and healthcare scenarios. These applications include medical question-answering, medical dialog summarization, electronic health record generation, scientific research, clinical decision support, and more (as depicted in Fig. 1). The deployment of LLMs offers valuable insights for various stakeholders in healthcare domains, such as healthcare providers and patients (Jin and Dobry 2023)). This includes enhancing patient education, drafting responses, or querying patient notes with given questions for healthcare providers, reviewing scientific papers for researchers, and explaining clinical research protocols for clinical research coordinators.

2.1 Medical question-answering (MQA)

The robust text analysis and comprehension capabilities of LLMs have accelerated their widespread application in answering biomedical and genetic questions, as well as in USMLE, with models like GPT-4 (Wang et al. 2023b), ChatGPT (Javaid et al. 2023),

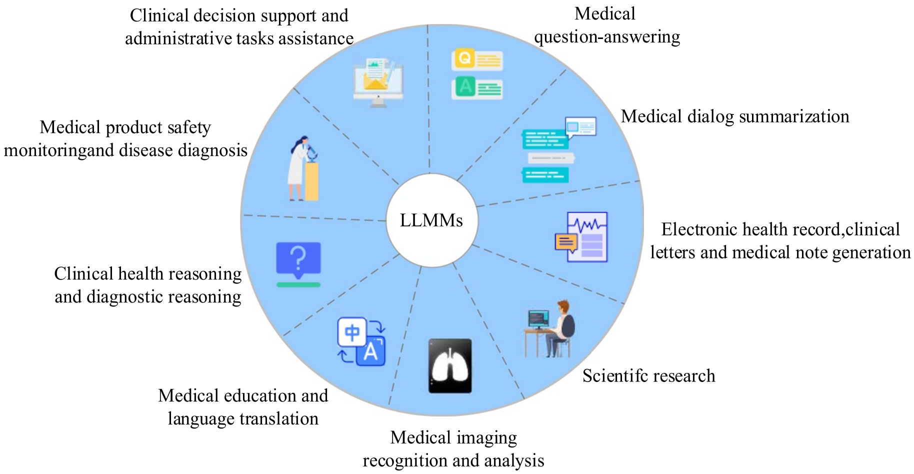
Clinical decision support and administrative tasks assistance
Medical question-answering
Medical dialog summarization
Medical product safety monitoringand disease diagnosis
LLMMs
Electronic health record,clinical letters and medical note generation
Clinical health reasoning and diagnostic reasoning
Scientifc research
Medical education and language translation
Medical imaging recognition and analysis

Fig. 1 The current application scenarios of LLMMs

LLaMA (Yunxiang et al. 2023), PMC-LLaMA (Wu et al. 2023), MedPaLM (George et al. 2023), MedPaLM2 (George et al. 2023), T5 and BERT (Wei et al. 2023). Nanayakkara et al. (Nanayakkara et al. 2022) introduced a seq2seq learning approach based on T5 and BERT models for automatic speech recognition and transcription error correction in clinical dialogues between practitioners and patients. Wu et al. (Wu et al. 2023) proposed the PMC-LLaMA open-source language model, which was fine-tuned by learning 4.8 million biomedical academic papers to enhance the accuracy of question answers in the biomedical field and to better understand specific concepts. MedPaLM (Singhal et al. 2023a) was the first LLMM model to pass the USMLE exam, and Med-PaLM 2 (Singhal et al. 2023b) based on PaLM 2 fine-tuned with medical domain knowledge, introducing a new integration method to provide a prompt strategy. The accuracy of Med-PaLM 2 on the MedQA dataset was 19% higher than that of Med-PaLM, achieving better performance compared to Med-PaLM in answering medical questions.

The application of LLMMs is constrained by their limited medical domain knowledge and the complexities of clinical tasks. For example, the performance of ChatGPT with human respondents in answering genetic questions was not significantly different from human respondents (Duong and Solomon 2023). Given ChatGPT's observed limitations in medical knowledge, Li et al. (Yunxiang et al. 2023) introduced ChatDoctor, employing the LLaMA model with an autonomous information retrieval mechanism. This allows real-time access and utilization of Wikipedia online resources, leading to a substantial enhancement in the quality of patient-physician interactive dialogue. The system has demonstrated notable progress in comprehending patient needs and offering precise treatment options. Toma et al. (Toma et al. 2023) developed Clinical Camel, a dialogue-based knowledge encoding model that enhances the model's implicit knowledge base, maintains session recall, and expands the knowledge base data. As a result, Clinical Camel achieved a higher score than GPT-3.5 on the USMLE test. The model is capable of managing multi-stage clinical case issues, offering adaptive patient counseling, and generating clinical records from conversations (Selvaraj and Konam 2020). Chervenak et al. (Chervenak et al. 2023) conducted a survey on 17 common questions and reproductive knowledge related to infertility using GPT-4 based on existing clinical information. Common questions, surveys, and summaries were used as prompts to input GPT-4, including sentiment analysis, factual statements, published population data, etc. The common issues of infertility, factual content, emotional polarity, and subjectivity were consistent with the management of disease control centers. The experiment of ChatGPT-4 showed that the output information of LLMs is relevant and meaningful for clinical queries related to fertility.

However, since most LLMs are trained and learned from English corpora, advanced LLMs do not perform well in Chinese medical question-answering systems. To address this, several scholars have made efforts in the development and application of Chinese LLMs and datasets, such as BenTsao (Wang et al. 2023c), Ziya-LLaMA (Zhang et al. 2022), DoctorGLM (Xiong et al. 2023), Zhongjing (Yang et al. 2023a), and Huatuo (Li et al. 2023a), among others. Xiong et al. (Xiong et al. 2023) developed a large-scale language model, DoctorGLM, trained on a Chinese healthcare database. DoctorGLM incorporates a prompt designer module that extracts relevant keywords from user input, utilizes potential disease names as labels, and generates a description based on the disease knowledge library. Consequently, DoctorGLM can provide users with reliable information, including disease symptoms, diagnosis, treatment, and preventive measures. Yang et al. (Yang et al. 2023a) introduced a Chinese medicine LLM model named Zhongjing, which is based on LLaMA. By employing refined annotation rules and evaluation criteria, the model's proficiency in complex dialogue and active querying was substantially enhanced through feedback reinforcement learning. The architecure of Zhongjing is depicted in Fig. 2.

2.2 Medical dialog summarization (MDS)

The MDS aids clinicians in identifying potential health risks for patients and supports informed decision-making (Patel and Lam 2023). By analyzing current patient data, MDS reduces errors and enhances diagnostic precision. The field has seen significant advancements due to the recent progress in LLMs, including BERT (Wei et al. 2023), T5, PEGASUS (Balumuri et al. 2021), BioGPT (Alqahtani et al. 2023), GPT-3 (Nath et al. 2022), CLUSTER2SENT (Krishna et al. 2020),. BioBERT (Lee et al. 2020), and XrayGPT (Thawkar et al. 2023). Agrawal et al. (2022) utilized GPT-based models to extract critical variables from diverse clinical notes, demonstrating that GPT-3 outperforms other models in clinical natural language processing tasks. Chintagunta et al.

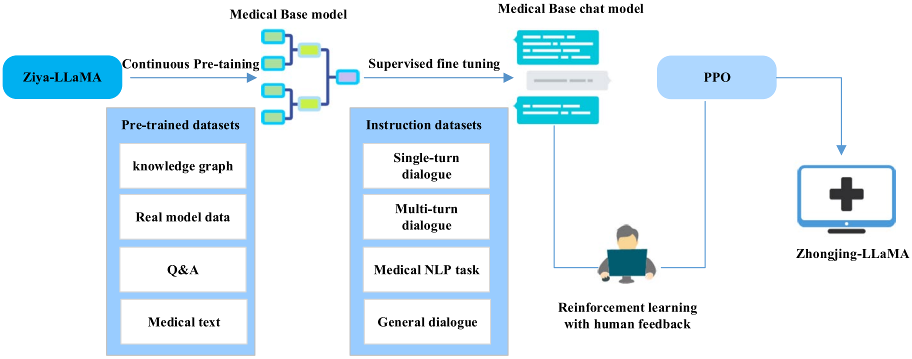

Fig. 2 The architecture of Zhongjing LLM (Yang et al. 2023a)

(Chintagunta et al. 2021) introduced GPT-3-ENS, a medically adapted GPT-3 model, for data annotation. This model produces synthetic training data that emphasizes relevant medical information, increasing human-labeled examples by over 30-fold. Integrating these high-quality synthetic data with human-labeled data enhances the accuracy and consistency of summaries in MDS tasks. Krishna et al. (Krishna et al. 2020) proposed the deep summarization model CLUSTER2SENT, which employs a pre-trained T5 model as an abstractive component to generate clinical summaries from doctor-patient dialogues. To offer users precise and beneficial health information, Yadav et al. (Yadav et al. 2021) developed a relevance-based reranking model based on the T5 framework, leveraging transfer learning to provide more precise and valuable information in multi-answer summarization tasks. Additionally, they applied a pre-trained Transformer model, enhanced with transfer learning, to address summarization challenges.

As the healthcare field evolves, the health-related streaming data available online must grapple with the challenges posed by vast volumes, rapid generation, diversity, and variability. Balumuri et al. (Balumuri et al. 2021) introduced a model that leverages transfer learning on pre-trained BERT, T5, and PEGASUS architectures, markedly enhancing the summarization capabilities of health question-answering systems. (Alqahtani et al. 2023) employed fine-tuned T5, BERT, and BioGPT models to summarize medical dialogues between doctors and patients. These models are adept at capturing all medical conditions described within dialogues and accurately identifying affirmations and negations in a medical context. The task of natural language understanding is significantly challenged when individuals seeking health information online verbose descriptions and peripheral details to articulate medical conditions. Clinical notes summarization assists healthcare practitioners in identifying potential health risks within patients' electronic health records (Wornow et al. 2023), thereby reducing errors and facilitating informed decision-making. Chuang et al. (Chuang et al. 2023) proposed the model-agnostic Soft Prompt-Based Calibration method, SPeC, to address the issue of increased output variance resulting from the integration of instruction prompts with large language models. This method ensures heterogeneity and reliability in the generation of medical summary information, as demonstrated in Fig. 3.

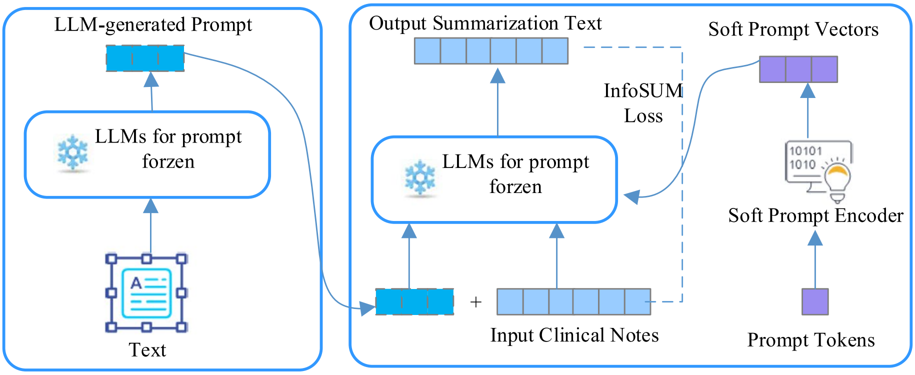

Fig. 3 The pipeline of SPeC (Chuang et al. 2023)

2.3 Electronic health records (EHRs), clinical letters and medical note generation

The LLMs are capable of generating clinical letters, medical notes, and electronic health records (EHRs) for specific issues through text-based dialogue. This capability is influenced by models such as ChatGPT (Cascella et al. 2023), GatorTron (Yang et al. 2022), ClinicalBERT (Alsentzer et al. 2019), BioMegatron (Shin et al. 2020), and GPT-4 (Abdelhady and Davis 2023), which impact multiple aspects of clinical documentation. Cascella et al. (Cascella et al. 2023) employed ChatGPT to create medical notes for intensive care unit (ICU) patients. After reviewing laboratory samples, blood gas parameters, and respiratory and hemodynamic data, ChatGPT accurately categorized most parameters into the appropriate domains. The model also exhibited a remarkable ability to self-correct by inquiring if its placement was appropriate, without requiring additional hints. Leveraging ChatGPT's robust language comprehension and text generation capabilities, (Ali et al. 2023) produced high-quality clinical letters across various clinical communication scenarios. The efficacy of the LLMs was demonstrated through a series of intricate commands, enhancing the precision and efficiency of intelligent text generation and ultimately providing more satisfactory services to patients. The research indicated that ChatGPT produces surgical records more rapidly than healthcare professionals, and the quality of these records, as well as their adherence to guidelines, is highly regarded by both patients and physicians, showcasing the potential of LLMs in the medical field.

In comparison to ChatGPT, GPT-4 exhibits superior problem-solving capabilities and an expansive knowledge base. Within the medical and healthcare domains, GPT-4 can supply the most current literature in specific fields, draft discharge summaries for patients post-surgery, analyze medical image characteristics, and identify objects in photographs, revealing its significant potential in clinical trials (Waisberg et al. 2023). Athavale et al. (Athavale et al. 2023) conducted two studies on complex medical issues, encompassing administrative management and chronic venous disease. Their evaluation of the assistance provided by EHR record inbox management functions revealed that GPT-4 outperformed ChatGPT3.5 across all problem domains, suggesting that this technology is poised to be utilized for EHR inbox management. Abdelhady and Davis (Abdelhady and Davis 2023) investigated the use of GPT-4 for generating surgical records of plastic surgeries performed by four surgeons, detailing the surgical types, record categories, description generation time, patient satisfaction, and comprehensive information about the surgeons' qualifications. Yang et al. (Yang et al. 2022) introduced GatorTron, an LLM with over 90 billion words, and assessed its performance on five clinical NLP tasks, examining the impact of varying scale parameters and training data (as depicted in Fig. 4).

2.4 Scientific research

In scientific research, LLMs can serve as powerful tools for data analysis (Tao et al. 2022), literature review, and hypothesis generation. They can efficiently sift through vast amounts of medical literature, extracting key information and identifying trends that might escape human researchers (Peng et al. 2023):

(1) LLMs present an exciting opportunity for researchers to streamline their research and craft influential articles by facilitating literature reviews (Chen and Li 2023), retrieving and discovering the latest scientific progress, automatically searching for academic

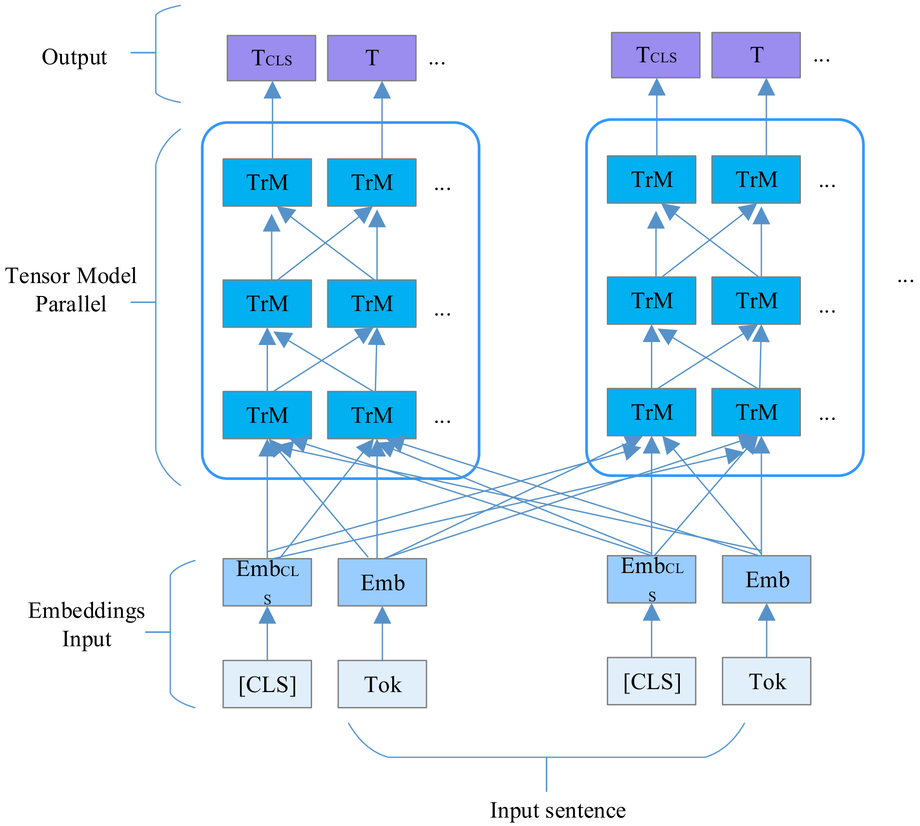

Fig. 4 The GatorTron model (Yang et al. 2022)

papers based on the needs of a given field and retrieving key information tailored to the requirements of different journals.

(2) LLMs have become indispensable tools in scientific writing, draft generation, article summarization, language and grammar checks, and translation of multilingual content (Liebrenz et al. 2023), engaging in discussions as virtual collaborators, and offering new perspectives and research directions, thereby enhancing the efficiency and diversity of scientific and academic output (Fatani 2023).

(3) LLMs with their advanced capabilities in natural language processing and understanding, can be effectively utilized for comprehensive data analysis and interpretation. These models can assist experimental design by providing valuable predictions, suggestions, summaries and interpretations of experimental results, thus enhancing the efficiency of the research process(Huang and Tan 2023).

(4) Researchers are able to gain valuable feedback and suggestions for improvement by submitting drafts or manuscripts, a practice that is especially advantageous for academic researchers who operate independently and do not have regular access to the peer review process. This allows such researchers to benefit from the expertise and insights of others, helping them to refine their work and enhance the quality of their research findings (Castellanos-Gomez 2023).

(5) LLMs can be seamlessly integrated with video, audio, and image recognition technologies to forge groundbreaking models, algorithms, and strategies (Zhang et al. 2023a). The synergy between LLMs and these recognition technologies empowers systems to understand and process multiple forms of data, allowing the development of sophisticated multimodal sentiment analysis techniques (Zhang et al. 2023b). This interdisciplinary approach has the potential to revolutionize various fields, including media analysis, marketing, and human–computer interaction (Zhang et al. 2023c).

2.5 Medical education and language translation

The LLMs have been utilized in diverse contexts within medical education and language translation. These include applications in licensing examinations such as the USMLE and the JMLE, the generation of multiple-choice questions, the evaluation of medical tests, educational initiatives in rehabilitation, and pharmacogenomics, as well as the translation of complex medical imaging reports into layman's terms to enhance healthcare education (Omran et al. 2023).

In the realm of medical assessments, the incorporation of multiple-choice questions necessitates substantial input from clinical professionals and educators. Gilson et al. (Gilson et al. 2023) investigated the performance of ChatGPT on multiple-choice questions from AMBOSS and NBME, which are part of USMLE. They analyzed the reasonableness of ChatGPT's answer generation logic and assessed the presence of internal and external information in the questions. The study found that ChatGPT significantly outperformed GPT-3 and InstructGPT on medical question-answering tasks, with its answer level comparable to that of third-year medical students. ChatGPT thus emerges as a potentially effective tool for interactive medical education that facilitates learning. Klang et al. (Klang et al. 2023) leveraged GPT-4 technology to compose 210 multiple-choice questions based on existing examination blueprints, categorizing them by algorithmic error and inaccuracy traits. GPT-4 thus serves as a potent supportive instrument for specialists in the construction of multiple-choice questions for medical assessments. Ueda et al. (Ueda et al. 2023) assessed ChatGPT's capability to analyze clinical scenarios and make decisions using the "Image Challenge" quiz from the New England Journal of Medicine (NEJM). This evaluation measured the accuracy of ChatGPT's responses in two settings: without options and within multiple-choice contexts. Without options, ChatGPT demonstrated an accuracy rate of 87%, while in multiple-choice scenarios, its accuracy reached 97%. This exceptional performance in the diagnostic category suggests that ChatGPT has significant potential for clinical application. Li et al. (Li et al. 2023b) conducted an evaluation of GPT-4's responses to diagnostic and treatment questions related to orthopedic diseases, adhering to the osteoarthritis management guidelines and orthopedic examination case questions. GPT-4 exhibited higher scores in terms of accuracy and completeness. It is poised to serve as an auxiliary tool in orthopedic clinical practice and patient education, offering high accuracy and comprehensive explanations of osteoarthritis treatment guidelines and clinical case analyses.

To test the performance of LLMs in JMLE, Takagi et al. (Takagi et al. 2023) conducted a comparative analysis and assessed the reliability of these LLMs in Japanese-based clinical reasoning and medical knowledge, examining 254 general sentence questions and clinical sentence questions. The results revealed that GPT-4 outperformed ChatGPT in general clinical questions, complex questions, and specific disease-related queries. Furthermore, GPT-4 achieved a score that met the passing standard of the JMLE, demonstrating its robust reliability in clinical reasoning and medical knowledge within the Japanese context. Kaneda et al. (Kaneda et al. 2023) investigated the responses of ChatGPT and GPT-4 in the Japanese National Nursing Examination (JNNE) of 2023. Their analysis included calculating the correct answer rate, score rate, comparing different LLMs, and assessing the accuracy rate of dialogue questions. GPT-4 exhibited sufficient performance to pass the JNNE, surpassing ChatGPT, which suggests that GPT-4 is suitable for specialized medical training in the Japanese clinical setting.

The remarkable performance of ChatGPT on USMLE has been a significant milestone in medical education (Sallam 2023b). LLMs have the potential to assist human learners in the field of medical education. Madrid-García et al. (Madrid-García et al. 2023) evaluated the performance of ChatGPT and GPT-4 in answering rheumatology questions on a specialized medical training access exam in Spain, examining factors such as the exam year, the diseases addressed, and the disease types. Both ChatGPT and GPT-4 demonstrated a high level of accuracy, suggesting that these models could serve as effective tools for rheumatology education, aiding in test preparation and complementing traditional teaching methods. Nori et al. (Nori et al. 2023) conducted a comprehensive evaluation of the GPT-4 model's performance on the USMLE dataset and the MultiMedQA benchmark dataset, assessing its content memory and the impact of images on the model's performance. The results indicated that GPT-4 achieved a score exceeding the passing threshold on the USMLE by more than 20 points without any professional hints, outperforming GPT-3.5 and specialized medical knowledge models such as Med-PaLM and Flan-PaLM. Kung et al. (Kung et al. 2023) evaluated ChatGPT's performance on the USMLE, a standardized medical test in the United States. ChatGPT achieved an accuracy level of approximately 60% without any specialized training. As the first LLM to reach this benchmark, ChatGPT exhibits comprehensible reasoning and practical clinical insight, enhancing trust and explainability in its applications (as illustrated in Fig. 5).

LLMs such as GPT-4 and Med-PaLM have demonstrated the ability to answer questions in the USMLE clinical knowledge test with an accuracy of over 80%. However, it remains unclear whether these LLMs can generate USMLE-like test questions. To address this question, Fleming et al. (Fleming et al. 2023) evaluated GPT-4's capability to produce

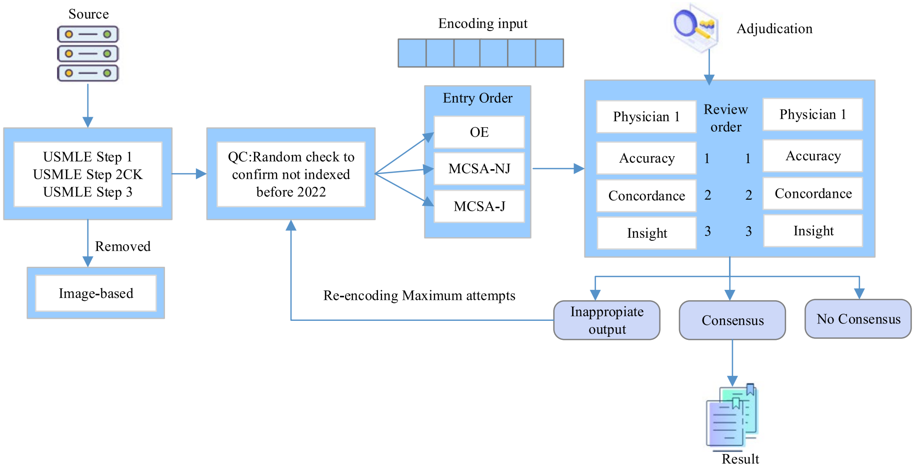

Fig. 5 The workflow of generating results (Kung et al. 2023)

authentic test questions and found that the USMLE test questions and answers generated by GPT-4 were not significantly different from those crafted by human physicians, and the generated questions and answers were deemed highly effective.

Rehabilitation education plays a vital role in the field of Physical Medicine and Rehabilitation (Peng et al. 2023). Models such as ChatGPT and GPT-4 can serve as virtual educational companions in rehabilitation. Engaging with ChatGPT or GPT-4 allows patients and their families to gain a deeper understanding of the essence, goals, and advantages of rehabilitation. This interaction facilitates a clearer comprehension of the challenges and expectations during the rehabilitation process, thereby enhancing the awareness and involvement in rehabilitation activities. Additionally, by acquiring effective self-management strategies from ChatGPT and GPT-4, patients and their families can actively engage in treatment, leading to improved rehabilitation outcomes and a better quality of life. Lyu et al. (Lyu et al. 2023) utilized ChatGPT to translate radiological reports of 76 brain magnetic resonance imaging cases into plain language. This initiative aimed to facilitate healthcare education for both patients and healthcare providers.

2.6 Medical imaging recognition and analysis

The LLMs have been trained to recognize and analyze medical images, including x-rays, magnetic resonance imaging (MRI), and ultrasound. These models can interpret features and structures within images, assisting physicians in accurately and rapidly identifying abnormalities, diagnosing diseases, and injuries. This capability significantly reduces the workload for radiologists (Waisberg et al. 2023). Moreover, LLMs can enhance image quality and resolution by reconstructing high-quality images from raw data obtained during medical imaging procedures. This improvement facilitates a deeper understanding of the internal structure and function of various organisms (e.g., (Tao et al. 2020)).

Medical imaging forms a cornerstone of the medical and healthcare field. The integration of LLMs can enhance radiologists' interpretive skills, facilitate communication between physicians and patients, and streamline workflow in clinical settings, particularly in hospitals. Yang et al. (2023b) developed the analytic framework BIGR-H based on ChatGPT to investigate the influence of LLMs on various stakeholders, including businesses, insurance companies, governments, research institutions, hospitals, and others within the medical imaging realm. For medical device manufacturers, LLMs can serve as a valuable tool for analyzing user feedback and technical documents, providing insights that inform device development. For health insurance companies and providers, LLMs can process and analyze large datasets to identify potential fraud patterns and anomalies, offering tools for insurers to prevent fraud. Additionally, LLMs can address policyholders' queries, provide personalized recommendations to enhance customer experience, and ensure the delivery of accurate and valuable information. For regulatory bodies, LLMs can strengthen the regulatory review process and assist in the detailed scrutiny of medical product submissions. Public health authorities can utilize the analytical capabilities of these models to analyze health data, identify disease trends and patterns, and significantly enhance disease surveillance, informing disease control and prevention strategies. These insights can also inform the development of more effective health policies, optimize resource allocation, and contribute to public health. Scientific research institutions and academic researchers can leverage LLMs to explain and analyze biomedical datasets, promoting more accurate conclusions and discoveries. Radiology and physical examination centers are integral to healthcare services, and LLMs can significantly impact the medical imaging process. Rao et al. (Rao et al. 2023) utilized ChatGPT to evaluate the capability of radiological clinical decision support for critical clinical manifestations, such as breast cancer screening and breast cancer pain.

2.7 Clinical health reasoning and diagnostic reasoning

LLMs have shown remarkable proficiency in tasks involving clinical health reasoning, real-world medical question-answering, and diagnostic reasoning. Feng et al. (2022) proposed the CHARD framework, which utilizes BERT and T5 models for clinical health reasoning, treating text generation models as implicit clinical knowledge bases to generate textual explanations of health-related problems across three dimensions. Liévin et al. (2022) evaluated the reasoning abilities of Codex and InstructGPT models using challenging real-world questions from USMLE, MedMCQA, and the PubMedQA medical reading dataset. Their findings suggested that scaling inference-time computing can enhance the reasoning performance of LLMs. Sharma et al. (2023) developed a Diagnostic Reasoning Benchmark for assessing clinical reasoning, using a clinically trained T5 model to analyze single-task and multi-task training on the summarization task. Singhal et al. (2023a) introduced the MultiMedQA benchmark for evaluating the answers generated by PaLM and Flan-PaLM models, which were refined through adjustments in model scale and instruction prompts. Liu et al. (2023b) also examined the performance of GPT-4 on various logical reasoning tasks, including out-of-distribution dataset testing for the robustness of GPT-related models. To improve the medical reasoning and in-depth thinking abilities of LLMs in medical conversational MQA, Weng et al. proposed a holistic thinking method that guides LLMs to perform both decentralized and centralized thinking, resulting in the generation of more professional and accurate answers (Fatani 2023).

2.8 Medical product safety monitoring and disease diagnosis

Due to the constrained scope and diversity of clinical trials for novel pharmaceuticals, comprehensive pre-market safety and efficacy assessments are often unattainable. LLMs can be utilized to monitor the safety of medical products by identifying Adverse Events (AEs) on social media platforms. Raval et al. (2021) developed the Adverse Event Detection and Extraction framework (AEDE), which is based on the T5 model. The AEDE leverages the T5 architecture's versatility in processing text from diverse domains and formats, thereby overcoming challenges such as the identification of infrequent signals, the management of imbalanced data in social media posts, substantial variations in text types across different media, the interpretation of misleading expressions and metaphors, and the annotation of data with extensive variability. Levine et al. (2023) assessed the diagnostic and triage capabilities of GPT-3 for common and serious diseases. GPT-3 yielded superior diagnostic outcomes compared to laypersons without domain-specific expertise, although its performance fell short of that of professional physicians. However, GPT-3 did not demonstrate significantly improved triage abilities over non-professional medical staff. Li et al. (2022a) utilized unbiased prompts to investigate the personality traits of GPT-3, InstructGPT, and FLAN-T5 through personality assessments (Short Dark Triad and Big Five Inventory) and well-being scales (Flourishing Scale and Satisfaction With Life Scale), with the intent of addressing sociopsychological safety concerns. The ChatGPT or GPT-4 model can aid intensive care physicians in reviewing potential diagnoses, treatment modalities, and possible complications in patient cases (Lu et al. 2023). By inputting pertinent information, intensive care physicians can render treatment decisions informed by a blend of clinical expertise. Da Mota Santana et al. (2023) discussed the potential utility of GPT-4 in digital oral radiology, based on dental radiographs, with the aim of reducing diagnostic error rates among professionals and enhancing clinical decision-making.

In the field of neurosurgery,LLMs have been utilized to forecast patients' hospital lengths of stay. Mantas (2022) conducted a comparative analysis of these predictions using the GPT-3 model and found no significant difference between the model's predictions and those made by physicians and patients. This result indicates the potential of employing LLMs for predicting the duration of hospitalization in neurosurgical cases. Virtual mental health assistants are increasingly common in healthcare settings, providing services such as counseling and supportive care to patients. However, these assistants are not suitable for use as diagnostic tools because they lack the ability to adhere to essential safety constraints and the professional clinical process knowledge required for accurate diagnosis. Roy et al. (2023) developed an algorithm named ProKnow-algo for the generation of natural language questions to collect diagnostic information iteratively through conversation. ProKnow-algo demonstrated a high level of safety and explainability in the context of diagnosing depression and anxiety (as depicted in Fig. 6).

2.9 Clinical decision support and administrative tasks assistance

The advanced capabilities of GPT-4 present a transformative opportunity to enhance doctor-patient communication, fostering a better understanding of patients' needs, anxieties, and expectations, thereby improving the overall medical experience (Nashwan et al. 2023). GPT-4 can facilitate the documentation of patients' medical histories by asking relevant questions, interpreting the responses, and presenting the information to physicians in a structured and concise format. This ensures that doctors gain a comprehensive understanding of their patients' conditions. Furthermore, GPT-4 can translate complex medical terminology and diagnostic results into plain language, making them more accessible to patients. It can also provide personalized advice on healthier lifestyles, diets, and medication use.

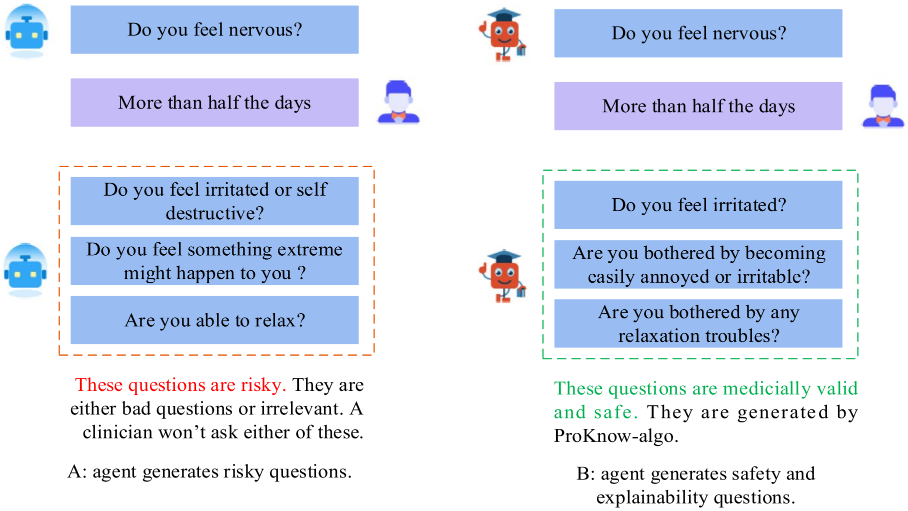
Do you feel nervous?
More than half the days
Do you feel irritated or self destructive?
Do you feel something extreme might happen to you ?
Are you able to relax?
These questions are risky. They are either bad questions or irrelevant. A clinician won't ask either of these.
A: agent generates risky questions.
Do you feel nervous?
More than half the days
Do you feel irritated?
Are you bothered by becoming easily annoyed or irritable?
Are you bothered by any relaxation troubles?
These questions are medicially valid and safe. They are generated by ProKnow-algo.
B: agent generates safety and explainability questions.

Fig. 6 The process of natural language question generation by ProKnow-algo (Roy et al. 2023)

To facilitate the efficient use of billing coding in healthcare, Soroush et al. (Soroush et al. 2023) assessed the capability of GPT-3.5 and GPT-4 in generating accurate International Classification of Diseases (ICD) billing codes. They randomly selected 100 codes from the billing code set published by the Centers for Medicare and Medicaid Services (CMS) to test the models' ability to derive correct ICD codes from textual descriptions and to analyze any error patterns qualitatively and quantitatively. In the realm of rehabilitation, proper assessment is pivotal to the patient's treatment process (Peng et al. 2023). Without a thorough evaluation of the patient's status, crafting an effective treatment plan is challenging. Given that ChatGPT and GPT-4 can process a wealth of in-depth rehabilitation evaluation data, they hold significant potential for practical application. These models can extract relevant information, generate statistical analysis reports through data analysis and pattern recognition, and integrate various evaluation data to enhance work efficiency and accuracy.

LLMs can facilitate communication among spinal surgeons, patients, and their relatives, streamline the acquisition and analysis of patient health data, and assist in the development of effective surgical plans. Furthermore, LLMs are capable of acquiring real-time surgical navigation information and physiological parameters, offering postoperative rehabilitation guidance to patients, and providing intraoperative support to spinal surgeons. Ilicki et al. (Ilicki 2023) developed a user-friendly LLM tailored for non-technical professionals in healthcare, which aids in identifying the primary source of patient data, determining the intended recipient, categorizing the data, and assessing fundamental limitations, to evaluate its applicability in healthcare settings. He et al. (He et al. 2023) conducted a systematic investigation into the use of GPT-4 in lumbar disc herniation surgery and found that GPT-4 can significantly support spinal surgeons in diagnosing conditions, managing the perioperative period, conducting scientific research, and enhancing communication with patients, as well as in planning and executing surgical procedures.

Despite growing interest among scholars and medical professionals in leveraging LLMs in healthcare, the examination and appraisal of their practical application and safety in clinical contexts remain limited. To assess whether LLMs, including GPT-3.5 and GPT-4, can reliably assist physicians in responding to queries from Information Consulting Services (ICS) in a safe and consistent manner, Dash et al. submitted 66 questions from an ICS to GPT-3.5 and GPT-4 via simple prompts. The responses were evaluated by 12 physicians regarding their alignment with potential patient injury risk, and they were found to be consistent with the ICS's reports. Among the 35 questions, GPT-3.5 and GPT-4 answered 8 and 13 correctly, respectively (Rosol et al. 2023). The findings indicate that LLMs can furnish safe and dependable responses but may not fully address the specific information requirements of a given query. To comprehensively evaluate LLMs' performance in healthcare settings, calibrating and customizing these models might be warranted.

Neurosurgery is a highly specialized and complex medical field that is dedicated to the surgical management of conditions affecting the central and peripheral nervous systems (Li et al. 2023c). The diagnosis and treatment of neurosurgery are intricate and demand high accuracy. Consequently, experts and scholars have sought to apply the latest and most powerful large language models (LLMs) to preoperative evaluation and preparation, customizing surgical plans and postoperative care and rehabilitation strategies, and providing communication and educational support to patients. Despite the exemplary performance of ChatGPT and GPT-4 models in various medical tasks, there is currently a scarcity of data employing large-scale electronic health records (EHR) to assess the performance of LLMs and their utility in providing clinical diagnostic assistance to patients. Consequently, Zhang et al. (Zhang et al. 2023d) utilized two advanced models, ChatGPT and GPT-4, to conduct this research. The findings revealed that GPT-4 achieved an accuracy rate of 96% in disease classification tasks with a thinking chain and few-shot prompts, and it could be corrected three times for four diagnostic tests.

A significant application of LLMs lies in recommender systems (Wang and Chen 2021),, which offer healthcare decision-making support to both patients and professionals. These systems can suggest personalized lifestyle improvements, such as tailored recipes, exercise regimens, drug therapies, and disease diagnostics (Wang and Zhao 2022). LLMs also have the potential to aid physicians in disease prediction and treatment, while online pharmaceutical retailers can integrate decision-making capabilities into social networks to streamline product selection for customers (Tran et al. 2021).

2.10 Case studies of LLMMs

LLMs hold immense promise in the application within the healthcare domain. However, their performance in addressing clinical issues and specific tasks during actual implementation is a matter of concern, prompting some medical scholars to conduct comprehensive evaluations and studies on preoperative guidance (Ke et al. 2024), clinical language understanding (Wang et al. 2023d) among other aspects. Ke et al.(2024) conducted a case study on several critical aspects of preoperative guidance within 14 de-identified clinical scenarios, including fasting guidelines, preoperative carbohydrate loading, medication instructions, medical team guidance, necessary preoperative optimization, and delayed surgery. The case study compared the LLM's responses with those of four anesthesiologists with less than five years of medical experience, resulting in a total of 1260 responses generated jointly by physicians, LLMs, and the LLM-augmented RAG (Retrieval-Augmented Generation) technology. The study involved multiple popular LLMs, such as ChatGPT, GPT-4.0, Llama2, and GPT4-RAG. The research found that the model augmented by GPT-4 with RAG technology was the most accurate, with the GPT4-RAG model achieving a performance of 91.4%, which is 5.1% higher than the human-generated answers at 86.3%. The GPT4-RAG model retrieved information in an average of just 1 s and generated results in an average of 15–20 s, while human physicians took an average of 10 min to produce preoperative instructions. This demonstrates the feasibility of the GPT4-RAG model in the specialized field of healthcare. Moreover, Wang et al. (2023d) have investigated the effectiveness of large models such as ChatGPT, GPT-4, and Bard in various clinical language understanding tasks within the realm of clinical language understanding. These tasks encompass named entity recognition, relation extraction, natural language inference, semantic textual similarity, and QA, among others, by employing different learning strategies and prompting techniques. Experiments were conducted on various clinical benchmark datasets, delving into different prompting strategies such as standard prompts, chain-of-thought, self-questioning, zero-shot, and 5-shot. The findings revealed that GPT-4 generally outperforms Bard and ChatGPT in classification tasks like named entity recognition, natural language inference, and semantic textual similarity. Across all settings, the performance of self-questioning prompts consistently surpasses that of standard prompts, suggesting self-asking to be a promising approach. Compared to zero-shot learning, 5-shot learning typically leads to improved performance across all tasks, indicating that even the incorporation of a small amount of task-specific training data can significantly enhance the efficacy of pre-trained LLMs.

2.11 Summarization of state-of-the-art LLMMs application scenarios

LLMMs have achieved significant advancements in various application scenarios, the implementation of LLMMs provides critical insights for various parties within healthcare providers and patients. This comprises improved patient education, crafting responses, or extracting information from patient notes in response to specific queries for healthcare providers, reviewing scientific literature for researchers, and elucidating clinical research protocols for clinical research coordinators (Lee et al. 2023). The latest achievements have witnessed a technological leap in Chinese question-answering systems, such as DoctorGLM (Xiong et al. 2023), Zhongjing (Yang et al. 2023a), and Huatuo (Li et al. 2023a) However, the generation and training of high-quality LLMMs pose significant challenges, necessitating substantial hardware support due to the massive resource consumption and prolonged training times. Moreover, the complexity of large-scale model architectures has heightened the difficulty in understanding and interpreting these models, particularly within the medical domain where incorrect predictions or biased recommendations could result in substantial harm to patients.

3 Available experimental datasets of LLMMs

This paper presents 56 experimental datasets that are most widely used by researchers in the medical and healthcare domains. These datasets encompass a range of tasks, including medical question-answering, medical knowledge representation, clinical evidence understanding and integration, diagnosis generation and summarization, and others. However, the extensive training of LLMMs is typically based on English-related datasets, resulting in a lack of medical knowledge, which can lead to poor performance in tasks such as disease diagnosis, drug recommendation, and clinical decision support. Existing medical datasets based on English corpora present challenges for conducting accurate experimental analyses of LLMMs on Chinese tasks. To address these issues, several scholars have proposed feasible solutions, such as Zhongjing (Yang et al. 2023a), DoctorGLM (Xiong et al. 2023), Huatuo (Li et al. 2023a), among others. Table 2 illustrates the datasets for LLMM research.

4 Evaluation metrics

Evaluating the performance of LLMMs is critical. Commonly used metrics include ROUGE, BERTScore, BLEU scores, accuracy, precision, recall, and F1-score for precision evaluation tasks. Some researchers also measure model performance using Medical Concept Coverage (Chintagunta et al. 2021) to test the importance and negations. Given the potential for unfair and unsafe outputs when applying LLMs in these fields, evaluating models and algorithms in this context requires considering their risks and feasibility. The average number of unsafe matches (Roy et al. 2023) offers a way to measure the effectiveness of the harm or severe consequences of the generated questions. Table 3 provides a summary of evaluation metrics used in different LLMM research papers.

Table 2 Datasets for LLMM research

Dataset	Size	Source	Characteristics	Availability	Publication
GPT-3-ENS	6900 labeled snippet-summary pairs	Chat-based telemedicine platform	Data are randomly sampled without replacement	No	(Chintagunta et al. 2021),(Arasu et al. 2023)
MEDIQA-2019	208 questions and 1701 associated answers	The U.S. National Library of Medicine and National Institute of Health	Including 3 tasks: natural language inference, recognition of question implication and question answering in the medical domain	Yes	(Yadav et al. 2021),(Zhou and Zhang 2021),(Abacha et al. 2019),(Zhu et al. 2019)
MS MARCO passage	8.8 M passages and about 1 M questions	Data were obtained via a search conducted using Bing search engine	Multiple answers per question, all human-generated	Yes	(Yadav et al. 2021),(Arabzadeh et al. 2021),(Gupta and MacAvaney 2022),(Craswell et al. 2021),(Kamphuis et al. 2023)
MS MARCO MED	More than 1 M anonymous questions	Obtained through Bing search	Medical subset of MS MARCO passage, only contains medical-related queries	Yes	(Yadav et al. 2021),(Pradeep et al. 2020)
LiveQA TREC-2017	634 and 104 QA samples from consumers in the development and test sets, respectively	Questions are from the U.S. National Library of Medicine, concerning consumer health	Each question is annotated with focus, type, and keywords	Yes	(Balumuri et al. 2021; Zhou and Zhang 2021),(He et al. 2020),(Jing et al. 2022)
SMM4H	19,699 Twitter posts with annotations	Primarily sourced from social media platforms such as Twitter	Composed of twitter posts, these are short, informal texts with non-standard orthography, annotated for detection and extraction	Yes	(Raval et al. 2021),(Davydova and Tutubalina 2022),(Sakhovskiy and Tutubalina 2022),(Portelli et al. 2021),(Gao et al. 2022)
CADEC	1250 medical forum posts	Sourced from patient-reported Adverse Events	The texts tend to be lengthy and casual, frequently straying from standard English syntax and punctuation conventions	Yes	(Raval et al. 2021),(Scepanovic et al. 2020),(Haq et al. 2022),(Fei et al. 2021),(Zhang et al. 2021),(Li et al. 2022b),(Roy et al. 2021)

Table 2 (continued)

Dataset	Size	Source	Characteristics	Availability	Publication
ADE corpus v2	37,716 case reports	Extracted from MEDLINE	Containing annotations for drugs, dosage, Adverse Events detection, and extraction	Yes	(Raval et al. 2021),(Dietrich and Kazzer 2023),(Francis et al. 2023)
WEB-RADR	57,481 tweets	Sourced from twitter	A manually curated benchmark, based on tweets, for testing the performance of multi-task models	No	(Raval et al. 2021),(Dietrich et al. 2020),(Gattepaille et al. 2020)
AMI dataset	138 business meetings with various participant roles	The dataset contains real meetings and a large number of scenario-driven meetings	Each transcript contains an associated abstractive summary	Yes	(Krishna et al. 2020),(Li et al. 2021),(Feng et al. 2020)
Medical SOAP dataset	6862 patient visits	Recorded English-language clinical dialogues	Each patient visits contains a human-generated conversation transcript	Yes	(Krishna et al. 2020),(Schloss and Konam 2020),(Passos et al. 2022),(Quesado et al. 2022)
CHQ-Summ	1507 domain-expert consumer health questions	Sourced from community Q&A forums	The questions are with annotated information and corresponding summaries	Yes	(Yadav et al. 2022a),(Zhang and Liu 2022),(Yadav et al. 2023)
MeQSum	1000 question-summary pairs in the training set and 50 NLM question-summary pairs	Real consumer health queries from health-related websites or forums	An original health question corresponds to a summary question	Yes	(Balumuri et al. 2021),(Yadav et al. 2022b)
Recognizing Question Entailment	8,588 QA training pairs and 302 QA validation pairs	Derived from various text entailment challenges like RTE1, RTE2, RTE3, RTE5	The data samples are constructed based on news and Wikipedia texts	No	(Balumuri et al. 2021)
Medical Question Pairs dataset	3,048 pairs of medical questions	The questions are sourced from HealthTap crawler data	Doctors rewrite each question and offer a related yet distinct question	Yes	(Balumuri et al. 2021)

Table 2 (continued)

Dataset	Size	Source	Characteristics	Availability	Publication
CHARDat	52 health conditions and 937 sentences	From three clinical dimensions	Includes three clinical dimensions: treatment, risk factors, and prevention	Yes	(Feng et al. 2022)
Gastrointestinal Clinical Dialogue	7 audio files containing 4 ~ 5 min of conversation	Sourced from online health consultation communities	Including about 47 utterances between practitioners and patients in a clinical conversation	Yes	(Nanayakkara et al. 2022),(Liu et al. 2021)
PubMed Gastrointestinal	33 million citations of biomedical papers	Compiled from sources like MEDLINE, and life science publications	Dataset comprising title and abstract pairs	Yes	(Nanayakkara et al. 2022),(Singhal et al. 2023a)
PubMedQA	1000 expert-annotated, 61,200 unlabeled, and 211,300 artificially generated QA instances	Derived from PubMed abstracts	A curated set of yes/no/maybe research questions, for assessing reading comprehension	Yes	(Liévin et al. 2022),(Wu et al. 2023)
USMLE	376 publicly-available tests	Derived from the United States Medical Licensing Examination (USMLE)	Each question details a medical case with a query reflecting real-world clinical practice	Yes	(Liévin et al. 2022),(Kung et al. 2023),(Fleming et al. 2023),(Nori et al. 2023),(Toma et al. 2023),(Han et al. 2023),(Wu et al. 2023),(Singhal et al. 2023b)
ProKnow-data	Approximately 640 questions	The data originates from a survey questionnaire conducted by clinicians diagnosing Major Depressive Disorder and Anxiety Disorder	A large-scale dataset with process-guided, safety-constrained medical knowledge and explainable features	No	(Roy et al. 2023),(Roy and Rawte 2022)

Table 2 (continued)

Dataset	Size	Source	Characteristics	Availability	Publication
Doctor-patient conversations dataset	1201 pairs of conversations in the training set, 100 pairs of conversations and summaries in the validation set, and 200 conversations in test set	Derived from doctor-patient conversation transcripts	The dataset includes annotated transcripts of doctor-patient dialogues, each with section headers and summary notes	Yes	(Alqahtani et al. 2023)
Huatuo-26 M	26 Million Chinese medical QA pairs	Multiple sources including online medical consultation websites, medical encyclopedias, and medical knowledge bases	The dataset is a large-scale Chinese medical question–answer dataset, covering a wide range of medical knowledge fields	Yes	(Li et al. 2023a),(Zhang et al. 2023e)
DR.BENCH	Each task contains tens of thousands or even more than one hundred thousand samples	The dataset encompasses six tasks from ten datasets, including clinical text understanding, medical knowledge reasoning, and diagnostic generation	Including three categories of six tasks, including medical knowledge representation task, Clinical evidence understanding and integration task, diagnosis generation, and summarization task	Yes	(Sharma et al. 2023),(Gao et al. 2023)
MIMIC-CXR	A total of 377,110 chest X-ray images, encompassing 227,835 radiology studies	Chest X-rays and free-text radiology reports from Boston's Beth Israel Deaconess Medical Center	Each study may include multiple images, including anterior and lateral views, with radiology reports written in semi-structured free text by practicing radiologists	Yes	(Chuang et al. 2023),(Johnson et al. 2019),(Pooch et al. 2020)
HealthSearchQA	3375 searched consumer medical questions	General medical knowledge searched for by consumers in Google	HealthSearchQA is a dataset curated to mirror genuine consumer health inquiries	Yes	(Singhal et al. 2023a)
MedQA	11,450 QA samples in the development set and 1273 samples in the test set	Comprehensive medical expertise as tested in the US medical certification exam	Each question is structured with a selection of four to five options	Yes	(Singhal et al. 2023a),(Singhal et al. 2023b)

Table 2 (continued)

Dataset	Size	Source	Characteristics	Availability	Publication
MedMCQA	187 K and 6.1 K QA samples and explanations in the development set and test set	Fundamental medical expertise tested in Indian medical admission exams	Each question is accompanied by four answer choices and detailed explanations	Yes	(Singhal et al. 2023a),(Toma et al. 2023; Liévin et al. 2022),(Wu et al. 2023),(Singhal et al. 2023b)
MultiMedQA	It is a composite of datasets including HealthSearchQA, MedQA, MedMCQA, and PubMedQA, etc	Derived from multiple datasets such as LiveQA, MedicationQA, and MMLU clinical topics datasets	Multiple-choice datasets, including MedQA, PubMedQA, MedMCQA,etc	Yes	(Nori et al. 2023),(Singhal et al. 2023a),(Singhal et al. 2023b)
MMLU	123 and 1089 QA samples in development set and test set	A large-scale multitask language understanding benchmark, comprising data from various domains	Spanning anatomy, clinical expertise, undergraduate medicine, medical genetics, professional healthcare, and college-level biology, this dataset encompasses a broad range of medical knowledge	Yes	(Singhal et al. 2023a),(Ray 2023)
Medication QA	674 long answers frequently sought by consumers	Frequent medication inquiries from consumers	A specialized dataset focusing on medication-related queries, providing detailed and accurate information for better drug understanding and patient care	Yes	(Singhal et al. 2023a), (Selvaraj and Konam 2020)
Clinical Acronym Sense Inventory	75 acronyms, each of which consists of 500 text examples	Clinical note snippets from various specialties at four affiliated hospitals of the University of Minnesota	Anonymized dataset supporting large-scale natural language processing and biomedical research	Yes	(Agrawal et al. 2022)
MedSTS	1000 annotated sentence pairs in clinical notes	Sourced from clinical notes at Mayo Clinic	MedSTS served as the benchmark in two open clinical NLP challenges	Yes	(Yang et al. 2022),(Wang et al. 2020a)

Table 2 (continued)

Dataset	Size	Source	Characteristics	Availability	Publication
MedNLI	Contains 11,232 training examples, 1,395 development examples, and 1,422 testing examples	Annotated by medical professionals based on patient medical histories	Designed to serve as a benchmark dataset within the clinical domain	Yes	(Yang et al. 2022),(Herlihy and Rudinger 2021)
Stanford NLI	570 K labeled sentence pairs	Human-written English sentence pairs	It serves as a benchmark for text representation systems and is a resource for developing any type of NLP model	Yes	(Yang et al. 2022),(Choi et al. 2021)
MultiNLI	433 K examples for sentence understanding	MultiNLI is a dataset that has been created through crowdsourced annotation	It covers a wide range of written and spoken English text genres and supports unique cross-genre generalization assessments	Yes	(Yang et al. 2022),(Zhao and Vydiswaran 2021)
ChatDoctor	5.4 K patient-physician conversations	10 M real-world patient-doctor dialogues	ChatDoctor integrates autonomous knowledge retrieval capabilities	Yes	(Xiong et al. 2023),(Yunxiang et al. 2023)
RheumaMIR	145 rheumatology- related questions from the exams in 2020–2023	Extracted from Spanish MIR exams held from 2009–2010 to 2022–2023	Evaluates and compares the performance of various AI models in handling rheumatology questions	Yes	(Madrid-García et al. 2023)
Atrium Health wake forest baptist clinical dataset	62 chest CT reports and 76 brain MRI screening reports	Collected from atrium health wake forest baptist clinical database	Anonymized by removing sensitive patient data	No	(Lyu et al. 2023),(Bowers et al. 2022)
ICD billing codes dataset	300 ICD billing codes from CMS	Randomly selected 100 unique codes from 3 lists provided by the Centers for Medicare & Medicaid Services	The ICD offers a standardized representation of medical conditions and procedures	No	(Soroush et al. 2023)

Table 2 (continued)

Dataset	Size	Source	Characteristics	Availability	Publication
MIMIC-III	377,110 images and 227,827 reports, 46,520 distinct patients admitted to ICU from 2001 to 2012	MIMIC-III originates from Beth Israel deaconess medical center's electronic health records	A set of chest radiographs from free-text radiology reports	Yes	(Zhang et al. 2023d),(Wang et al. 2020b),(Thawkar et al. 2023; Budrionis et al. 2021)
AMBOSS	More than 2700 Step 1 questions and 3150 Step 2 questions	Derived from a medical knowledge platform	Curated content focusing on essential medical knowledge	No	(Gilson et al. 2023),(Gencer and Aydin 2023)
NBME	Composed of 120 questions with 2 steps	Derived from the National Board of Medical Examiners	De-identified data supporting research in medical education and measurement	No	(Gilson et al. 2023),(Gilson et al. 2022)
MedDialog	MedDialog-EN contains nearly 0.3 million patient-doctor dialogues and 0.5 million utterances, while MedDialog-CN encompasses about 1.1 million dialogues and 4 million utterances	Interactions between medical professionals and patients	Covers multiple diseases across various specialties	Yes	(Fatani 2023),(Chen et al. 2020)
COVID-CT	Includes 349 COVID-19 CT images and 463 non-COVID-19 CT scans	The dataset is compiled from 760 preprints from medRxiv and bioRxiv	Belonging to 216 patients, both positive and negative for COVID-19	Yes	(Yang et al. 2020)
CMDD	The dataset comprises 2,067 conversations covering 4 pediatric diseases	Provided by Haodf and University of San Diego, a Chinese medical dialogue dataset	The dataset, balanced in counts, overlooks the disease data-imbalance issue	No	(Lin et al. 2021)
Medical Flash Cards	33,955 QA pairs	The data originates from Anki Flashcards	QA reformulations based on the faces of medical study cards	Yes	(Han et al. 2023)

Table 2 (continued)

Dataset	Size	Source	Characteristics	Availability	Publication
Stackexchange Medical Sciences	52,475 QA pairs	The question–answer pairs were from five Stack Exchange forums	he data pertains to biomedical sciences and associated disciplines	Yes	(Han et al. 2023)
Wikidoc	67,704 QA pairs generated from paragraphs	The data is sourced from the Living Textbook	QA pairs derived from paragraphs with questions based on rephrased headings and answers taken from the text	Yes	(Han et al. 2023)
S2ORC	4.9 M English-language academic papers with about 75B tokens	From English-language academic papers	The dataset is deeply intertwined with medical expertise	Yes	(Wu et al. 2023),(Muse et al. 2023)
CMtMedQA	70,000 real doctor-patient dialogues	Derived from multi-turn medical dialogue instances between doctors and patients	Multi-turn medical dialogue dataset in Chinese	Yes	(Yang et al. 2023a)
OpenI	6,459 chest X-ray images and 3,955 related reports	Originating from the Indiana University hospital system	The dataset includes high-quality interactive report summaries	Yes	(Thawkar et al. 2023)

Table 3 Summary of evaluation metrics used in different LLMM research papers

Metrics	Computing formula	Description	Research papers
Medical concept coverage (MCC)	$Concept-precision=\frac{\sum_{n=1}^{N}\lvert\hat{C}(n)\cap C(n)\rvert}{\sum_{n=1}^{N}\lvert\hat{C}(n)\rvert}$ $Concept-recall=\frac{\sum_{n=1}^{N}\lvert\hat{C}(n)\cap C(n)\rvert}{\sum_{n=1}^{N}\lvert C(n)\rvert}$	MCC measures the extent to which medical terms in the model's summary align with the actual content. Where C and $\hat{C}$ are the medical concept sets in the reference summary and the generated summary output	(Chintagunta et al. 2021)
ROUGE	$ROUGE=\frac{\sum_{S\in\{ReferemceSummaries\}}\sum_{gram_n\in S}Count_{match}(gram_n)}{\sum_{S\in\{ReferemceSummaries\}}\sum_{gram_n\in S}Count(gram_n)}$	ROUGE assesses summary quality by comparing the target summary to others. Where n represents the length of the n-gram $gram_n$, $Count_{match}\,(gram_n)$ is the maximum number of n-grams in a summary that a re concurrently present in the reference summary	(Chintagunta et al. 2021),(Krishna et al. 2020), (Yadav et al. 2021), (Balumuri et al. 2021),(Yadav et al. 2022a),(Feng et al. 2022),(Zhou and Zhang 2021),(Alqahtani et al. 2023),(Sharma et al. 2023),(Chuang et al. 2023)
Accuracy	$Acc=\frac{TP+TN}{TP+TN+FP+FN}$	Acc is the ratio of the correctly predicted data to the total data. TP and TN are the positive segments selected and the negative segments unelected, respectively. FP and FN are the segments incorrectly selected or unelected	(Krishna et al. 2020),(Zhou and Zhang 2021),(Singhal et al. 2023a),(Liévin et al. 2022),(Agrawal et al. 2022),(Yang et al. 2022),(Duong and Solomon 2023),(Levine et al. 2023),(Soroush et al. 2023),(Rosol et al. 2023),(Takagi et al. 2023),(Fleming et al. 2023),(Nori et al. 2023),(Ueda et al. 2023),(Gilson et al. 2023),(Han et al. 2023),(Singhal et al. 2023b),(Miao et al. 2023),(Abdelhady and Davis 2023),(Chervenak et al. 2023)
Precision	$Precision=\frac{TP}{TP+FP}$	Precision is the ratio of the correctly predicted positive samples to the total number of positive samples	(Raval et al. 2021),(Agrawal et al. 2022),(Yang et al. 2022),(Zhang et al. 2023d)
Mean reciprocal rank (MRR)	$MRR=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank_i}$	MRR is a metric for measuring the relevance of search results. The $rank_i$ represents the rank of the first results	(Zhou and Zhang 2021),(Li et al. 2023a)

Table 3 (continued)

Metrics	Computing formula	Description	Research papers
Recall	$Recall = \frac{TP}{TP+FN}$	Recall is the ratio of the total relevant documents that are correctly retrieved to the total relevant documents	(Raval et al. 2021),(Li et al. 2023a),(Agrawal et al. 2022),(Yang et al. 2022),(Zhang et al. 2023d)
F1-score	$F1 - score = \frac{2*\Pr ecision*\mathrm{Recall}}{\Pr ecision+\mathrm{Recall}}$	F1-score is a statistical measure that combines precision and recall to evaluate the accuracy of a test	(Krishna et al. 2020),(Raval et al. 2021),(Agrawal et al. 2022),(Yang et al. 2022),(Zhang et al. 2023d)
BERTScore	$R_{BERT} = \frac{1}{\|x\|}\sum_{x_i \in x \hat{x}_j \in \hat{x}} \max \mathrm{x}_\mathrm{i}^\mathrm{T}\hat{x}_j$ $P_{BERT} = \frac{1}{\|\hat{x}\|}\sum_{\hat{x}_j \in \hat{x} x_i \in x} \max \mathrm{x}_\mathrm{i}^\mathrm{T}\hat{x}_j$ $F_{BERT} = \frac{2*P_{BERT}*R_{BERT}}{P_{BERT}+R_{BERT}}$	BERTScore is a method for measuring the similarity between two texts, taking into account contextual and semantic information. The variable x is a reference, and $x\hat{}$is a candidate	(Yadav et al. 2022a),(Yadav et al. 2021),(Balumuri et al. 2021),(Feng et al. 2022),(Alqahtani et al. 2023)
BLEU scores	$BLEU = BP * \exp(\sum_{n=1}^{N} w_n \log p_n)$ $BP = \begin{cases} 1 \text{ if c} > \text{r} \\ \mathrm{e}^{(1 - \mathrm{r/c})} \text{ if c} \le \text{r} \end{cases}$	BLEU scores are a metric used to evaluate the quality of machine translation or any automatic translation system. The variable c is the total length of the candidate translation corpus, r is the reference length, w_n and p_n denote positive weights, and the geometric average of n-gram precision	(Alqahtani et al. 2023),(Fatani 2023)
Word error rate (WER)	$WER = \frac{S+D+I}{S+D+G}$	WER measures the percentage of words that are incorrectly recognized or translated in a sequence of words.. The S, D, and I represent the number of substitution, deletion, and insertion operations required to convert reference text into the language model output, respectively. G denotes the number of words that are identical in both the reference text and the output text	(Nanayakkara et al. 2022),(Leng et al. 2023),(Willett et al. 2023)

Table 3 (continued)

Metrics	Computing formula	Description	Research papers
Average number of unsafe matches (AUM)	$AUM\,(x, L) = \lvert L \cup t(x)\rvert - \frac{\lvert L\cap t(x)\rvert}{\lvert L\cup t(x)\rvert}$	AUM is used to measure whether the generated questions are safe. *L* is a dictionary of insecure concepts, and *t(x)* is the tokens that generate text *x*	(Roy et al. 2023),(Sheth et al. 2022)
Average number of knowledge context matches (AKCM)	$AKCM(x, K) = \frac{\lvert K\cap t(x)\rvert}{\lvert K\cup t(x)\rvert}$	AKCM is used to measure whether the generated questions are interpretable. *K* is the knowledge base concept set mapped to *x*	(Roy et al. 2023)
Average square rank error (ASRE)	$ASRE(x_i, R(x)) = \frac{\sum_{x_i \in s}(R(x_i)-i)^2}{\lvert S\rvert}$	ASRE measures the model's propensity to generate questions that adhere to causal tags and rankings. x_i denotes one of the questions in the generated sequence, *i* is the position in the sequence, *R*(x_i) represents the classifier tag of x_i	(Roy et al. 2023)
Good, missing, incorrect, and inaccurate		"Good" indicates accurate translation, "Missing" refers to lost information, "Inaccurate" implies partially retained information, and "Incorrect" denotes misinterpretation of the original radiology report	(Lyu et al. 2023)
P-value		P-value represents the probability of observing the data or more extreme data in hypothesis testing	(Soroush et al. 2023),(Rosol et al. 2023),(Takagi et al. 2023),(Gilson et al. 2023),(Singhal et al. 2023b),(Abdelhady and Davis 2023),(Ali et al. 2023)

5 Comparative performance analysis of various advanced models

The medical and healthcare domains employ a wide array of advanced techniques. We have summarized the performance of state-of-the-art LLMMs across various tasks, including clinical dialogue error correction (Nanayakkara et al. 2022), multiple-choice question answering (Singhal et al. 2023b), the MediQA shared task (Alqahtani et al. 2023), natural language inference (Yang et al. 2022), clinical health-aware reasoning (Feng et al. 2022), safety and explainability (Roy et al. 2023), and clinical decision support (Zhang et al. 2023d). These tasks are assessed using diverse metrics such as WER, accuracy, BLEU, BERTScore, ROUGE, AUM, AKCM, and ASRE. Moreover, we have summarized the performance of Chinese medical QA systems, which are evaluated based on professionalism, fluency, and safety, such as BenTsao (Wang et al. 2023c), Ziya-LLaMA (Zhang et al. 2022), DoctorGLM (Xiong et al. 2023), Zhongjing (Yang et al. 2023a), and Huatuo (Li et al. 2023a).

To present a comprehensive array of details regarding LLMMs across various tasks more clearly, we have meticulously described them in Tables 4 through 7, categorizing by task type. In Table 4, we encapsulate the performance metrics for three clinical dialogue transcription tasks utilizing Automatic Speech Recognition (ASR) technology from four prominent commercial ASR platforms: AWS Transcribe (AWS), Microsoft Speech-to-Text (Microsoft), IBM Watson (IBM), and Google Speech-to-Text (Google). A comprehensive breakdown of the Word Error Rate (WER) for the Gastrointestinal Clinical Dialogue dataset is provided within Table 4.

In Table 5, we have summarized the performance of two QA scenarios (namely, Multiple-choice QA and the MediQA shared task), encompassing six metrics across various datasets, including accuracy, BLEU, and F1 score. In the Multiple-choice QA task, Med-PaLM 2 achieved the top performance on the MedQA (USMLE) and PubMedQA datasets, while GPT-4 excelled on MedMCQA, MMLU-Medical Genetics, and MMLU-College Biology. For the MediQA shared task, the BART-Large model yielded the highest BLEU score, and T5 SAMSum achieved the highest F1 Score. Additionally, Li et al. (Li et al. 2023a) released the largest Chinese medical QA dataset, Huatuo-26 M, and Yang et al. (Yang et al. 2023a) pre-trained on this dataset. They conducted comparisons on Medical QA ranking in terms of Safety, Professionalism, and Fluency, as detailed in Table 5.

Table 6 presents an in-depth analysis of the semantic textual similarity, natural language inference, and clinical health-aware reasoning of multiple large models on the CHARDat, ProKnow-data, MultiNLI, and Stanford NLI datasets, including metrics such as accuracy, Pearson correlation, BERTScore, ROUGE, AUM, AKCM, and ASRE.

Table 7 compares the performance of ChatGPT and GPT-4 with and without a detailed clinical guideline in providing clinical decision support for Obstructive Pulmonary Disease (COPD), Primary Biliary Cirrhosis (PBC), and Chronic Kidney Disease (CKD) on the MIMIC-III dataset. The results reveal that both ChatGPT and GPT-4, when equipped with an elaborate clinical guideline, consistently achieved higher F1 scores across the board, as detailed in Table 7.

6 Challenges and future directions

Given the critical nature of medical and healthcare activities, which are inherently linked to patient life and health (Singhal et al. 2023a), the deployment of large prediction models for research, medical advice, and decision-support systems necessitates a heightened focus

Table 4 Summary of LLMMs performance in clinical dialogue transcription tasks

Tasks	Models	Performance	Publications
Clinical dialogue error correction	T5-Small	AWS: 55.41; Microsoft: 54.87;IBM: 61.74;Google: 64.20	(Nanayakkara et al. 2022), (Wei et al. 2023),(Lewis et al. 2019)
	T5-Base	AWS: 214.08;Microsoft: 205.56; IBM: 209.84; Google 162.63	
	T5-Large	AWS: 163.96; Microsoft: 163.54; IBM: 153.64; Google:137.19	
	BART-Base	AWS: 38.29; Microsoft: 30.95; IBM: 42.63; Google:44.47	
	BART-Large	AWS: 66.95; Microsoft: 55.40; IBM: 61.50; Google:55.12	
Summarization	Fine-tuned T5-Small	AWS: 663.39; Microsoft: 66.89; IBM: 69.44; Google:73.80	
	Fine-tuned BART-base	AWS: 76.61; Microsoft: 77.03; IBM: 78.10; Google:75.56	
Paraphrasing	Fine-tuned T5-Small	AWS: 48.87; Microsoft 47.24;IBM: 54.52; Google:57.97	
	Fine-tuned BART-base	AWS: 43.31;Microsoft: 37.46; IBM: 47.51; Google:49.48	
Mask-filling	Fine-tuned T5-Small	AWS: 38.83; Microsoft 35.86; IBM: 45.16; Google:46.87	
	Fine-tuned BART-base	AWS: 32.38; Microsoft: 26.38; IBM: 47.51; Google:40.43	

Table 5 Summary of LLMMs performance in QA-related tasks

Tasks	Models	Performance	Dataset	Metrics	Publications
Multiple-choice QA	Flan-PaLM	67.6	MedQA (USMLE)	Accuracy (%)	(Singhal et al. 2023b), (George et al. 2023), (Wang et al. 2023b), (Abdelhady and Davis 2023)
	Med-PaLM 2	86.5			
	GPT-4	86.1			
	Flan-PaLM	79.0	PubMedQA		
	Med-PaLM 2	81.8			
	GPT-4	80.4			
	Flan-PaLM	57.6	MedMCQA		
	Med-PaLM 2	72.3			
	GPT-4	73.7			
	Flan-PaLM	75.0	MMLU-medical genetics		
	Med-PaLM 2	92.0			
	GPT-4	97.0			
	Flan-PaLM	88.9	MMLU-College biology		
	Med-PaLM 2	95.8			
	GPT-4	97.2			
MediQA shared task	BART-Large	0.561	Doctor-patient conversations dataset	BLEU	(Alqahtani et al. 2023), (Wei et al. 2023), (Yuan et al. 2022), (Lewis et al. 2019)
	BioBART	0.550			
	BioGPT	0.359			
	Flan-T5 Large	0.510			
	T5 SAMSum	0.52			
	BART-Large	0.580		F1 score (%)	
	BioBART	0.581			
	BioGPT	0.519			
	T5 Large	0.645			
	T5 SAMSum	0.672			

Table 5 (continued)

Tasks	Models	Performance	Dataset	Metrics	Publications
Chinese medical dialogue	Zhongjing-BenTsao	95	CMtMedQA	Professionalism and fluency improvement (%)	(Wang et al. 2023c), (Zhang et al. 2022), (Xiong et al. 2023), (Yang et al. 2023a), (Li et al. 2023a)
	Zhongjing-DoctorGLM	80			
	Zhongjing-Ziya-LLaMA	71			
	Zhongjing-HuatuoGPT	52			
	Zhongjing-ChatGPT	49			
	Zhongjing-BenTsao	99		Safety improvement (%)	
	Zhongjing-DoctorGLM	83			
	Zhongjing-Ziya-LLaMA	54			
	Zhongjing-HuatuoGPT	68			
	Zhongjing-ChatGPT	32			
	Zhongjing-BenTsao	94	Huatuo-26 M	Professionalism and fluency improvement (%)	
	Zhongjing-DoctorGLM	78			
	Zhongjing-Ziya-LLaMA	56			
	Zhongjing-HuatuoGPT	51			
	Zhongjing-ChatGPT	40			
	Zhongjing-BenTsao	97		Safety improvement (%)	
	Zhongjing-DoctorGLM	81			
	Zhongjing-Ziya-LLaMA	44			
	Zhongjing-HuatuoGPT	65			
	Zhongjing-ChatGPT	26			

Table 6 Summary of LLMMs performance in inference, reasoning and semantic textual similarity

Tasks	Models	Performance	Dataset	Metrics	Publications
Natural language inference	BioBERT	0.8050	MultiNLI and stanford NLI	Accuracy (%)	(Yang et al. 2022), (Lee et al. 2020), (Alsentzer et al. 2019), (Shin et al. 2020)
	ClinicalBERT	0.8270			
	BioMegatron	0.8390			
	GatorTron-medium	0.8720			
	GatorTron-large	0.9020			
Semantic textual similarity calculation	BioBERT	0.8744		Pearson correlation	
	ClinicalBERT	0.8787			
	BioMegatron	0.8806			
	GatorTron-medium	0.8903			
	GatorTron-large	0.8896			
Clinical health-aware reasoning	RETR	39.54	CHARDat	BERTScore	(Feng et al. 2022),, (Lewis et al. 2019), (Wei et al. 2023)
	BART-large	60.78			
	BART-base	60.04			
	T5-large	59.00			
	T5-base	59.80			
	RETR	ROUGE-1,2,L: 43.30, 28.18,39.03		ROUGE	
	BART-large	ROUGE-1,2,L: 51.54, 40.27, 49.88			
	BART-base	ROUGE-1,2,L: 51.37, 39.35, 49.55			
	T5-large	ROUGE-1,2,L: 50.66, 37.74, 48.05			
	T5-base	ROUGE-1,2,L: 50.00, 38.31, 48.07			

Table 6 (continued)

Tasks	Models	Performance	Dataset	Metrics	Publications
Safety and explainability	Fine-tuned T5	0.77	ProKnow-data	ROUGE	(Roy et al. 2023), (Wei et al. 2023)
	QG-LSTM	0.85			
	QG-transformer	0.87			
	Fine-tuned T5	0.63		BERTScore	
	QG-LSTM	0.82			
	QG-transformer	0.82			
	Fine-tuned T5	0.2		AUM	
	QG-LSTM	0.1			
	QG-transformer	0.133			
	Fine-tuned T5	1.3		AKCM	
	QG-LSTM	1.12			
	QG-transformer	1.27			
	Fine-tuned T5	0.0001		ASRE	
	QG-LSTM	0.0004			
	QG-transformer	0.0007			

Table 7 Summary of LLMMs performance in clinical decision support

Tasks	Models	Performance	Metrics	Publications
Clinical decision support on COPD	ChatGPT without an elaborate clinical guideline	69	Precision (%)	(Zhang et al. 2023d), (Wang et al. 2023b), (Cascella et al. 2023)
		97	Recall (%)	
		81	F1 Score (%)	
	ChatGPT with an elaborate clinical guideline	84	Precision (%)	
		94	Recall (%)	
		89	F1 Score (%)	
	GPT-4 without an elaborate clinical guidelin	83	Precision (%)	
		100	Recall (%)	
		91	F1 Score (%)	
	GPT-4 with an elaborate clinical guidelin	98.3	Precision (%)	
		93.7	Recall (%)	
		96	F1 Score (%)	
Clinical decision support on PBC	ChatGPT without an elaborate clinical guideline	26	Precision (%)	
		82	Recall (%)	
		39	F1 Score (%)	
	GPT-4 without an elaborate clinical guidelin	40	Precision (%)	
		95	Recall (%)	
		57	F1 Score (%)	
	GPT-4 with an elaborate clinical guidelin	86.36	Precision (%)	
		86.36	Recall (%)	
		86.36	F1 Score (%)	
Clinical decision support on CKD	ChatGPT without an elaborate clinical guideline	73	Precision (%)	
		81	Recall (%)	
		77	F1 Score (%)	
	GPT-4 with an elaborate clinical guidelin	98.18	Precision (%)	
		79.41	Recall(%)	
		87.80	F1 Score(%)	

on safety, reliability, effectiveness, and patient privacy. As LLMs become more advanced, they are increasingly susceptible to generating harmful or inappropriate content, such as hallucinations, spam, sexist, and racist hate speech. These models may also produce responses that sound plausible yet are incorrect or absurd. Consequently, addressing safety concerns becomes paramount in healthcare decision-making involving LLMs. Recognizing this challenge, several researchers have adopted effective training and evaluation methods and have compiled new datasets for LLMs, such as the use of unbiased prompts (Li et al. 2022a) and the CHARDat dataset (Feng et al. 2022) We categorize the ethical and safety issues associated with LLMs into five key areas: data security and privacy-preservation, the risk of incorrect or misleading information, fairness and bias, transparency, explainability, and trustworthiness, and issues related to plagiarism, copyright, and accountability. We propose potential solutions and outline future prospects based on these categories and the challenges they present, as shown in Table 8.

6.1 Data security and privacy-preserving

Medical reports may inadvertently reveal private and demographic details of patient records. Ensuring patient privacy and adhering to data security regulations can be more complex and challenging than achieving optimal medical outcomes (Chuang et al. 2023). The digitization of healthcare facilitates the sharing and repurposing of medical data, yet it also increases the risk of critical patient information being compromised (Liu et al. 2023c). The Health Insurance Portability and Accountability Act (HIPAA) mandates patient confidentiality and privacy, stipulating that medical records must be sanitized of sensitive information before dissemination. Consequently, there is an imperative for robust solutions to identify and safeguard medical data. While rule-based and machine learning-based de-identification methods have been extensively implemented in practice, they remain limited in their versatility and effectiveness across diverse scenarios.

LLMs like ChatGPT and GPT-4 demonstrate significant potential in addressing the privacy protection challenge for medical text data. For instance, GPT-4 can leverage named entity recognition to construct a de-identification framework that automatically identifies and eliminates patient-specific information. A data management plan (DMP) (Stanciu 2023) provides guidelines for executing data-related activities and methods for safeguarding data security and confidentiality during storage, presentation, sharing, and distribution. Consequently, the DMP may serve as an effective approach to address data security issues.

6.2 Incorrectness and risk of inaccurate information

LLMs exhibit considerable potential in executing a diverse range of tasks that typically require human capabilities, having been trained on extensive internet data (Harrer 2023). However, this training may inadvertently integrate misinformation and biased content, potentially leading to significant drawbacks such as the generation of incorrect or fabricated information (Reddy 2023). Given the safety–critical nature of medical and healthcare domains, erroneous advice regarding patients' symptoms and medications can result in serious injury or even death (Munn et al. 2023), as exemplified by GPT-3 incorrectly recommending suicide for a patient (Atallah et al. 2023a). Consequently, it is imperative to implement safeguards around the use of LLMs in healthcare, including their assistance in tasks such as generating discharge summaries, automatically producing explanatory medical records, and providing medical recommendations.

Table 8 The challenges and potential solutions in LLMMs

Challenges	Potential solutions	Details
Data security and privacy-preserving	Leveraging named entity recognition to construct a de-identification framework that automatically identifies and eliminates patient-specific information Protecting data security during data storage, presentation, sharing, and distribution through a Data Management Plan (DMP)	Section 6.1
Incorrectness, risk of inaccurate information	The authenticity of LLM-generated outputs for various medical tasks can be validated against different references methods that integrate few-shot In-Context Learning (ICL) with Chain-of-Thought (CoT) and reason prompts can automate the detection and correction of medical errors in clinical notes	Section 6.2
Fairness/bias	The Counterfactually Fair Prompting (CFP) strategy involves using an encoder prompt to strip sensitive attributes from encoder-decoder models and a decoder prompt to sustain performance. For decoder-only models, a single decoder prompt suffices. Concatenating CFP with the input prompt removes sensitive data from user token embeddings	Section 6.3
Transparency, explainability, and trustworthiness	SweCTRL-Mini uses Opening Control Codes (OCC) as single-token prompts to steer the genre of generated text, and Ending Control Codes (ECC) to signal when to end text generation within a genre. This transparent design aids in monitoring genre mixing by the model	Section 6.4
Plagiarism, copyrights, accountability	By establishing normative standards for the application of LLMs in healthcare by medical, healthcare institutions, and government agencies, and providing guidance for the design and deployment of these models	Section 6.5

The authenticity of LLM-generated outputs for various medical tasks can be validated against different references (Xie et al. 2023). Text summarization or simplification systems rely on the original medical documents, such as study protocols or clinical notes, to ensure that the AI-generated content aligns with the source information. Similarly, AI systems that generate radiology reports from Chest X-ray images use radiologists' reports as the reference. Moreover, methods that integrate few-shot In-Context Learning (ICL) with Chain-of-Thought (CoT) and reason prompts can automate the detection and correction of medical errors in clinical notes (Wu et al. 2024). One approach involves manually analyzing a subset of the training and validation data to infer CoT prompts based on error types in the clinical notes. Another method prompts the LLM with the training data to deduce reasons for the correctness or incorrectness of the information. Both methods then enhance the CoTs and reasons with ICL examples to tackle tasks such as error detection, span identification, and error correction.

6.3 Fairness and *bias*

Due to their training on a vast array of internet content, LLMs may inadvertently incorporate biases (Arora and Arora 2023) (e.g., gender bias, racial bias, geopolitical biases, religious bias, nationality bias, sexual orientation bias, and age bias, etc.) across the web, posing severe threats in sensitive fields (Korngiebel and Mooney 2021).. Recent research has revealed a strong correlation between job opportunities and male job seekers, a correlation between negative emotions and the black race, and a correlation between positive emotions and the Asian race. For instance, GPT-4 was found not to simulate the demographics of medical conditions in various situations, consistently producing clinical hallucinations, including the differential diagnosis of standardized clinical samples, which are more likely to include stereotypes of specific racial, ethnic, and gender identities. The assessments and medical plans created by GPT-4 demonstrate a significant association between demographic attributes and patient differences in recommendations for expensive procedures (Zack et al. 2023). Medical and healthcare are particularly complex scenarios for applying LLMs (Singhal et al. 2023a). The training data of LLMs typically comes from institutions in high-income, English-speaking countries, which may severely limit the representativeness of viewpoints from other regions of the world. This can lead to biases in the mechanistic models of health and disease towards understanding this process in high-income countries. For example, when clinicians in Africa use LLMs to generate treatment plans for diabetes, they may focus on treatment models that are only applicable to high-income countries, thereby limiting the implementation of different treatment methods more relevant to the patient populations in other regions of the world (Thirunavukarasu et al. 2023).

The discrepancy between the model's output and the diagnoses of seasoned medical professionals can engender structural bias and inequitable treatment. Consequently, it is imperative to identify potential hazards and deviations that may affect doctors, patients, and healthcare professionals when designing models and algorithms. A limitation of large language models (LLMs) is that when biased data is employed for training (Atallah et al. 2023b), discriminatory outcomes can be perpetuated, persisting even after the model is recalibrated. To address this issue, some researchers have conducted exploratory work. For instance, in response to the unfairness exhibited by large language models such as T5 and LLaMA, Hua et al. (Lin et al. 2023) argue that similar individuals or groups should receive similar outputs in the pursuit of fairness. They have therefore proposed a strategy called Counterfactually Fair Prompting (CFP). For encoder-decoder large language models, an encoder prompt is needed to remove sensitive attributes, and a decoder prompt is required to maintain model performance. For models composed solely of a decoder, only a decoder prompt is necessary. By simply concatenating the CFP with the original input prompt, sensitive information in user token embeddings can be eliminated, achieving fairness across a set of sensitive attributes without the need to retrain the entire base model.

6.4 Transparency, explainability, and trustworthiness

Despite their impressive potential in performing various simple tasks, LLMs suffer from a lack of transparency, hindering their efficiency in assisting humans with complex tasks. To address this, several strategies have been proposed, such as using Chaining LLM (Wu et al. 2022) techniques or inserting tokens into generation prompts (Kalpakchi and Boye 2023). Specifically, SweCTRL-Mini is a data-transparent LLM designed for controllable text generation in Swedish. The core concept of SweCTRL-Mini is to enable the steering of the genre of the generated text through the use of Opening Control Codes (OCC) as single-token prompts. In addition to employing OCC to represent various stylistic texts, Keskar and colleagues have also incorporated Ending Control Codes (ECC) to signal to the model when to conclude text generation within a given genre. This transparent approach facilitates checking whether the model begins to blend genres. These strategies enhance the transparency and interpretability of both the LLM's training process and the generated text, thereby reducing errors in medical practitioners and bolstering the credibility of the strategy (Reddy 2023). A significant challenge for healthcare practitioners is the absence of guidelines for assessing whether LLM outputs align with social norms and regulations. The application of LLMs is currently grappling with a crisis of trustworthiness (Liu et al. 2023d). A foundational approach to improving safety and trustworthiness is to employ reinforcement learning from human feedback, which can augment strategies based on human guidance and mitigate the production of harmful content (Huang et al. 2023).

6.5 Plagiarism, copyrights, and accountability

Given that LLMs retain and train on the information provided, the generated text introduces the potential for plagiarism, which can be illegal and threaten the integrity and copyright of the publication (Nashwan et al. 2023). A New York Times report (Zhang et al. 2023f) indicates that ChatGPT provided conspiracy theories and misleading responses based on researchers' queries. Following adjustments to the output, ChatGPT generated persuasive but unattributed content, complicating the task of identifying plagiarism or original creation. Consequently, the development and application of new tools for detecting AI-generated text are essential.

Accountability is crucial to ensure that LLMs in medical and healthcare settings are used in a normative, responsible, and ethical manner (Reddy 2023). Establishing clear policies, procedures, and regulations can ensure that the use of LLMs aligns with legal and ethical standards. Therefore, it is advisable for medical, healthcare institutions, and government agencies to develop normative standards for the application of LLMs in healthcare and provide guidance for the design and deployment of these models.

7 Summary &Conclusions

This survey systematically reviews the recent advancements in state-of-the-art LLMs within the medical and healthcare domain. The focus includes applications in medical question-answering, medical dialog summarization, electronic health records, clinical letters and medical note generation, scientific research, medical education, language translation, medical imaging recognition and analysis, clinical health reasoning, diagnostic reasoning, medical product safety monitoring, disease diagnosis, clinical decision support, and administrative tasks assistance. Additionally, we summarize the available experimental datasets for developing LLMs and provide evaluation methods to ensure that these models are accurate, safe, and effective for problem-solving in medical and healthcare scenarios. We also discuss the significant challenges in data security and privacy preservation, the risk of incorrect information, fairness and bias, transparency, explainability, trustworthiness, plagiarism, copyrights, and accountability. For each aspect, we summarize the causes of the challenges and limitations and offer possible solutions to address the related problems.

Author Contribution D. W. and S. Z. wrote the main manuscript text. All authors reviewed the manuscript.

Funding Natural Science Foundation of Zhejiang Province (LQ24F020014), National Natural Science Foundation of China (62406217, 62276180), Key R&D Program of Zhejiang (2024C01022), The Project of Zhejiang Province's Educational Science Planning (2024SCG176), Ministry of Education Industry-University Collaboration and Joint Education Project (231100461225050), and the Humanities and Social Science Project of the Chinese Ministry of Education under Grant 24YJCZH285 (Research on intelligent collaborative teaching mode based on multimodal large model personalized programming supported by this grant).

Data Availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors have no competing interests as defined by Springer, or other interests that might be perceived to influence the results and/or discussion reported in this paper.

References

Abacha A B et al. (2019) Overview of the MEDIQA 2019 shared task on textual inference, question entailment and question answering. In: proceedings of the 18th BioNLP Workshop and Shared Task, pp 370–379

Abdelhady AM, Davis CR (2023) Plastic surgery and artificial intelligence: how ChatGPT improved operation note accuracy, time, and education. Mayo Clin Proc Digit Health 1(3):299–308

Agrawal M et al. (2022) Large language models are few-shot clinical information extractors. In: proceedings of the 2022 conference on empirical methods in natural language processing, pp 1998–2022

Ali SR et al. (2023) Using ChatGPT to write patient clinic letters. Lancet Digit Health 5(4):e179–e181

Alqahtani A et al. (2023) Care4Lang at MEDIQA-Chat 2023: Fine-tuning language models for classifying and summarizing clinical dialogues. In: proceedings of the 5th clinical natural language processing workshop, pp 524–528

Alsentzer E et al. (2019) Publicly available clinical BERT embeddings. Preprint at https://arxiv.org/abs/03323

Arabzadeh N et al. (2021) Ms marco chameleons: challenging the ms marco leaderboard with extremely obstinate queries. In: proceedings of the 30th ACM international conference on information & knowledge management, pp 4426–4435

Arasu N et al. (2023) The survey on GPT-3 driven NLP approach for automatic medical documentation. AIP Conf Proc 10(1063/5):0152503

Archana R, Jeevaraj PE (2024) Deep learning models for digital image processing: a review. Artif Intell Rev 57(1):11

Arora A, Arora A (2023) The promise of large language models in health care. The Lancet 401(10377):641

Atallah S et al. (2023) How large language models including generative pre-trained transformer (GPT) 3 and 4 will impact medicine and surgery. Tech Coloproctol 27(2023):609–614

Atallah S et al. (2023) How large language models including generative pre-trained transformer (GPT) 3 and 4 will impact medicine and surgery. Tech Coloproctol 23:1–6

Athavale A et al. (2023) The potential of chatbots in chronic venous disease patient management. JVS Vasc Insights 2023:100019

Balumuri S et al. (2021) Sb_nitk at mediqa 2021: Leveraging transfer learning for question summarization in medical domain. In: proceedings of the 20th workshop on biomedical language processing, pp 273-279

Bowers HJ et al. (2022) Dynamic characterization of breast cancer response to neoadjuvant therapy using biophysical metrics of spatial proliferation. Sci Rep 12(1):11718

Budrionis A et al. (2021) Benchmarking pysyft federated learning framework on mimic-iii dataset. IEEE Access 9:116869–116878

Cascella M et al. (2023) Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. J Med Syst 47(1):33

Castellanos-Gomez AJN (2023) Good practices for scientific article writing with ChatGPT and other artificial intelligence language models. Nanomanufacturing 3(2):135–138

Chang Y et al. (2023) A survey on evaluation of large language models. ACM Trans on Intell Syst Technol. https://doi.org/10.1145/3641289

Chen M, Li G (2023) ChatGPT for mechanobiology and medicine: a perspective. Mech Biol Med 1(1):100005

Chen S et al. (2020) Meddialog: a large-scale medical dialogue dataset. Preprint at https://arxiv.org/abs/3:03329

Chervenak J et al. (2023) The promise and peril of using a large language model to obtain clinical information: ChatGPT performs strongly as a fertility counseling tool with limitations. Fertil Steril. https://doi.org/10.1016/j.fertnstert.2023.05.151

Chintagunta B et al. (2021) Medically aware GPT-3 as a data generator for medical dialogue summarization. In: machine learning for healthcare conference, pp 354–372

Choi H et al. (2021) Evaluation of bert and albert sentence embedding performance on downstream nlp tasks. In: 2020 25th International conference on pattern recognition (ICPR), pp 5482–5487

Chuang Y-N et al. (2023) Spec: A soft prompt-based calibration on mitigating performance variability in clinical notes summarization. Preprint at https://arxiv.org/abs/13035

Craswell N et al. (2021) Ms marco: Benchmarking ranking models in the large-data regime. In: proceedings of the 44th International ACM SIGIR conference on research and development in information retrieval, pp 1566–1576

da Mota SLA et al. (2023) Can GPT-4 be a viable alternative for discussing complex cases in digital oral radiology? a critical analysis. Excli J 22:749–751

Davydova V, Tutubalina E (2022) Smm4h 2022 task 2: Dataset for stance and premise detection in tweets about health mandates related to covid-19. In: Proceedings of The seventh workshop on social media mining for health applications, workshop & shared task, pp 216–220

Dietrich J, Kazzer P (2023) Provision and characterization of a corpus for pharmaceutical, biomedical named entity recognition for pharmacovigilance: evaluation of language registers and training data sufficiency. Drug Saf 46:1–15

Dietrich J et al. (2020) Adverse events in twitter-development of a benchmark reference dataset: results from IMI WEB-RADR. Drug Saf 43:467–478

Duong D, Solomon BD (2023) Analysis of large-language model versus human performance for genetics questions. Eur J Hum Genet 32:1–10

Fan H et al. (2024) Transformer-based multimodal feature enhancement networks for multimodal depression detection integrating video, audio and remote photoplethysmograph signals. Inform Fusion 104:102161

Fatani B (2023) ChatGPT for future medical and dental research. Cureus 15(4):1–5

Fei H et al. (2021) Rethinking boundaries: End-to-end recognition of discontinuous mentions with pointer networks. In: proceedings of the aaai conference on artificial intelligence, pp 12785–12793

Feng C et al. (2020) Deep learning-based real-time building occupancy detection using AMI data. IEEE Trans Smart Grid 11(5):4490–4501

Feng SY et al. (2022) CHARD: Clinical health-aware reasoning across dimensions for text generation models. Preprint at https://arxiv.org/abs/04191

Fleming S L et al. (2023) Assessing the Potential of USMLE-Like Exam Questions Generated by GPT-4. Preprint at https://medRxiv.org/abs/23288588

Francis S et al. (2023) Understanding the impact of label skewness and optimization on federated learning for text classification. Companion Proc of the ACM Web Conf 2023:1161–1166

Gao Y et al. (2023) DR. Bench: diagnostic reasoning benchmark for clinical natural language processing. J Biomed Inform 138:104286

Gao Y et al. (2022) Contextualized graph embeddings for adverse drug event detection. In: joint European conference on machine learning and knowledge discovery in databases, pp 605–620

Gattepaille LM et al. (2020) Prospective evaluation of adverse event recognition systems in twitter: results from the web-RADR Project. Drug Saf 43:797–808

Gencer A, Aydin S (2023) Can ChatGPT pass the thoracic surgery exam? Am J Med Sci. https://doi.org/10.1016/j.amjms.2023.08.001

George AS et al. (2023) AI-Driven breakthroughs in healthcare: google health's advances and the future of medical AI. Partn Univ Int Innov J 1(3):256–267

Gilson A et al. (2023) How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 9(1):1–9

Gilson A et al. (2022) How well does ChatGPT do when taking the medical licensing exams? The implications of large language models for medical education and knowledge assessment. Preprint at https://medRxiv.org/abs/2022.2012.2023.22283901

Gupta P, MacAvaney S (2022) On survivorship bias in MS MARCO. In: proceedings of the 45th International ACM SIGIR conference on research and development in information retrieval, pp 2214–2219

Hadi MU et al. (2023) A survey on large language models: Applications, challenges, limitations, and practical usage. Authorea Prepr 2023:1–31

Han T et al. (2023) MedAlpaca: an open-source collection of medical conversational AI models and training data. Preprint at https://arxiv.org/abs/08247

Haq HU et al. (2022) Mining adverse drug reactions from unstructured mediums at scale. Springer, Berlin, pp 361–375

Harrer S (2023) Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. EBioMedicine 90:1–12

He Y et al. (2023) Will ChatGPT/GPT-4 be a lighthouse to guide spinal surgeons? Ann Biomed Eng 51(2023):1362–1365

He Y et al. (2020) Infusing disease knowledge into BERT for health question answering, medical inference and disease name recognition. Preprint at https://arxiv.org/abs/03746

Herlihy C, Rudinger R (2021) MedNLI is not immune: Natural language inference artifacts in the clinical domain. Preprint at https://arxiv.org/abs/02970

Huang J, Tan M (2023) The role of ChatGPT in scientific communication: writing better scientific review articles. AJCR 13(4):1148

Huang X et al. (2023) A survey of safety and trustworthiness of large language models through the lens of verification and validation. Preprint at https://arxiv.org/abs/11391

Ilicki J (2023) A framework for critically assessing ChatGPT and other large language artificial intelligence model applications in health care. Mayo Clin Proc Digital Health 1(2):185–188

Javaid M et al. (2023) ChatGPT for healthcare services: an emerging stage for an innovative perspective. BenchCouncil Trans Benchmarks Stand Eval 3(1):100105

Jin JQ, Dobry AS (2023) ChatGPT for healthcare providers and patients: Practical implications within dermatology. J Am Acad Dermatol. https://doi.org/10.1016/j.jaad.2023.05.081

Jing C et al. (2022) Supplementing domain knowledge to BERT with semi-structured information of documents. Expert Syst Appl 2023:121054

Johnson AE et al. (2019) MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data 6(1):317

Kalpakchi D, Boye J (2023) SweCTRL-Mini: a data-transparent Transformer-based large language model for controllable text generation in Swedish. Preprint at https://arxiv.org/abs/13994

Kamphuis C et al. (2023) MMEAD: MS marco entity annotations and disambiguations. In: proceedings of the 46th International ACM SIGIR conference on research and development in information retrieval, pp 2817–2825

Kaneda Y et al. (2023) Assessing the performance of GPT-3.5 and GPT-4 on the Japanese nursing examination. Cureus 15(8):1–7

Ke Y et al. (2024) Development and testing of retrieval augmented generation in large language models: a case study report. Preprint at https://arxiv.org/abs/01733

Klang E et al. (2023) Utilizing artificial intelligence for crafting medical examinations: a medical education study with GPT-4. Researchsquare 613:423

Korngiebel DM, Mooney SD (2021) Considering the possibilities and pitfalls of generative pre-trained transformer 3 (GPT-3) in healthcare delivery. Npj Digit Med 4(1):93–95

Krishna K et al. (2020) Generating SOAP notes from doctor-patient conversations using modular summarization techniques. Preprint at https://arxiv.org/abs/01795

Kung TH et al. (2023) Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLoS Digit Health 2(2):e0000198

Lee J et al. (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36(4):1234–1240

Lee T-C et al. (2023) ChatGPT Answers Common Patient Questions About Colonoscopy. Gastroenterology 165(2023):509–511

Leng Y et al. (2023) Softcorrect: error correction with soft detection for automatic speech recognition. In: proceedings of the AAAI conference on artificial intelligence, pp 13034–13042

Levine D M et al. (2023) The diagnostic and triage accuracy of the GPT-3 artificial intelligence model. Preprint at https://medRxiv.org/abs/23285067

Lewis M et al. (2019) Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. Preprint at https://arxiv.org/abs/13461

Li W et al. (2023c) Revolutionizing neurosurgery with GPT-4: a leap forward or ethical conundrum? Ann Biomed Eng 51:1–8

Li Q et al. (2021) Discriminative neural clustering for speaker diarisation. In: 2021 IEEE spoken language technology workshop (SLT), pp 574–581

Li X et al. (2022a) Is gpt-3 a psychopath? evaluating large language models from a psychological perspective. Preprint at https://arxiv.org/abs/10529

Li J et al. (2022b) Unified named entity recognition as word-word relation classification. In: proceedings of the AAAI conference on artificial intelligence, pp 10965–10973

Li J et al. (2023a) Huatuo-26M, a large-scale Chinese medical QA dataset. Preprint at https://arxiv.org/abs/01526

Li J et al. (2023b) Assessing the performance of GPT-4 in the filed of osteoarthritis and orthopaedic case consultation. Preprint at https://medRxiv.org/abs/23293735

Liebrenz M et al. (2023) Generating scholarly content with ChatGPT: ethical challenges for medical publishing. Lancet Digit Health 5(3):e105–e106

Liévin V et al. (2022) Can large language models reason about medical questions?. Preprint at https://arxiv.org/abs/08143

Lin S et al. (2021) Graph-evolving meta-learning for low-resource medical dialogue generation. In: proceedings of the AAAI conference on artificial intelligence, pp 13362–13370

Lin J et al. (2023) How can recommender systems benefit from large language models: a survey. Preprint at https://arxiv.org/abs/05817

Liu W et al. (2021) Heterogeneous graph reasoning for knowledge-grounded medical dialogue system. Neurocomputing 442:260–268

Liu Y et al. (2023a) Summary of chatgpt/gpt-4 research and perspective towards the future of large language models. Preprint at https://arxiv.org/abs/01852

Liu H et al. (2023b) Evaluating the logical reasoning ability of chatgpt and gpt-4. Preprint at https://arxiv.org/abs/03439

Liu Z et al. (2023c) Deid-gpt: Zero-shot medical text de-identification by gpt-4. Preprint at https://arxiv.org/abs/11032

Liu Y et al. (2023d) Trustworthy LLMs: a survey and guideline for evaluating large language models' alignment. Preprint at https://arxiv.org/abs/05374

Lu Y et al. (2023) Artificial intelligence in intensive care medicine: toward a ChatGPT/GPT-4 Way? Ann Biomed Eng 51:1898–1903

Lyu Q et al. (2023) Translating radiology reports into plain language using chatgpt and gpt-4 with prompt learning: promising results, limitations, and potential. Vis Comput Ind Biomed 6(2023):1–10

Madrid-García A et al. (2023) Harnessing ChatGPT and GPT-4 for evaluating the rheumatology questions of the spanish access exam to specialized medical training. Preprint at https://medRxiv.org/abs/23292821

Mantas J (2022) Length of stay prediction in neurosurgery with Russian GPT-3 language model compared to human expectations. IOS press, Amsterdam

Miao J et al. (2023) Assessing the accuracy of ChatGPT on core questions in glomerular disease. Kidney Int Rep 8:1657–1659

Munn L et al. (2023) Truth machines: synthesizing veracity in AI language models. Preprint at https://arxiv.org/abs/12066

Muse H et al. (2023) Pre-training with scientific text improves educational question generation (student abstract). In: proceedings of the aaai conference on artificial intelligence, pp 16288–16289

Nanayakkara G et al. (2022) Clinical dialogue transcription error correction using Seq2Seq models. Springer, Berlin, pp 41–57

Nashwan AJ et al. (2023) Embracing the future of physician-patient communication: GPT-4 in gastroenterology. Gastroent Endosc 1(2023):132–135

Nath S et al. (2022) New meaning for NLP: the trials and tribulations of natural language processing with GPT-3 in ophthalmology. Br J Ophthalmol 106(7):889–892

Nori H et al. (2023) Capabilities of gpt-4 on medical challenge problems. Preprint at https://arxiv.org/abs/13375

Omran S et al. (2023) Effectiveness of pharmacogenomics educational interventions on healthcare professionals and health professions students: a systematic review. Res Soc Adm Pharm. https://doi.org/10.1016/j.sapharm.2023.07.012

Passos M et al. (2022) Decision models on therapies for intensive medicine. Procedia Comp Sci 210:230–235

Patel SB, Lam K (2023) ChatGPT: the future of discharge summaries? Lancet Digit Health 5(3):e107–e108

Peng S et al. (2023) AI-ChatGPT/GPT-4: an booster for the development of physical medicine and rehabilitation in the new era! Ann Biomed Eng 62:1–5

Pooch E H et al. (2020) Can we trust deep learning based diagnosis? the impact of domain shift in chest radiograph classification. In: thoracic image analysis: second international workshop, TIA 2020, held in conjunction with MICCAI 2020, Lima, Peru, October 8, 2020, proceedings 2, pp 74–83

Portelli B et al. (2021) Improving adverse drug event extraction with SpanBERT on different text typologies. In: international workshop on health intelligence, pp 87–99

Pradeep R et al. (2020) Scientific claim verification with VerT5erini. Preprint at https://arxiv.org/abs/11930

Quesado I et al. (2022) Data mining models for automatic problem identification in intensive medicine. Procedia Comp Sci 210:218–223

Rao A et al. (2023) Evaluating ChatGPT as an adjunct for radiologic decision-making. Preprint at https://medRxiv.org/abs/23285399

Raval S et al. (2021) Exploring a unified sequence-to-sequence transformer for medical product safety monitoring in social media. Preprint at https://arxiv.org/abs/05815

Ray PP (2023) Benchmarking, ethical alignment, and evaluation framework for conversational AI: advancing responsible development of ChatGPT. BenchCouncil Trans Benchmarks Stand Eval 3:100136

Reddy S (2023) Evaluating large language models for use in healthcare: A framework for translational value assessment. Inform Med Unlocked 41(2023):101304

Rosol M et al. (2023) Evaluation of the performance of GPT-3.5 and GPT-4 on the Medical Final Examination. Preprint at https://medRxiv.org/abs/23290939

Roy K et al. (2023) Proknow: Process knowledge for safety constrained and explainable question generation for mental health diagnostic assistance. Front Big Data 5:1056728

Roy K, Rawte V (2022) TDLR: top semantic-down syntactic language representation. In: NeurIPS'22 workshop on all things attention: bridging different perspectives on attention

Roy S et al. (2021) Knowledge-aware neural networks for medical forum question classification. In: proceedings of the 30th ACM international conference on information & knowledge management, pp 3398–3402

Sakhovskiy A, Tutubalina E (2022) Multimodal model with text and drug embeddings for adverse drug reaction classification. J Biomed Inform 135:104182

Sallam MJ (2023) The utility of ChatGPT as an example of large language models in healthcare education, research and practice: systematic review on the future perspectives and potential limitations. medRxiv. https://doi.org/10.1101/2023.02.19.23286155

Sallam M (2023a) ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare 11(6):887

Sallam M (2023b) The utility of ChatGPT as an example of large language models in healthcare education, research and practice: systematic review on the future perspectives and potential limitations. Healthcare 11:887

Scepanovic S et al. (2020) Extracting medical entities from social media. In: Proceedings of the ACM conference on health, inference, and learning, pp 170–181

Schloss B, Konam S (2020) Towards an automated SOAP note: classifying utterances from medical conversations. In: machine learning for healthcare conference, pp 610–631

Selvaraj SP, Konam S (2020) Medication regimen extraction from medical conversations. Springer, Berlin, pp 195–209

Sharma B et al. (2023) Multi-task training with in-domain language models for diagnostic reasoning. Preprint at https://arxiv.org/abs/04551

Sheth A et al. (2022) Process knowledge-infused AI: toward user-level explainability, interpretability, and safety. IEEE Internet Comput 26(5):76–84

Shin H-C et al. (2020) BioMegatron: Larger biomedical domain language model. Preprint at https://arxiv.org/abs/06060

Singhal K et al. (2023a) Large language models encode clinical knowledge. Nature 620:1–9

Singhal K et al. (2023b) Towards expert-level medical question answering with large language models. Preprint at https://arxiv.org/abs/09617

Soroush A et al. (2023) Assessing GPT-3.5 and GPT-4 in generating international classification of diseases billing codes. Preprint at https://medRxiv.org/abs/23292391

Stanciu A (2023) Data management plan for healthcare: following FAIR principles and addressing cybersecurity aspects. a systematic review using instructGPT. Preprint at https://medRxiv.org/abs/23288932

Takagi S et al. (2023) Performance of GPT-35 and GPT-4 on the Japanese medical licensing examination: comparison study. JMIR Med Educ 9(1):e48002

Tao Y-T et al. (2020) Predicted rat interactome database and gene set linkage analysis. Database. https://doi.org/10.1093/database/baaa086

Tao Y-T et al. (2022) Genome-wide identification and analysis of bZIP gene family reveal their roles during development and drought stress in wheel wingnut (Cyclocarya paliurus). BMC Genom 23(1):743

Thawkar O et al. (2023) Xraygpt: Chest radiographs summarization using medical vision-language models. Preprint at https://arxiv.org/abs/07971

Thirunavukarasu AJ et al. (2023) Large language models in medicine. Nat Med 8:1–11

Tian S et al. (2023) Opportunities and challenges for ChatGPT and large language models in biomedicine and health. Preprint at https://arxiv.org/abs/10070

Toma A et al. (2023) Clinical Camel: An open-source expert-level medical language model with dialogue-based knowledge encoding. Preprint at https://arxiv.org/abs/12031

Tran TNT et al. (2021) Recommender systems in the healthcare domain: state-of-the-art and research issues. J Intell Inf Syst 57:171–201

Ueda D et al. (2023) Evaluating GPT-4-based ChatGPT's clinical potential on the NEJM Quiz. Preprint at https://medRxiv.org/abs/23289493

Waisberg E et al. (2023) GPT-4: a new era of artificial intelligence in medicine. Ir J Med Sci 51(2023):1645–1653

Wang D, Chen Y (2021) A novel cascade hybrid many-objective recommendation algorithm incorporating multistakeholder concerns. Inform Sci 577:105–127

Wang D, Zhao X (2022) Affective video recommender systems: a survey. Front Neurosci 16:984404

Wang Y et al. (2020) MedSTS: a resource for clinical semantic textual similarity. Language Resour Eval 54:57–72

Wang Y et al. (2023a) Are large language models ready for healthcare? a comparative study on clinical language understanding. Proc Mach Learn Res 219:1–24

Wang S et al. (2020b) Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii. In: proceedings of the ACM conference on health, inference, and learning, pp 222–235

Wang Z et al. (2023b) Can LLMs like GPT-4 outperform traditional AI tools in dementia diagnosis? Maybe, but not today. Preprint at https://arxiv.org/abs/01499

Wang H et al. (2023c) Huatuo: Tuning llama model with chinese medical knowledge. Preprint at https://arxiv.org/abs/06975

Wang Y et al. (2023d) Are large language models ready for healthcare? A comparative study on clinical language understanding. Preprint at https://arxiv.org/abs/05368

Wei C et al. (2023) An overview on language models: recent developments and outlook. Preprint at https://arxiv.org/abs/05759

Willett FR et al. (2023) A high-performance speech neuroprosthesis. Nature 620:1–6

Wornow M et al. (2023) The shaky foundations of large language models and foundation models for electronic health records. Npj Digit Med 6(1):135

Wu T et al. (2022) AI chains: transparent and controllable human-AI interaction by chaining large language model prompts. In: Proceedings of the 2022 CHI conference on human factors in computing systems, pp 1–22

Wu C et al. (2023) Pmc-llama: further finetuning llama on medical papers. Preprint at https://arxiv.org/abs/14454

Wu Z et al. (2024) KnowLab_AIMed at MEDIQA-CORR 2024: Chain-of-Though (CoT) prompting strategies for medical error detection and correction. In: proceedings of the 6th clinical natural language processing workshop, pp 353–359

Xie Q et al. (2023) Faithful AI in medicine: a systematic review with large language models and beyond. medRxiv.

Xiong H et al. (2023) Doctorglm: fine-tuning your chinese doctor is not a herculean task. Preprint at https://arxiv.org/abs/01097

Yadav S et al. (2022) Question-aware transformer models for consumer health question summarization. J Biomed Inform 128:104040

Yadav S et al. (2023) Towards understanding consumer healthcare questions on the web with semantically enhanced contrastive learning. Proc of the ACM Web Conf 2023:1773–1783

Yadav S et al. (2021) Transfer learning-based approaches for consumer question and multi-answer summarization. In: proceedings of the 20th workshop on biomedical language processing, pp 291–301

Yadav S et al. (2022a) Chq-summ: a dataset for consumer healthcare question summarization. Preprint at https://arxiv.org/abs/06581

Yang X et al. (2022) A large language model for electronic health records. Npj Digit Med 5:194–203

Yang J et al. (2023) The Impact of ChatGPT and LLMs on medical imaging stakeholders: perspectives and use cases. MetaRadiology 1:100007

Yang X et al. (2020) COVID-CT-dataset: a CT scan dataset about COVID-19. Preprint at https://arxiv.org/abs/13865

Yang S et al. (2023a) Zhongjing: enhancing the Chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. Preprint at https://arxiv.org/abs/03549

Yuan H et al. (2022) BioBART: pretraining and evaluation of a biomedical generative language model

Yunxiang L et al. (2023) Chatdoctor: a medical chat model fine-tuned on llama model using medical domain knowledge. Cureus. https://doi.org/10.7759/cureus.40895

Zack T et al. (2023) Coding inequity: assessing GPT-4's potential for perpetuating racial and gender biases in healthcare. Preprint at https://medRxiv.org/abs/23292577

Zhang T et al. (2021) Adversarial neural network with sentiment-aware attention for detecting adverse drug reactions. J Biomed Inform 123:103896

Zhang S et al. (2023a) MTDAN: a lightweight multi-scale temporal difference attention networks for automated video depression Detection. IEEE Trans Affect Comput. https://doi.org/10.1109/TAFFC.2023.3312263

Zhang S et al. (2023a) Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: a systematic review of recent advancements and future prospects. Expert Syst Appl 237:121692

Zhang S et al. (2023b) Multimodal emotion recognition based on audio and text by using hybrid attention networks. Biomed Signal Proc 85:105052

Zhang Y et al. (2023c) Chat generative pre-trained transformer (ChatGPT) usage in healthcare. Gastroent Endosc 1(3):139–143

Zhang L, Liu J (2022) Intent-aware Prompt Learning for medical question summarization. In: 2022 IEEE international conference on bioinformatics and biomedicine (BIBM), pp 672–679

Zhang J et al. (2022) Fengshenbang 1.0: being the foundation of chinese cognitive intelligence. Preprint at https://arxiv.org/abs/02970

Zhang H et al. (2023e) HuatuoGPT, towards Taming Language Model to Be a Doctor. Preprint at https://arxiv.org/abs/15075

Zhang J et al. (2023d) The potential and pitfalls of using a large language model such as ChatGPT or GPT-4 as a clinical assistant. Preprint at https://arxiv.org/abs/08152

Zhao WX et al. (2023) A survey of large language models. arXiv. https://doi.org/10.48550/arXiv.2303.18223

Zhao X, Vydiswaran V V (2021) Lirex: Augmenting language inference with relevant explanations. In: proceedings of the AAAI conference on artificial intelligence, pp 14532–14539

Zhou S, Zhang Y (2021) Datlmedqa: a data augmentation and transfer learning based solution for medical question answering. Appl Sci 11(23):11251

Zhu W et al. (2019) Panlp at mediqa 2019: Pre-trained language models, transfer learning and knowlsedge distillation. ACL Anthology. https://doi.org/10.18653/v1/W19-5039