Journal of Healthcare Informatics Research
https://doi.org/10.1007/s41666-024-00171-8

RESEARCH ARTICLE

Large Language Models in Biomedical and Health Informatics: A Review with Bibliometric Analysis

Huizi Yu[1] · Lizhou Fan[1] · Lingyao Li[1] · Jiayan Zhou[2] · Zihui Ma[3] · Lu Xian[1] · Wenyue Hua[4] · Sijia He[1] · Mingyu Jin[4] · Yongfeng Zhang[4] · Ashvin Gandhi[5] · Xin Ma[6]

Received: 24 May 2024 / Revised: 16 August 2024 / Accepted: 22 August 2024

Abstract

Large language models (LLMs) have rapidly become important tools in Biomedical and Health Informatics (BHI), potentially enabling new ways to analyze data, treat patients, and conduct research. This study aims to provide a comprehensive overview of LLM applications in BHI, highlighting their transformative potential and addressing the associated ethical and practical challenges. We reviewed 1698 research articles from January 2022 to December 2023, categorizing them by research themes and diagnostic categories. Additionally, we conducted network analysis to map scholarly collaborations and research dynamics. Our findings reveal a substantial increase in the potential applications of LLMs to a variety of BHI tasks, including clinical decision support, patient interaction, and medical document analysis. Notably, LLMs are expected to be instrumental in enhancing the accuracy of diagnostic tools and patient care protocols. The network analysis highlights dense and dynamically evolving collaborations across institutions, underscoring the interdisciplinary nature of LLM research in BHI. A significant trend was the application of LLMs in managing specific disease categories, such as mental health and neurological disorders, demonstrating their potential to influence personalized medicine and public health strategies. LLMs hold promising potential to further transform biomedical research and healthcare delivery. While promising, the ethical implications and challenges of model validation call for rigorous scrutiny to optimize their benefits in clinical settings. This survey serves as a resource for stakeholders in healthcare, including researchers, clinicians, and policymakers, to understand the current state and future potential of LLMs in BHI.

Keywords Artificial intelligence · Biomedical informatics · Health informatics · Large language models

Huizi Yu, Lizhou Fan contributed equally to this work.

Extended author information available on the last page of the article

Published online: 14 September 2024

1 Introduction

Large language models (LLMs) have emerged as pivotal technologies, redefining the landscape of natural language processing (NLP) and showing significant potential in the intersection of artificial intelligence (AI) and other domains, such as Biomedical and Health Informatics (BHI) [1–3]. The advent of groundbreaking models, including OpenAI's Generative Pre-trained Transformer (GPT) [4] has demonstrated its capabilities to process, understand, and generate human-like text by leveraging extensive datasets and sophisticated neural network architectures [5, 6]. These advances have set the stage for transformative applications within BHI, a domain where the accuracy and nuance of language understanding significantly impact patient care, medical research, and healthcare delivery.

Since the introduction of models like ChatGPT, the role of LLMs in BHI has been increasingly recognized. These potential applications include clinical decision support, patient engagement enhancement, and medical literature analysis [7–9]. These developments have provided enormous possibilities for not only augmenting traditional methodologies but also paving the way for novel approaches to addressing complex challenges in the healthcare sector.

Our review uniquely contributes to the discourse by offering a comprehensive analysis of LLM applications in BHI in 1698 papers from January 2022 to December 2023. Through an examination of research themes, scholarly networks, and the evolution of LLM technologies, we delve into the integration and impact of LLMs across various BHI fields. The scope of this study is twofold:

- *Research themes and topics*: We explore the development of LLM algorithms through the lenses of NLP and medical tasks, as well as the LLMs applications in various disease categories, identifying LLM-based applications in BHI.
- *Scholarly networks and partnerships*: Our analysis includes an examination of the collaborative efforts and research networks, underlying the dynamics of research paradigms of LLM research in the BHI domains.

By examining current literature, this review aims to highlight key trends and gaps in current research and further points out the opportunities. Our findings aim to provide a foundation for future research, giving stakeholders important insights to understand and contribute to this rapidly developing field. This review not only shows the enormous prospects of LLMs improving healthcare outcomes but also emphasizes the need to consider ethics and address practical challenges in the case of using LLMs in BHI.

The rest of the paper is organized as follows: We begin by providing background on the intersection of LLMs and BHI from three perspectives, i.e. the evolution of LLMs, their applications in BHI, and the synthesized knowledge of LLMs in BHI (Sect. 2). The methods section outlines our review approaches (Sect. 3), including data collection and description, topic classification, network analysis, and visualization techniques employed. The result sections are organized in an overall-to-specific manner. First, we provide a two-fold overview

(Sect. 4): the first fold is about content analysis, focusing on research themes and topics; the second one is on network analyses, focusing on scholarly networks and partnerships. Based on the analysis of *research themes and topics*, we further highlight three findings, including (1) *the distributed methodologies* (Sect. 5), (2) *the diverse prospects of LLM applications* (Sect. 6), and (3) *specific disease categories* where LLMs have shown promise (Sect. 7). Finally, the conclusions and discussion section (Sect. 8) summarizes our key findings, addresses limitations, and provides recommendations for future work in this rapidly evolving field[1].

2 Backgrounds

The intersection of LLMs and BHI represents a frontier of innovation. To better understand the application prospects of LLMs in the BHI domain, we conducted a background investigation from three perspectives: (1) *the evolution of LLMs*, (2) *applications of LLMs in the domain of BHI*, and (3) *synthesized knowledge of LLMs in BHI*.

2.1 Evolution of Large Language Models

LLMs represent a sophisticated category of language models that utilize neural networks with multi-billion parameter architectures. These models are trained on vast unlabeled textual data using self-supervised learning techniques [10, 11]. An earlier milestone was made in 2017 when Google released the Transformer model. This model introduced the self-attention mechanism, which was fundamental for LLMs by capturing contextual relationships and nuanced information among input tokens [12]. Following this model, the introduction of Bidirectional Encoder Representations from Transformers (BERT) in 2018 was another milestone that revolutionized the way that machines understand human language [13].

Later, the evolution of LLMs witnessed a significant moment with the release of OpenAI's GPT-3 in 2020, which has been widely regarded as a game-changer in the field. Having trained using 175 billion parameters, GPT-3's transformer-based model demonstrated an unprecedented capacity for generating text that resembles human writing [14]. This period also gave rise to other significant models such as T5 [15], ERNIE [16], and EleutherAI's GPT-Neo [17], each contributing uniquely to the LLM landscape.

In recent years, the development of LLMs has pivoted towards enhancing both efficiency and contextual understanding. This shift has unlocked more sophisticated and nuanced applications [18, 19]. In particular, recent models are not only linguistically adept but also integrate multimodal capabilities, processing both text and other forms of data [20]. This advancement has led to the emergence of various generative AI models, both in closed-source and open-source domains. Prominent closed-source LLMs include ChatGPT by OpenAI [4], Claude 2 by Anthropic [21], and Gemini by Google [22]. Typical models in the open-source domain include LLaMa 2 by Meta [23] and Phi-family models by Microsoft [24].

[1] We also provide the workflow and relations among sections in Appendix 1.

2.2 Applications of LLMs in BHI

In the early stages, NLP applications in BHI primarily focused on extracting and categorizing information from electronic medical records and medical literature. These applications aimed to improve information retrieval [25, 26], learn semantic relations of clinical text [27], and train word embeddings [28, 29]. These early implementations of NLP have set the stage for the integration of sophisticated models that could handle a broader range of linguistic tasks.

With the advancement of LLMs, the scope of NLP in healthcare has expanded dramatically. In particular, the research on the BERT model in BHI has transitioned from rule-based text processing to more advanced applications [30]. One of its notable applications is text classification, where BERT's contextual analysis significantly enhances the accuracy of categorizing clinical notes, research papers, and patient feedback into relevant medical categories [31–34]. The BERT model has been extensively applied in named entity recognition (NER) and relation extraction within the BHI domain [35–37]. In addition, there has been significant progress in fine-tuning the BERT model for specific applications within BHI. Noteworthy among these are BioBERT and ClinicalBERT, introduced by [38] and [39], respectively.

Compared to BERT models, the advanced LLMs have shown general-purpose capabilities, which enable them to excel across a broad set of NLP tasks in BHI [40], rather than being designed solely for a single NLP task, such as NER or text classification. For example, LLMs have shown potential for interpreting complex patient data and suggesting medical diagnoses [41–45]. This capability is useful for synthesizing unstructured patient information and supporting clinical decisions. They are also integral to drug-disease identification and drug discovery, where they have shown promise in identifying drug candidates and their effects [46, 47]. In addition, the customization abilities of LLMs have unlocked new possibilities in medical education [48–51]. These models could adapt to the learning pace and style of individual students, providing personalized learning experiences.

Among these applications, there are several studies to highlight. For example, Kung et al. [52] evaluated the performance of ChatGPT on the United States Medical Licensing Exam (USMLE). Their findings revealed that ChatGPT achieved scores at or near the passing threshold across all three sections of the exam without any training or reinforcement. Singhal et al. [1] proposed an approach for the evaluation of LLMs in the context of medical question-answering. Their study showed the promise of LLMs in clinical knowledge and question-answering capabilities.

2.3 Synthesized Knowledge of LLMs in BHI

Several review papers on applications of LLMs in BHI have appeared [40, 53–57]. We present an overview of the reviewed papers in Table 1. Two of the earliest review papers of applied research on LLMs in BHI surveyed how LLM applications could be developed and leveraged in clinical settings [40].

As a systematic review of ChatGPT in healthcare, Li et al. [53] selected papers on PubMed with keywords "ChatGPT." A two-sided taxonomy (application-oriented

Table 1 Representative review studies using NLP in the domain of biomedical informatics

Paper	Type of review	Paper count	Scope	Contribution	Limitations	Bibliometric analysis
[40]	Commentary	None	None	One of the earliest reviews; discusses the mixed results of LLMs in medical contexts	No systematic paper collection and analysis framework; subjective opinion potentially introduces bias	No
[57]	Commentary	None	None	Describes opportunities and pitfalls associated with employing LLMs in biomedical research	No systematic paper collection and analysis framework; subjective opinion potentially introduces bias	No
[53]	Systematic review	58	Paper on PubMed with keyword "ChatGPT"	Provides a 3-level application- and user-oriented taxonomy	Limited scope (only research about ChatGPT); could omit most emergent research by only including publications on PubMed	No
[54]	Comprehensive survey	582	Paper on PubMed with keywords "large language models"/"ChatGPT"	Comprehensive analysis of the diverse applications, including information retrieval, question answering, medical text summarization, information extraction, and medical education	Limited scope (missing some important applications such as multimodal LLM); could omit most emergent research by only including publications on PubMed	No
[55]	Bibliometric review	5752	Paper on Web of Science with keywords: (("large"/"big"/"massive") AND ("language model"/"language models")) AND ("BERT"/"GPT-1"/"GPT-2"/"GPT-3"/"ChatGPT")	Identifies patterns in research paradigms, collaboration networks, and thematic trends in LLM research, covering core algorithm developments, NLP tasks, and diverse applications across fields such as medicine, engineering, social sciences, and humanities	Not dedicated to reviewing papers in the field of BHI; review could be too broad	Yes

Table 1 (continued)

Paper	Type of review	Paper count	Scope	Contribution	Limitations	Bibliometric analysis
[56]	Systematic review	329	Paper on OpenAlex with (large language model/LLM/BERT/RoBERTa/T5/XLNet/Mistral/Mixtral/Falcon/Qwen/BLOOM/Vicuna/LLaMA/GPT/Claude/Bard/Google PaLM/Gemini) and (electronic medical record/electronic health record)	Categorizes and discusses the reviewed papers into seven major topics: named entity recognition, information extraction, text similarity, text summarization, text classification, dialogue systems, and diagnosis and prediction	Limited amount of papers included	No

and user-oriented) was provided to categorize three levels of papers (generic comment about the applications in healthcare as level 1, one or more example uses in specific medical specialty as level 2, and qualitative and quantitative evaluation of ChatGPT in a specialty as level 3). The comprehensive survey by Tian et al. [54] particularly focused on the areas of biomedical information retrieval, question answering, medical text summarization, information extraction, and medical education. Their study found significant advances made in the area of text generation but modest advances in other applied research in BHI, such as multimodal LLM. Moreover, they selected papers with keywords LLM and ChatGPT only within PubMed. Some emergent research may be omitted because of the limited scope of PubMed.

Additionally, although the review papers about LLMs involved multiple electronic resource libraries [54, 55], applied research on LLMs in healthcare was only one aspect of their broader research. Conversely, Li et al. [56] solely focused on the applied study of LLMs in investigating electronic health records (EHRs). They categorized 329 papers on OpenAlex with LLM keywords (LLM, Bert, et al., and Electronic Medical Record, et al.) into seven major topics: named entity recognition, information extraction, text similarity, text summarization, text classification, dialogue systems, and diagnosis and prediction. However, concentrating only on EHRs might not fully explore the broader impact and versatility of LLMs in various facets of healthcare, including clinical decision support and medical imaging analysis.

Our survey of 1698 papers with bibliometric analysis offers several distinct advantages by providing a more comprehensive and systematic examination of the current state of applied research on LLMs in BHI. We employ a hybrid approach that not only offers a panoramic overview of the field but also facilitates a detailed exploration of specific research themes. This includes both general LLM research themes and their applied research on major diagnostic categories within BHI domains. By integrating a bibliometric analysis, we could be able to quantify and visualize trends, research hotspots, and the impact of various studies, providing a data-driven perspective that enhances the depth and rigor of our review.

Another key advantage of our survey is its dedicated focus on emerging LLMs, specifically the ChatGPT model family. This allows us to delve deeply into the unique characteristics and capabilities of these models, which are at the forefront of technological advancement in natural language processing. By concentrating on these state-of-the-art models, we provide valuable insights that are directly relevant to the current and future applied studies of LLMs in BHI. Recently, multimodal LLMs have emerged in the domain of BHI, which integrate and process multiple modal data types such as text and images and offer significant potential for more comprehensive and accurate data analysis, diagnosis, and personalized treatment planning. Our survey highlights the transformative potential of these multimodal models and underscores the need for further exploration and application in the field of BHI.

Our investigation into the existing review papers highlights a research gap in the literature: there remains a need for a survey that encapsulates the full spectrum of LLM developments and their specific applications. Our review paper stands out for its multifaceted contributions. Firstly, it offers a detailed survey and bibliometric analysis of the latest LLM applications in BHI, providing a perspective on the evolving trends and challenges within this field. Secondly, the data-driven nature of our review allows for a deeper understanding of the interdisciplinary connections within the published literature and assists in locating the key contributors through semantic network analysis. Thirdly, unlike previous reviews that may have concentrated on particular facets, our work presents a holistic perspective on the trajectory of LLMs in BHI, elucidating how these models have both shaped and been shaped by the needs and advancements in biomedical sciences and health practices.

1. It offers a detailed survey and bibliometric analysis of the latest LLMs' applied research in BHI, providing a perspective on the evolving trends and challenges within the BHI field.
2. The data-driven nature of our review allows for a deeper understanding of the interdisciplinary connections within the published literature and assists in locating the key contributors through semantic network analysis.
3. Unlike previous reviews that may have provided an overview of LLM in multiple fields (e.g., engineering, humanities) [55] or one particular domain within BHI (e.g., EHR) [56], our work presents a holistic perspective on the trajectory of LLMs in BHI, elucidating how these models have both shaped and been shaped by the needs and advancements in biomedical sciences and health practices.

3 Methods

In this section, we provide an overview of the methodologies employed in the review, which include data collection and description, topic classification, network analysis, and visualization techniques.

3.1 Data Collection and Analytics Workflow

Figure 1 shows the data retrieval, cleaning, and analysis workflow. In this study, the primary data source is OpenAlex, a comprehensive database known for its extensive collection of academic publications. OpenAlex includes both published papers and preprints on platforms like arXiv and medRxiv. This feature allows us to access a broader range of research, including early-stage findings and contributions yet to be peer-reviewed, thereby enriching our dataset with a wider variety of scholarly work. The specific query[2] employed to extract relevant data was:

$$\text{(large language model OR GPT) AND (health OR medical)} + [2022 - 2023]$$

This query was chosen to ensure the inclusion of relevant documents that discuss or mention LLMs, including GPT, in the context of health or medical fields. The time frame of 2022–2023 was selected to gather the most recent and relevant insights[3]. Importantly, the decision to avoid explicitly including model names such

[2] We conducted the data collection on 01/05/2024.

[3] Some papers that were officially published in 2024 had their original versions published on arXiv in either 2022 or 2023.

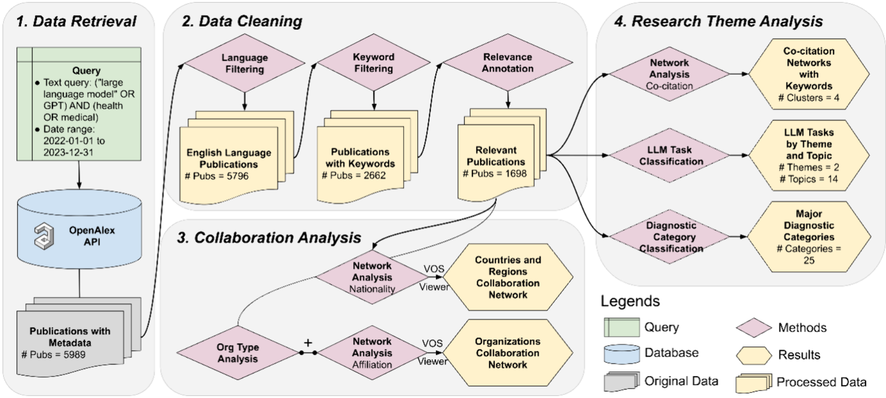

Fig. 1 Data selection and analytic workflow

as "llama" in the query was deliberate. While "llama" is a term associated with certain models (i.e., Llama by Meta) [23], it also commonly refers to the animal. Including it could dilute the relevance and focus of the research. By structuring the query in this manner, we were able to efficiently isolate documents that are specifically relevant to the intersection of LLMs like GPT and health or medical studies, without the interference of unrelated topics.

Following this, we implemented a more focused local restriction based on the English-language papers and terms "large language model" or "GPT." According to the information on the OpenAlex help page, the API scans through titles, abstracts, and full texts of documents while searching. However, it employs techniques like the removal of stop words and the use of stemming (specifically, the Kstem token filter) to enhance search results. Although these techniques are generally effective, they could sometimes lead to the inclusion of non-relevant documents, particularly after the stemming process. To counteract this issue, we performed a second round of cleaning, aiming to retain only those documents that explicitly mention the model query terms in their titles and abstracts. This step was crucial in refining the results to ensure the relevance and precision of our dataset.

The final step in our filtering process involved the removal of irrelevant papers through human annotation. Even with the advanced algorithmic filtering, some false positives—particularly non-health and non-medical articles—may be retained. To address this issue, we engaged two human annotators who independently reviewed the dataset. Their task was to identify and eliminate any remaining irrelevant papers. After this independent annotation, we measured the agreement rate between the two annotators, which stood at 96%. This human element of the filtering process was vital in ensuring the highest possible accuracy and relevance of the final collection of 1698 papers for our research.

3.2 Topic Classification for Content Analysis

3.2.1 RoBERTa Text Classification

For the paper topic classification task, we employed the "roberta-large-mnli" model, a pre-trained transformer-based neural network designed for natural language understanding tasks. While unsupervised methods like topic modeling are generally valuable for exploratory analysis, we've empirically tried a SBERT-based topic modeling method named BERTopic, but the specificity of the BHI domain made the general semantic-based unsupervised clustering hard to distinguish topics [58, 59]. Another reason for the challenge with SBERT-based models is that the clusters are not as distinguishable for closely related tasks within one field compared to more distinct topics across multiple fields, such as engineering and social science [60]. This lack of clear differentiation further complicates the effective classification of BHI research topics using unsupervised methods. Additionally, SBERT methods typically do not classify a single paper into multiple categories, which can be a significant limitation given that research papers often span multiple topics [61].

Instead, supervised models such as RoBERTa offer enhanced precision for well-defined categories. Specifically, we choose roberta-large-mnli for its high performance on the Multi-Genre Natural Language Inference (MNLI) benchmark and capability to leverage pre-trained knowledge, which makes it well-suited for zero-shot learning tasks [62–65]. This model is especially adept at categorizing LLM research papers, which may encompass multiple topics within a single document.

The zero-shot classification process involved defining a set of target topics related to LLMs, such as "model evaluation," "sentiment analysis," "education," and "ethics." There are 14 topics in total, selected by combining research themes of prominent NLP conferences, such as Empirical Methods in Natural Language Processing (EMNLP) and Association for Computational Linguistics (ACL). The final topic list was reviewed by three researchers independently. The purpose of the selection of these topics was to capture a wide spectrum of impactful applied research on LLMs in BHI. We have carefully curated the 14 topics from the original sub-domain lists of EMNLP and ACL. In addition, while these topics may not cover every aspect of the literature corpus, they represent key areas of interest and innovation in the applied research of LLMs in the BHI field.

Using the roberta-large-mnli model, each title and abstract was classified into one or more of the 14 predefined topics. The model inferred the relevance of each topic to a given text by predicting the likelihood that the text would be a hypothetical premise for a human-written hypothesis representing each topic[4]. To select the most likely set of predefined topics, we restrict the likelihood to be above 0.1[5].

[4] In our analysis, the hypothesis is "The topic of this paper is {}." The classification did not require any fine-tuning or training on a labeled dataset, as the model leveraged its pre-trained knowledge to make inferences about the unseen topics.

[5] We tested various thresholds by sampling 100 papers to manually inspect their relevance. The threshold of 0.1 was chosen to balance between specificity and sensitivity in the zero-shot classification process.

3.2.2 Major Diagnostic Categories

To evaluate the applied research of LLMs in medical domains, we extracted the specific diseases and symptoms from paper abstracts and grouped them into their corresponding Major Diagnostic Categories (MDC). The MDC is a system of classification that organizes diseases and medical conditions into 25 mutually exclusive diagnosis areas that are related to the affected organ system or the etiology of the condition. As the diseases and symptoms mentioned in the abstract directly align with the specific research objectives or questions each study aims to address, this process classifies research papers into their corresponding broader diagnostic categories[6]. For example, epilepsy, Parkinson's disease, and Alzheimer's disease are under "nervous system" disorders.

Specifically, we employed a multi-step approach to categorize diseases mentioned in abstracts, ensuring accuracy and reliability with collaborative and systematic methods. First, two researchers with biomedical backgrounds reviewed the abstract and identified mentions of disease, disorder, symptoms, and public health crises. Following the identification phase, another pair of researchers group the identified diseases, disorders, and symptoms into their corresponding MDC. Next, to ensure the reliability and consistency of the categorization process, an intercoder reliability check is performed with Cohen's Kappa of 0.9. We then include a third annotator, who is an experienced researcher in the BHI fields, to judge the annotation result and resolve discrepancies in data labeling.

3.3 Network Analysis Algorithm and Visualization

To construct the bibliometric networks, we employed the VOSviewer [66] software. These networks' entities include organizations, researchers, or individual publications, and the analysis is based on co-citation, bibliographic coupling, or co-authorship relations. VOSviewer utilizes a clustering algorithm based on the Visualization of Similarities (VOS) technique, which effectively maps and visualizes complex bibliometric networks. This algorithm begins by calculating the similarity between items (such as publications, authors, or journals) based on criteria such as co-citation or co-authorship. These similarities then form a matrix, which is used to spatially arrange items that reflect their mutual similarities. Leveraging modularity-based techniques, the algorithm groups items into clusters, which allows for an intuitive representation of the relationships and patterns within BHI. In each network, the size of the node represents the total link strength[7], indicating the cumulative strength of the connections an entity has with entities. The edge represents the connections or links between the nodes, illustrating the specific relationships such as co-citation, bibliographic coupling, or co-authorship.

[6] For detailed disease to MDC mapping, refer to Table 1.

[7] The mathematical definition of total link strength is provided in Appendix 2.

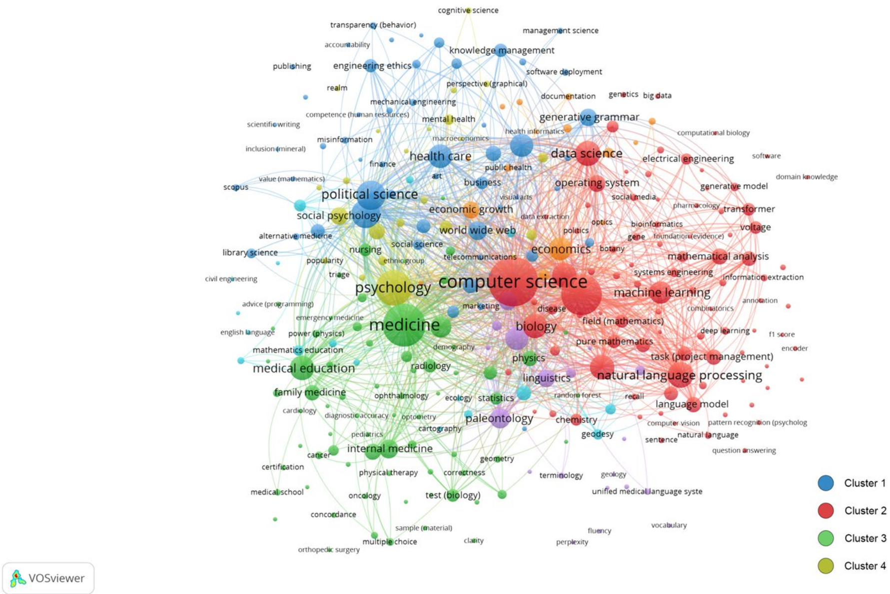

Fig. 2 Keyword co-occurrence network

4 Mapping the Terrain: an Overview of the Diverse Ecosystem of LLM Research in BHI

This section delves into the comprehensive landscape of LLM research within the realm of BHI. Our exploration is structured into two sections: first, the core research themes and topics employing LLMs, and second, the scholarly networks and partnerships that facilitate this research. Through the overview, we identify representative papers that exemplify significant developments and findings. These selected papers are discussed in subsequent results sections (Sects. 5–7) to highlight their contributions and innovations.

4.1 Research themes and topics

In the burgeoning field of BHI, LLMs have emerged as pivotal tools, enabling the transformation of data into actionable insights. As shown in Fig. 2, the keyword co-occurrence network adeptly represents the diverse research themes and topics that converge in this multidisciplinary domain[8]. At the center of this complex network lies the interdisciplinary interplay between technologies and BHI fields: social science (cluster 1: blue), computer science (cluster 2: red), biomedical science (cluster 3: green), and psychological science (cluster 4: yellow). Their synergy illustrates the multifaceted nature of the application of LLMs in BHI research.

[8] Appendix 2 shows the top 50 keywords in the network ranked by the total link strength in descending order.

Cluster 1 highlights the social implications of deploying LLMs in the biomedical and health sciences, including terms such as "engineering ethics," "data transparency," and "knowledge management," which are indicative of a keen awareness of the social dimensions intrinsic to the deployment of technology in sensitive fields. Cluster 2 is strongly associated with the core technical disciplines of LLMs, such as computer science and mathematics. This cluster's prominence underscores a significant research focus on the theoretical and computational foundations that are necessary for the development and refinement of LLM algorithms. The high level of connectivity within this cluster suggests a concerted effort toward advancing the capabilities of LLMs in handling and interpreting complex biomedical data. Cluster 3 emphasizes the potential practical medical applications of LLMs and encompasses various medical specialties and fields, such as internal medicine and medical education. This cluster signifies the prospective role of LLMs in clinical practice, medical training, and patient care. Cluster 4 shows concepts at the crossroads of psychological science and its applications within the biomedical and health sectors. This cluster signifies an emerging trend where LLMs have been used to obtain insights into patient psychology, public health, and the societal impact of health interventions.

Overall, this keyword network provides an overview of the state-of-the-art LLM application in BHI. It shows the main topics being studied and the interdisciplinary collaborations that are crucial for making progress in this field. The following sections will examine each of these topics in-depth, explaining their contributions and highlighting the interconnected research efforts that could drive the continued advancement of BHI.

4.1.1 LLM Research Themes

The categorization of tasks associated with LLMs in the context of BHI into methodology and outcome is a strategic way of organizing the research papers' focus areas[9], which delineates between technical development and practical applications/evaluation. Figure 3 shows the number of papers within each research theme, with red bars indicating the outcome theme and blue bars indicating methodologies.

In terms of methodology (blue), LLM topics such as information extraction, inference, summarization, sentiment analysis, and named entity recognition show the nuanced capabilities of LLMs in processing and analyzing textual data, which could support various aspects of clinical and research activities in the biomedical sector. The topic of multilinguality and the topic of text generation are also well-represented, illustrating the technical versatility of LLMs and their potential for creating understandable medical content in multiple languages, which is vital for diverse patient communication and international research collaboration. From a technical standpoint, the topic of image, vision, video, and multimodality acknowledges the integration of LLMs with other data forms, which is an important step towards comprehensive analytics in diagnostics and patient care.

For outcome (red), the highest number of papers centered on the model evaluation category, which suggests that there is a significant emphasis on validating and testing the

[9] In Appendix 3, we present the representative papers for each LLM task.

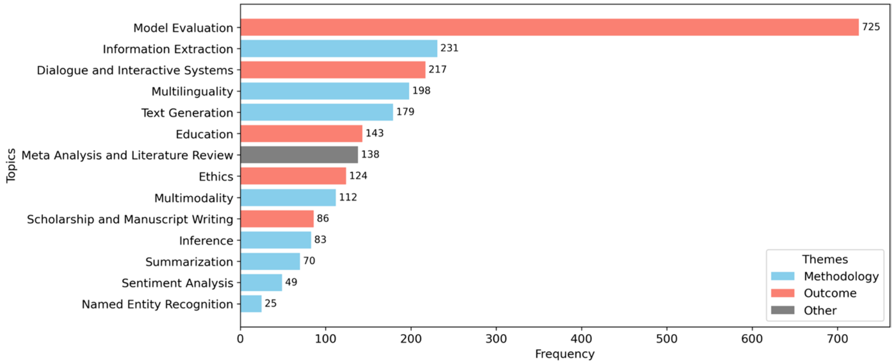

Fig. 3 LLM tasks by research theme and topic (We include *Meta-analysis and literature review* to classify the papers. However, since it is not within the scope of LLM methodology or outcome, detailed analysis of papers of this category is presented in Appendix 4)

effectiveness and reliability of LLMs within the biomedical field. Model Evaluation is critical because the outputs of such models often inform decision-making in health-related matters where accuracy is paramount. Other LLM tasks in the outcome category include Sects. 6.1, 6.3, and 6.2, representing the substantial interest in using LLMs to distill medical information from various data sources to enhance patient interaction, medical education, and research. The topic of ethics also has a dedicated focus, which is crucial given the sensitive nature of medical data and the implications of AI in healthcare decisions.

4.1.2 Major Diagnostic Categories

Table 2 categorizes the research papers according to the health issues they address, showcasing the wide-ranging capabilities and applications of LLMs in BHI. Research has predominantly focused on mental health conditions, including depression and ADHD. Similarly, diseases of the nervous system also attract considerable attention, with studies covering disorders from Parkinson's to Alzheimer's disease. The application of LLMs in tracking and managing infectious and parasitic diseases, such as complications from infections and COVID-19, underscores their importance in infectious disease surveillance, particularly in light of recent global health emergencies. Furthermore, research on the circulatory system targets widespread conditions such as heart disease, which continues to be a leading cause of death globally. Other less-represented diseases, such as those affecting the musculoskeletal and endocrine systems, metabolic and digestive disorders, and urinary tract issues, demonstrate LLMs' versatility in tackling a broad spectrum of chronic and acute health challenges.

4.2 Scholarly networks and partnerships

The visualization of the citation network shown in Fig. 4 offers a detailed perspective on the emergent field of LLMs in healthcare. The network includes

Table 2 Major diagnostic category count with examples

Category	No. of collected papers	Examples
Mental diseases and disorders	89	Depression, post-traumatic stress disorder (PTSD), attention-deficit/hyperactivity disorder (ADHD)
Nervous system	87	Epilepsy, vestibular schwannoma, carpal tunnel syndrome (CTS), Parkinson's disease, Alzheimer's disease
Infectious and parasitic diseases and disorders	60	Post-infectious complications, signs and symptoms of adverse events following immunization (AEFIS), symptomatic COVID-19 infections
Skin, subcutaneous tissue, and breast	41	Breast cancer, melanoma, skin disease
Circulatory system	40	Congenital heart disease, atrial fibrillation, heart failure
Musculoskeletal system and connective tissue	31	Shoulder impingement syndrome, anterior cruciate ligament (ACL) injury, rheumatology-related diseases, osteoarthritis (OA), gout
Endocrine, nutritional, and metabolic system	28	Anorexia, thyroid cancer, type 2 diabetes mellitus, diabetes
Digestive system	26	Colorectal cancer, inflammatory bowel disease (IBD), inflammatory bowel disease, digestive diseases
Eye	24	Primary acquired nasolacrimal duct obstruction, myopia, cataract
Respiratory system	19	Lung cancer, asthma, metastases, non-resolving pneumonia
Hepatobiliary system and pancreas	17	Cirrhosis, hepatocellular carcinoma, liver cirrhosis, liver disease
Kidney and urinary tract	16	Urolithiasis, end-stage renal disease, transplant chronic dysfunction, graft loss, urinary tract infection (UTI)
Blood and blood-forming organs and immunological disorders	16	Sickle cell anemia, chronic myeloid leukemia, non-Hodgkin's lymphoma, acute bleeding, anemia severity
Ear, nose, mouth, and throat	13	Oral potentially malignant disorders (OPMDS), necrotizing otitis externa, neoplastic rhinopharyngeal lesion
Female reproductive system	12	Infertility, ovarian cancer
Alcohol/drug use or induced mental disorders	11	Substance use disorders, drug abuse, addiction, smoking cessation, addiction
Male reproductive system	10	Prostate cancer, erectile dysfunction
Factors Iinfluencing health status	9	Drug–drug interaction (DDI)
Pregnancy, childbirth, and puerperium	7	Postpartum hemorrhage (PPH)
Injuries, poison, and toxic effects of drugs	5	Acute organophosphate poisoning
Multiple significant trauma	5	Joint contractures, internal organ dysfunction

Table 2 (continued)

Category	No. of collected papers	Examples
Newborn and other neonates (perinatal period)	4	Neonatal diseases
Myeloproliferative diseases and disorders (poorly differentiated neoplasms)	4	Chronic myeloproliferative neoplasms
Burns	4	1st-degree burns
Human immunodeficiency virus (HIV) infection	3	HIV

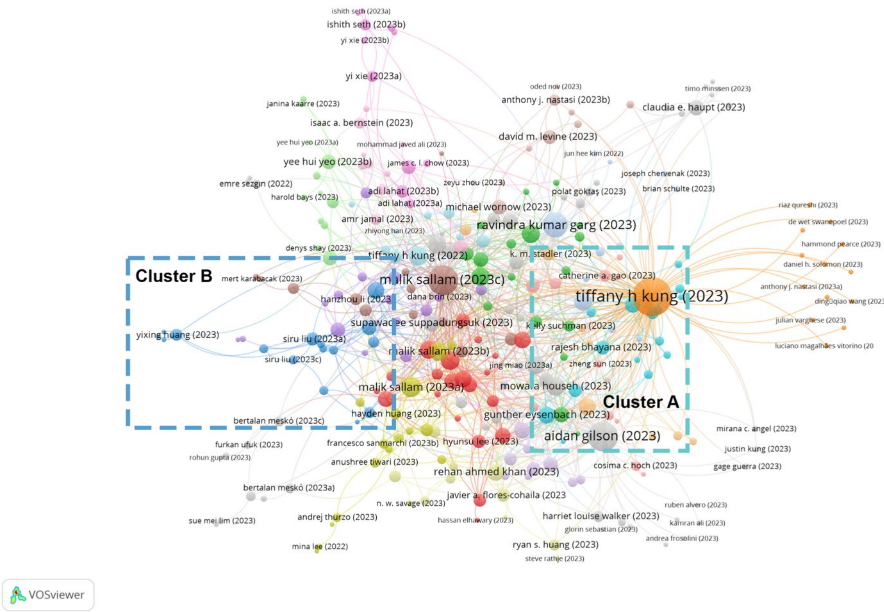

Fig. 4 Paper co-citation network

312 papers, each with at least five citations, which ensures that the visualization emphasizes the more influential and recognized studies within the field. The structure of the network indicates a close connection among studies, with certain seminal papers emerging as central nodes with their high citations. Sallam [50], Kung et al. [52], and Gilson et al. [67] are particularly prominent, suggesting that their works on model performance evaluation and systematic literature reviews have been widely recognized across the field. Additionally, the network shows the emergence of subfields or specialized areas of research, as illustrated by distinct clusters. For instance, cluster A (cyan) highlights the focus on radiology reports [68–71], whereas cluster B (blue) is dedicated to educational applications within medical specialties, such as dentistry [72–74].

The dynamic and collaborative nature of the citation network indicates the ongoing development within this field of study. New theories and methodologies are continuously being integrated. This dynamic is typical for an emerging field where the foundational work is still being established and where there is significant potential for discoveries and applications.

4.2.1 Organization Collaboration Network

The network map in Fig. 5 provides a visual representation of the co-authorship links (with more than 5 co-occurrences) that exist among research organizations across the globe. We observe that the nodes are predominantly universities and research institutions. However, the presence of hospitals and healthcare organizations within this network cannot be overlooked; it signals an integrated research

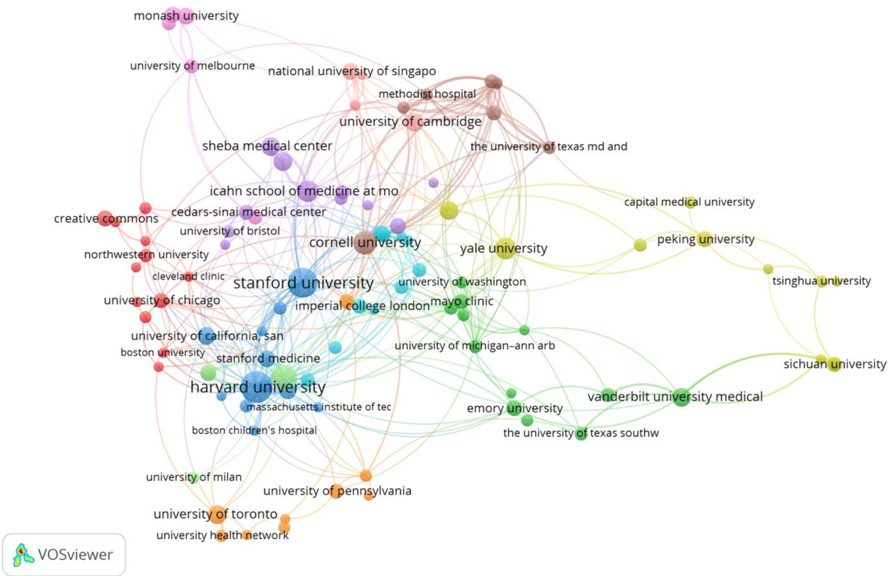

Fig. 5 Organization collaboration network

approach where clinical settings play a crucial role in the translation of academic findings into healthcare advancements. The inclusion of these healthcare entities not only diversifies the nature of the collaborations but also enhances the potential for practical, patient-centered outcomes to emerge from these scholarly partnerships.

Certain institutions appear as pivotal nodes within this network. These nodes, often representing universities and research centers like Harvard University, Stanford University, and the University of Oxford, are heavily interconnected with a multitude of other nodes. This finding suggests a high degree of collaborative engagement, which is often a reflection of the institutions' broad research portfolio and its pivotal role in facilitating multidisciplinary studies.

The network also includes tightly interconnected research clusters indicated by colors, suggesting the existence of consortia or research groups that may be working in concert towards a common scientific objective. The network includes edges connecting institutions from multiple continents and countries, which signifies the extent of international collaboration efforts.

4.2.2 Collaboration Network Among Countries

Figure 6 provides a visual representation of a collaboration network among countries and regions, with an overlay that indicates the average publication year of papers from each country and region. This visualization not only shows the collaborations that exist among countries but also provides a temporal dimension of how the research landscape has evolved. There are three main findings regarding the early pioneers, the major collaborators, and the dynamic and evolving network.

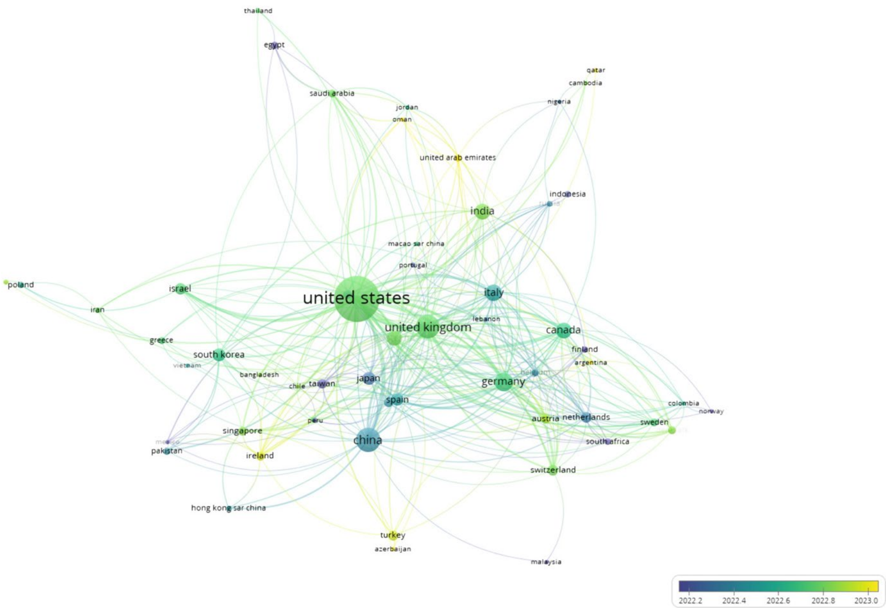

Fig. 6 Collaboration network among countries

Early Pioneers It is shown that countries such as Japan and the Netherlands have begun research on LLMs earlier, making them pioneers in this field. Their early start suggests that these countries have established a strong foundation in LLM research, contributing significantly to the early development and understanding of these technologies in the field of BHI.

Major Collaborators As shown in Fig. 6, the USA and the UK are depicted with a large total link strength (as indicated by the size of the nodes), which is indicative of their strong influence and the density of their collaborative networks. A large link strength suggests these countries are central nodes in the network, engaging in numerous collaborative research projects and often being the driving force behind pushing the frontiers of LLMs. Their central role in the network underscores their importance in both producing and disseminating LLM knowledge.

Dynamic and Evolving Network The network is dynamic and evolving, with countries like Ireland, Turkey, and the United Arab Emirates emerging as participants. It indicates that the field of LLMs is growing, attracting a diverse set of contributors, and expanding the geographic diversity of research. The participation of these countries may bring new perspectives and innovations to the field, and their increasing involvement highlights the global interest in and importance of LLM research.

5 Navigating the Spectrum: the Distributed LLM Methodologies in BHI

The study of LLMs in the area of BHI covers a wide range of methods and use cases, showing a major change in how AI is used in the biomedical and health fields. This detailed review starts with how LLMs change information extraction, making it easier to handle and understand different types of data, like clinical notes and radiology reports. The discussion then moves on to multilinguality, looking at how well LLMs perform in different languages and the challenges and solutions to creating content in multiple languages. The next parts focus on text generation, highlighting how LLMs play a role in both medical writing and communication with patients. The study also looks at how LLMs can handle multiple types of data, like images and genomic data, which helps improve diagnosis and prediction. As the discussion continues, it emphasizes how LLMs are important for drawing conclusions and analyzing sentiment, showing their significant impact on understanding complex medical data and human feelings. The study ends with a look at how LLMs are used in named entity recognition, pointing out current progress and potential for future improvements. Overall, each part highlights the diverse and specific applications of LLM methods in changing BHI research and practice.

5.1 Information Extraction (Including Sentiment Analysis and Named Entity Recognition)

The utilization of LLMs has rapidly reshaped the BHI research landscape, notably in the domain of information extraction. Recent literature underscores their transformative impact across multiple applications, which we will discuss in three main areas: structured information extraction, sentiment analysis, and NER. Sentiment analysis and NER are sub-tasks of information extraction that play crucial roles in understanding and organizing unstructured data.

For structured information extraction, some studies demonstrate LLMs' proficiency in enhancing diagnostic accuracy in hematology [75], extracting structured information (e.g., diseases, symptoms, and signs) from vast textual data (e.g., clinical notes, EMR notes, and radiological reports) in various languages [76–80], and identifying narrative entities in the news domain [81–83]. In addition, a part of the research illustrated the ability of LLMs to assist in the extraction of evidence-based explanations and enable the accurate retrieval of information from clinical documentation, providing support for medical practitioners' decision-making [84–88]. Our review also indicated that LLMs are instrumental in extracting medication mentions [89], classifying events and contexts in clinical notes [90], and improving the understanding of medication adherence through the detection of drug discontinuation events from social media data [91]. They also excel at generating structured outputs on medications and temporal relations, further aiding in disease prediction and clinical decision support [92–94]. These advancements, coupled with self-verification techniques [77] and the extraction of demographics and social determinants of health from EHRs [95–99] illustrate LLMs' capacity to integrate and analyze healthcare data effectively.

In BHI domains, LLMs contribute to the refinement of sentiment analysis tools [100–102]. For instance, a model utilizing weights from a publicly available zero-shot classifier, which is built from the BART LLM and fine-tuned on the MNLI dataset, has been employed to evaluate linguistic nuances during psychological therapy sessions [103]. Similarly, other research finds that LLMs could be used to analyze patient feedback, clinical notes, and public health discussions, thereby gauging public sentiment on health-related matters [104], understanding patient experiences [105], monitoring mental health trends [106, 107], and identifying cognitive distortions or suicidal tendencies [108]. Additionally, LLMs in sentiment analysis facilitate medical education [102, 109–111] by fostering interactions between medical trainees and educators, detecting thematic differences and potential biases, and revealing how feedback language may reflect varying attitudes toward learning and improvement [112]. LLMs could also contribute to the sentiment analysis of research articles and medical journals, offering insights into the research community's responses to novel findings or treatments [113, 114].

Moreover, LLMs have been applied to improve the efficiency and performance of NER in BHI domains. For instance, LLMs have helped identify ancient Chinese medical prescriptions from the Song Dynasty [115, 116]. While there is not too much representative literature compared to other methodology subdomains, [117] identifies the need to further develop supervised medical NER models, especially when human-annotated data are unavailable.

5.2 Multilinguality

Our review highlights the emerging research applications of multilingual LLMs in BHI. Some research has explored how to use multilingual LLM to generate multilingual content in BHI. The content generation tasks include using multilingual LLMs for dataset generation [115, 118, 119] and question generation [115, 118, 119][10]. Multilingual LLMs are also leveraged to identify personal health information in Chinese-English code-mixed clinical text and ancient Chinese medical prescriptions [120, 121]. These studies demonstrate the versatility and potential of multilingual LLMs in processing low-resource multilingual and cross-cultural biomedical and health information.

Other research papers concentrate on evaluating LLM performance across various languages, including English, Korean, Spanish, Turkish, and Chinese [122–128]. Studies explore multilingual question answering using the Japanese National Examination for Pharmacists (JNEP) [129], the Korean dermatology specialty certificate examination, and the Persian medical residency examination [125]. LLMs, including ChatGPT, were also tested for their ability to generate multilingual health-related questions [115] and their ability to facilitate multilingual communication [130]. By comparing the results obtained from different language settings, these studies focus on the correctness, consistency, and verifiability of LLMs' responses.

[10] The generation tasks here exclude text generation, which is discussed in Sect. 5.3.

5.3 Text Generation

Research LLM-based text generation in BHI concentrates on two main purposes: medical scientific writing and clinical patient-facing writing.

In medical scientific writing, current research on text generation predominantly focuses on two areas. The first area focuses on the potential utility of LLMs, particularly GPT-4, as tools for authoring various scientific publications. The general consensus is that human-written texts are more concrete, diverse, and typically contain more useful information [131–133]. In contrast, medical texts generated by GPT-4 prioritize fluency and logic, often using general terminologies instead of context-specific information [131, 134]. AI-generated texts may include inaccurate information, fabricated references, and lack the inclusion of recent literature [135–137].

The second area is the effectiveness of distinguishing LLM-generated texts through human evaluation or AI-driven output detection mechanisms. Some studies focus on detecting AI-written text in specific sections of BHI papers, such as the abstract and background [138, 139]. While LLM-based methods are generally useful in distinguishing AI-written abstracts from original ones, they struggle in the field of radiology where both human reviewers and output detectors fail to differentiate GPT-generated abstracts from original ones [139]. It has also been claimed that distinguishing AI-written backgrounds from human-written ones is challenging [139]. More robust output detectors have been developed to distinguish AI-generated text from human-generated text [140, 141]. Overall, researchers advocate for chatbots to serve as assistants rather than authors in scholarly work, emphasizing the importance of transparency if chatbots are involved in generating academic content [142].

For clinical patient-facing writing, efforts have been made to evaluate the feasibility of using GPT-4 for generating case reports and responses to various patient inquiries about surgical procedures and health-related matters. These include responding to postoperative questions [143], generating health messages [144], aesthetic surgery advice [145], pro-vaccination message generation [146], and communication in palliative care [8]. Most studies show positive results regarding GPT-4's ability to generate coherent, easily comprehensible answers. One study even suggests that AI-generated messages are comparable to human-generated ones in terms of sentiment, reading ease, semantic content, and suggestions [144]. However, its accuracy, completeness, and extent of personalization still need improvement [145]. Therefore, AI models cannot replace a human agent at present [8].

5.4 Multimodality

Multimodality in large language models within BHI refers to the ability of these models to understand and process multiple types of data beyond text, which includes imaging, audio, and genomic data. In our scoping review, papers on multimodal LLMs have been applied to various aspects of BHI, including healthcare in general [147, 148], medical image analysis [149–152], radiology [153, 154], pharmaceutical sciences [155], dentistry [73], and public health informatics [156]. Methods used in these papers can be crudely classified into pretrain-from-scratch [157–163] as well as finetuning based on the pre-trained or instruction-tuned models [164–170] such as Vicuna, SAM, BLIP, Llama, OpenLlama, etc.

As healthcare and medicine are highly specialized fields, many multimodal models are uniquely adapted to enhance tasks in vision, audio, and genomic analysis. In vision applications, models are designed for tasks including image-to-text medical report generation [164, 167, 171–173], medical image captioning [159, 161, 174], medical video retrieval [175], and video anomaly detection [169]. In [173], LLMs integrate Vision Transformers (ViT) and Faster R-CNN with GPT-2 to analyze brain images for dementia, enhancing diagnostic accuracy by capturing intricate visual features and generating detailed textual reports. Specified models are also developed in audio and genomic applications: LLMs such as the Diagnosis of Thought (DoT) model [176] assist in psychotherapy by detecting cognitive distortions from patient speech and aiding therapists in understanding and addressing mental health issues more effectively. In the field of genomics, protein language models predict the impact of genetic variations on protein structure and function, identifying potential compensatory mutations in pathogenic variants [177].

5.5 Inferences

In addition to LLMs' application in correlational or empirical studies in BHI, they have also been instrumental in inferences, with a focus on analyzing associations and causal relationships. For example, LLMs facilitated a Socratic dialogue with ChatGPT to analyze the causal effects of PM2.5 on human mortality risks. After extensive fine-tuning and addressing confounding factors, a causal link was established [100]. Moreover, LLMs have been adapted to develop a natural language inference system specifically for clinical trial reports. This system focuses on extracting and interpreting medical evidence to enhance the accuracy and reliability of these reports [101]. In a different application, the GPT model has been utilized for medical image analysis. Demonstrating its capabilities as a plug-and-play transductive inference tool, GPT has proven effective in detecting prediction errors and improving accuracy in medical images, highlighting its potential for broader applications in this field [102].

6 Expanding the Horizon: the Diverse Outcomes of LLMs in BHI Applications

The integration of LLMs has also expanded the horizons of BHI, leading to a diverse array of outcomes and applications. Beyond enhancing NLP capabilities, LLMs have facilitated a more personalized and nuanced approach to patient engagement, enabling healthcare providers to tailor their communication and interventions based on individual patient profiles through dialogue and interactive systems. In addition, LLMs have revolutionized scholarship and manuscript writing, which are also applicable to BHI fields. Furthermore, the evaluation and ethics assessment of LLMs have become essential research topics in BHI, given the high standards of precision and stability in healthcare and medical systems. This section explores the multifaceted impact of LLMs across various BHI applications, highlighting their potential to revolutionize patient care and medical research.

6.1 Dialog and Interactive Systems

The LLMs have been implemented in the newly developed chat box as an AI assistant for healthcare conversion, including personalized health diagnosis and intervention in BHI. Typically, the chat assistant, based on either naïve conversational AI or generative AI systems, was designed to help in the analysis of the message from dialogs [178–181], the estimation and evaluation of the health status [178], and the generation of high-quality responses [178, 179] after considering the possible knowledge, including the patient's EHRs and medical knowledge in the clinical setting. For example, an LLM-derived chatbot called CareCall [178] was designed to support people and alleviate feelings of loneliness. It leads to frequent open-ended conversations, generates replies by using a pre-trained LLM model, captures health metrics and emergency alerts, and displays the reports for social works. Another newly developed application powered by the ChatGPT-3.5 model [179] allows callers to receive up-to-date personalized medical suggestions based on the conversation. In addition, a prospective use of ChatGPT within healthcare, especially during the pandemic period, was proposed, which helps with answering the patients' health-related questions [182]. The high-quality performance of using the AI assistant confirms that the models can understand and reply to people's needs. However, privacy, ethics, and information accuracy are the major concerns while the LLM/AIs are involved in generating professional responses regarding disease diagnoses and drug suggestions [182, 183]. More rigorous tests are needed to guarantee the safety of using the LLM in clinics [183].

6.2 Scholarship and Manuscript Writing

As more researchers in the BHI domain use ChatGPT and other AI technologies in writing manuscripts, the discussions around the use of LLMs in scientific writing have been emphasized, accompanied by a rise in various concerns. Although LLMs can improve writing quality, summarize relevant articles, and facilitate manuscript translation [184], they face challenges in accurate referencing [185], unintentional plagiarism, and data biases [186]. Establishing regulations and guidelines for the use of LLMs in scientific writing is crucial for assessing both effectiveness and ethical considerations [48, 187].

6.3 Education

Researchers have assessed LLMs' abilities to enhance medical education, discussing their potential to improve the current education and decision-making process. LLMs exhibit similar performance in comparison to human achievement without specialized training on both the USMLE [52] and more specialized domains such as neurology board-style examinations [188]. LLMs can also enhance student engagement and learning experiences [49], especially personalized curriculum development and study plans [189], albeit with considerations of ethical challenges [49, 50], algorithmic bias, and plagiarism [50, 189]. Additional efforts are required from educators, students, and model developers to establish clear guidelines and rules for their applications ethically and safely in academic activities [189]. These perspectives on using LLMs highlight both the potential benefits and ethical considerations surrounding the integration of LLMs in medical education and practice.

6.4 Model Evaluation

As demonstrated in the previous sections, LLMs are widely utilized in a range of applications within BHI. To assess their effectiveness, new frameworks, benchmarks, and metrics for evaluating the performance of these models have been developed. Frameworks such as the Translational Evaluation of Healthcare AI (TEHAI) have been proposed by research teams to evaluate the capability, utility, and adoption of such systems in healthcare [190]. Papers also set benchmarks by assessing the performance of LLMs on various tasks [111, 191, 192], using relevant datasets such as MIMIC for general medical information and OpenI for radiographs [193]. In their evaluation, metrics such as ROUGE-L have been frequently used [194]. In some cases, additional human evaluations are introduced, which rely on the qualitative coding of LLM outputs. For example, for LLMs applied to summarize medical evidence [195], human efforts to evaluate the model-generated summaries involve the open coding of qualitative descriptions of error types for medical evidence summarization, drawn from qualitative methods in grounded theory. As another example, human evaluation involves recruiting human subjects to interact with chatbots and solicit their responses [196, 197].

In Appendix 5, we present a thorough analysis of the specialized and contextualized model evaluation in specific disease categories. Taking mental health disease as an example, we highlight evaluation techniques in mental health applications against various metrics and datasets.

6.5 Ethics

Ethical discussions on LLMs caution against the application of LLMs in high-stakes contexts and center around issues of misinformation, bias, inequalities, privacy, and transparency [198, 199]. The use of LLMs as a clinical decision support tool as well as a service-providing tool through chatbots can potentially harm patients when they make false recommendations, diagnoses, or prescriptions [198, 200]. Such harms, while unintended, are rooted in the corpus of training data embedded in unequal social processes [201]. Moreover, those negative consequences can also be compounded when human health professionals' judgments and decision-making processes are influenced by such biased diagnoses [198]. In particular, the use of AI-generated texts or conversational chatbots in medical contexts often involves patient-specific medical information [198, 202, 203]. This might introduce additional privacy harm to patients since these technologies often require access to patients' sensitive information and medical record data [204]. For the responsible use of such technologies, clinicians will need to critically review and validate generated texts or outputs before deploying them in practical settings. Moreover, the lack of consent sharing poses another concern around data privacy and security in healthcare [205].

7 Applying LLMs in Specific Disease Categories: Popular Fields and Open Opportunities

This section provides a detailed exploration of the transformative impact of LLMs on various disease categories, focusing particularly on mental health, nervous system disorders, and other open opportunities. Mental health and nervous system disorders are the top two widely represented topics in the collected corpus, as indicated by the counts in Table 1. We focus on these two areas as examples to analyze the trending LLM-based BHI applications while uncovering additional domains as open opportunities. By understanding how LLMs can be effectively applied in these well-represented domains, we can extend these insights to other disease categories, thereby broadening the scope and impact of LLM technology in healthcare.

7.1 Mental Health

As shown in Fig. 7, LLMs are poised to revolutionize mental healthcare by enhancing diagnostic processes, intervention strategies, and overall mental health and well-being promotion. The potential for LLMs in these domains is vast, ranging from facilitating early detection of mental health issues to providing scalable interventions.

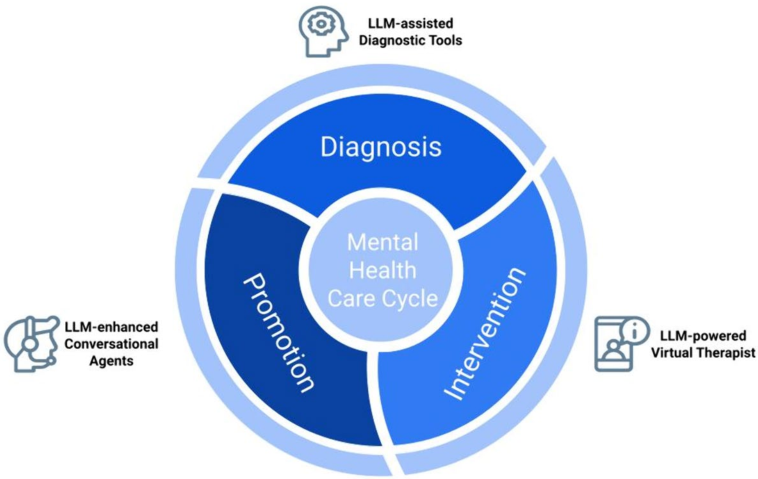

Fig. 7 Integration of LLM technology in the mental healthcare cycle

7.1.1 Diagnosis

The application of AI in mental health diagnostics has been rapidly advanced with tools like GPTFX [206], which exhibits a remarkable ability to classify mental health disorders and generate relevant explanations. This approach not only enhances the performance of mental health disorder detection but also provides valuable interpretability for the predictions, which is a crucial aspect of clinical applications. The study *Advancing mental health diagnostics: GPT-based method for depression detection* [7] leverages transformer networks like BERT, GPT-3.5, and GPT-4 to analyze clinical interviews. They have shown strong abilities to understand complex linguistic patterns and contextual cues.

These pioneering studies indicate that LLMs could be instrumental in mental healthcare by providing nuanced, scalable, and efficient tools for diagnosis. By analyzing language with unprecedented depth and breadth, LLMs could uncover mental health patterns that may be imperceptible to humans, assist in early detection, and offer continuous support for individuals struggling with mental health issues.

7.1.2 Intervention

The field of mental health intervention has benefited through the integration of LLMs and digital health technologies. In [109], researchers proposed a mobile app that utilizes GPT technology for tracking psychological mood changes and providing e-therapy. By offering a platform for users to record and analyze their psychological fluctuations, it aids in identifying triggers for negative mood changes, effectively functioning like a virtual therapist. The app's evaluation underscores its efficacy in journaling and basic AI-driven mental health guidance, exemplifying the potential of LLMs in personal mental health management.

Community-based mental health support can also leverage the advanced capabilities of AI and LLMs, providing more healthcare resources. The paper titled *Enhancing psychological counseling with a LLM: a multifaceted decision-support system for non-professionals* [207] highlights the need for psychological interventions in the social media sphere, where expressions of negative emotions, including suicidal intentions, are alarmingly prevalent. The model leverages the advanced capabilities of AI and LLMs to empower non-professionals or volunteers to provide psychological support. By analyzing online user discourses, the system assists non-professionals in understanding and responding to mental health issues with a degree of accuracy and strategy akin to that of professional counselors.

These pioneering applications of LLMs in mental health interventions demonstrate their immense potential in both personal and community settings. Supporting nuanced, user-friendly, and scalable solutions, LLMs have reshaped the landscape of mental health care. They offer innovative tools for real-time emotional tracking, mood analysis, and intervention, facilitating broader access to mental health support and enabling effective responses to complex emotional expressions.

7.1.3 Promotion

Healthcare promotion, particularly in the realm of mental health and well-being, has undergone a significant transformation with the advent of AI-based conversational agents (CAs) [208, 209]. The integration of these advanced technologies has not only reshaped therapeutic approaches but also expanded access to mental health resources. This shift is well-articulated in the comprehensive paper titled *Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being* [210]. The study underscores that the quality of human–AI therapeutic relationships, content engagement, and effective communication significantly shape the user experience. It implies that while AI-based CAs could be highly effective, their impact is greatly influenced by the quality of interaction and the relevance of the content they provide.

Additionally, LLMs play a crucial role in healthcare promotion, not only by raising overall awareness but also by offering patient-centric recommendations [211]. They effectively address and dispel common misconceptions and myths about mental health, significantly contributing to the reduction of stigma associated with mental health issues. By educating the public in a non-judgmental and informative manner, LLMs help cultivate a more understanding and supportive community. Furthermore, these models are adept at disseminating a wealth of health-related information in formats that are easily comprehensible. They offer insights on a wide range of topics, from general wellness and stress management to the critical importance of mental health. This comprehensive approach aids in heightening awareness and educating people about the importance of maintaining good mental health, as well as recognizing the early signs of potential issues.

7.2 Nervous System

In the realm of neurological disorders, leveraging LLMs for disease prediction signifies a groundbreaking shift toward harnessing the intricacies of human language and clinical data. Two pivotal studies exemplify this innovative approach, particularly focusing on multimodal data to predict diseases of the nervous system.

The study, *Predicting dementia from spontaneous speech using large language models*, [212] delves into the predictive potential of LLMs by analyzing physicians' clinical notes for signs indicative of seizure recurrence in children following an initial unprovoked seizure. Their work demonstrates that the nuanced understanding captured from electronic medical records could significantly augment the predictive accuracy of seizure recurrence. Another paper, *Multimodal approaches for Alzheimer's detection using patients' speech and transcript* [213], ventures into the domain of Alzheimer's disease detection by employing a multimodal strategy that integrates patients' speech and transcript data. This study underscores the immense potential of multimodal data in enhancing Alzheimer's detection and sheds light on the complexities and opportunities inherent in leveraging speech data for the prediction of neurological diseases, paving the way for more effective and nuanced diagnostic tools.

Both of the above studies underscore the significant advancements made in the domain of neuroscience, particularly through the use of LLMs and multimodal data analysis. By capturing and integrating diverse data types, from clinical notes to speech and transcripts, researchers could unveil previously obscure patterns and indicators of disease, offering promising new avenues for early detection and treatment strategies for conditions affecting the nervous system.

7.3 Open Opportunities

The application of LLMs in BHI holds promising potential to revolutionize disease diagnosis, prediction, and intervention, other than the mental health and neurological disorders that have been extensively researched. Though their use in clinical fields is still in the beginning stages, there are several opportunities for LLMs to significantly enhance patient care and disease prognosis, particularly in areas such as hospital management, adverse drug reactions, infectious diseases, and health promotion.

In clinical settings, LLMs could be instrumental in identifying correlations or even casual relationships [214] by referencing vast datasets such as clinical notes, emergency care reports, and poison control center data. It could lead to the development of more effective triage systems in emergency departments [215] and quicker, more accurate diagnoses [216]. Ultimately, it would help reduce the time needed to administer antidotes or interventions that alleviate symptoms and monitor drug/treatment reactions. Additionally, through the in-depth analysis of the language and semantic information embedded in these full EHRs, LLMs could predict potential personalized treatments [217] to mitigate adverse drug reactions [218].

In the management of infectious diseases with or without pandemic potential, such as sexually transmitted disease (STD), influenza, and COVID-19, LLMs could play a pivotal role in improving patient engagement, promoting adherence to antiretroviral/antibacterial therapy, and monitoring disease progression [219]. By analyzing patient interactions, social media, and support group communications, LLMs could identify language indicative of treatment fatigue or social determinants affecting adherence [98]. Furthermore, through the analysis of clinical narratives over time, LLMs could detect subtle changes in patient status, predict potential comorbidities, and personalize patient education and intervention programs [220]. It could lead to improved health outcomes and quality of life for individuals affected by diseases that currently have no cure.

Finally, LLMs could also extend their contributions beyond disease settings. For example, LLMs can also support the training of medical professionals through simulations and interactive learning platforms, providing personalized learning experiences and improving the quality of medical education [189]. LLM can also benefit public health promotion by enabling more precise and targeted health communication strategies [221].

The potential of LLMs in these medical domains is vast, offering opportunities for enhanced diagnostic accuracy, personalized treatment, and patient care. As the technology and methodologies behind LLMs continue to advance, their integration into clinical workflows and research initiatives will likely become increasingly prevalent, driving forward the capabilities of modern medicine. As the capabilities and applications of LLMs in healthcare expand, there will be a growing need for research into their ethical, legal, and social implications to ensure they are used responsibly and equitably.

8 Conclusions and Discussion

Our review has shown important trends and developments in using LLMs for BHI. Applying LLMs has changed the methods and potential outcomes in the healthcare field. Particularly from January 2022 to December 2023, there has been a big increase in the number of research articles, showing rapid progress in this field. This applied research includes better diagnostic tools, improved patient engagement, more efficient management of EHRs, and the emerging field of personalized medicine.

The use of LLMs in BHI has captured advanced natural language processing capabilities, potentially improving medical diagnosis, patient care, and research methods. Our network analysis shows that LLMs have also fostered collaborative networks across different disciplines, including academia, healthcare, and technology industries. This multidisciplinary approach is vital for the responsible growth and ethical application of LLMs. Our review also highlights an increasing focus on addressing practical challenges and ethical implications, such as data privacy and AI bias, underlining the need for robust policy frameworks. The potential impact of LLMs in BHI is significant, but it requires a balanced approach considering both the technological capabilities and the ethical, legal, and social implications.

In summary, our review provides a comprehensive resource for stakeholders in the healthcare sector. It offers an overview of the current state of LLMs in BHI and insights into future directions. As LLMs continue to evolve and integrate further into healthcare, understanding their development could be crucial for researchers, clinicians, policymakers, industry leaders, and all stakeholders. It is also important to remain committed to the ethical and responsible use of LLMs in advancing healthcare.

8.1 Limitations

This review is subject to certain limitations. First, our classification methodology, while able to conduct multi-label classification, primarily focuses on identifying the most relevant topics within each article. This approach is effective in streamlining the analysis but may overlook the multi-faceted nature of some research papers, where secondary topics could also hold significant relevance.

Second, the scope of our review is centered on LLMs, potentially excluding foundation models operated in other modalities such as vision and voice. Additionally, the specific use of biomedical and health-related keywords in our search criteria may have inadvertently excluded relevant studies that do not explicitly use these terms but are pertinent to the field.

Another potential limitation stems from the data-collection process. At the time of our data collection, OpenAlex did not facilitate a refined search based on keyword matches within titles or abstracts. Therefore, we applied several predefined rules, such as filtering articles based on key search terms in the abstracts. We also note that a significant portion of the collected papers are preprints, which have not undergone the peer-review process and whose findings and assertions are not established. Although studies, such as [222], have found that over 75% of preprints are eventually published in peer-reviewed journals, we recognize the need for additional validations to ensure the reliability and accuracy of the information presented in these preprints.

These limitations present several opportunities for future work to refine the review. One future work could investigate the application of foundation models in other modalities in BHI fields, including vision and voice. Another future work could continue to collect articles and track trends in this area.

8.2 Future Work

Looking ahead, LLMs have recognizable potential to transform healthcare delivery and patient outcomes. As LLM capabilities continue to evolve, our future work will focus on exploring more advanced ways to integrate LLMs into BHI. This will involve addressing emerging ethical and operational challenges, such as ensuring responsible and fair use of LLMs in healthcare, which is crucial for fully realizing their potential.

The field is evolving rapidly, so ongoing monitoring and analysis will be necessary. We anticipate a surge in publications and citations related to LLMs in the near future. Therefore, continuously updating our review will be essential to maintaining its relevance and impact. Our future work will also explore foundation models beyond LLMs, acknowledging the growing importance of multimodal systems in healthcare. By expanding our research focus, we aim to provide a more comprehensive understanding of the role of advanced computational models in BHI, thereby contributing to the development of more effective and ethical healthcare solutions.

Appendix 1 Paper Organization Workflow

Here, we provide a visual representation of the workflow used for organizing this research paper Fig. 8.

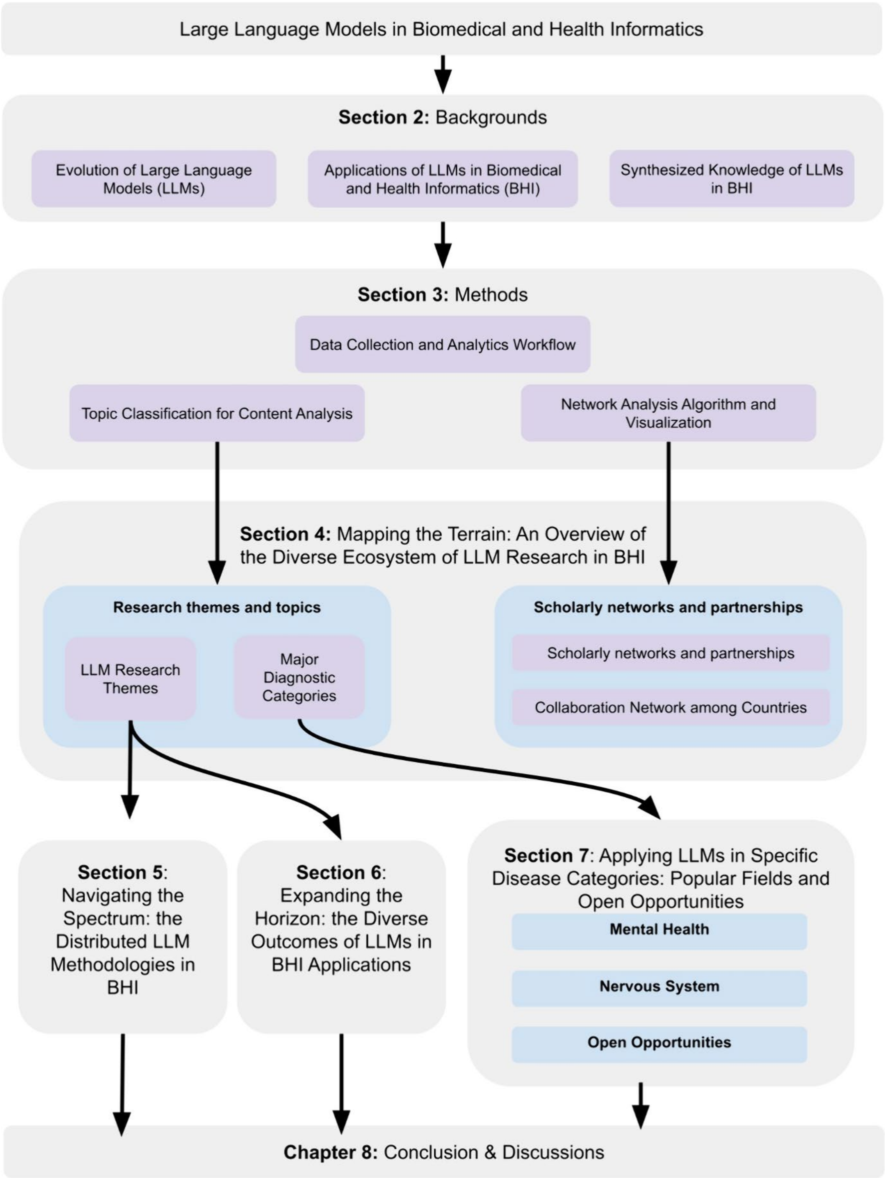

Fig. 8 Paper overview

Appendix 2 Top 50 Concepts in Keyword Co-occurrence Network Ranked by Total Link Strength

Table 3 shows the top 50 keywords represented in the keyword co-occurrence network by total link strength. The total link strength refers to the sum of the link strengths of one keyword over all the other keywords. The greater the frequency of

Table 3 Top 50 concepts in keyword co-occurrence network

Keyword	Occurrences	Total link strength
Computer science	1394	16,099
Artificial intelligence	932	11,246
Medicine	1033	10,710
Psychology	719	8286
Political science	483	5850
Law	461	5602
Natural language processing	371	4845
Mathematics	344	4749
Sata science	363	4606
Biology	344	4585
Programming language	334	4471
Economics	303	4262
Healthcare	324	4239
Philosophy	294	3869
Medical education	349	3813
Engineering	301	3752
Machine learning	285	3699
Paleontology	238	3309
Pathology	244	2945
Context (archaeology)	187	2593
World Wide Web	208	2514
Economic growth	161	2273
Operating system	166	2242
Linguistics	172	2220
Mathematical analysis	145	2134
Social psychology	174	2118
Task (project management)	136	2117
Internal medicine	208	2044
Generative grammar	170	2027
Epistemology	142	1929
Physics	139	1906
Language model	133	1845
Domain (mathematical analysis)	123	1813
Computer security	140	1812
Geography	124	1741
Management	114	1732
Psychiatry	143	1693
Set (abstract data type)	111	1552
Medical physics	144	1540
Sociology	112	1492
Quantum mechanics	102	1458
Information retrieval	113	1448
Family medicine	134	1447

Table 3 (continued)

Keyword	Occurrences	Total link strength
Knowledge management	104	1439
Pure mathematics	102	1422
Field (mathematics)	98	1380
Test (biology)	96	1347
MEDLINE	112	1205
Radiology	103	1204
Engineering ethics	103	1198

the co-occurrence, the higher the link strength. Occurrence is the number of times a given keyword appears across the corpus.

Appendix 3 Representative Papers for Each LLM Task

Table 4 presents the representative papers for each LLM task and their respective DOI.

Appendix 4 Analysis of LLM Task: Meta-analysis and Literature Review

Systematic reviews and meta-analyses in this domain critically assess LLMs, focusing on their capacity to revolutionize various aspects of medical practice [9, 223–226] and providing guidelines on their applications [227]. One mainstream in this sub-topic focused on the comprehensive evaluations of different model performances, highlighting the strengths of LLMs in processing medical information and their potential to augment clinical decision-making while also acknowledging their limitations, such as occasional inaccuracies and biases [228–231]. Detailed investigations into the methodologies reveal how advanced techniques like generative pre-trained transformers [232] and fine-tuning [233] on medical datasets are applied to create innovative applications, from automated medical reporting to virtual patient engagement tools [234]. The other literature suggests future developments, such as emphasizing the need for richer training data [235, 236], enhancing interdisciplinary research collaborations [237], and setting up stringent ethical standards to ensure that LLMs can be safely integrated into patient care [228, 238]. However, they ultimately pave the way for more personalized and efficient healthcare solutions. This collective body of work benchmarks current LLM capabilities and charts a strategic course for their evolution in the healthcare domain.

Table 4 Representative papers for each LLM task

LLM task	Representative paper	DOI
Model evaluation	Evaluating large language models on medical evidence summarization	https://doi.org/10.1038/s41746-023-00896-7
	Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model	https://doi.org/10.21203/rs.3.rs-2566942/v1
	How large language models perform on the United States Medical Licensing Examination: a systematic review	https://doi.org/10.1101/2023.09.03.23294842
Information extraction	Exploring zero-shot capability of large language models in inferences from medical oncology notes	https://doi.org/10.48550/arxiv.2308.03853
	Potential of ChatGPT and GPT-4 for data mining of free-text CT reports on lung cancer	https://doi.org/10.1148/radiol.231362
	Development of a privacy preserving large language model for automated data extraction from thyroid cancer pathology reports	https://doi.org/10.1101/2023.11.08.23298252
Dialog and interactive systems	Understanding the benefits and challenges of deploying conversational AI leveraging large language models for public health intervention	https://doi.org/10.1145/3544548.3581503
	A novel AI-based chatbot application for personalized medical diagnosis and review using large language models	https://doi.org/10.1109/rmkmate59243.2023.10368616
	ChatGPT: a novel AI assistant for healthcare messaging—a commentary on its potential in addressing patient queries and reducing clinician burnout	https://doi.org/10.1136/leader-2023-000844
Multilinguality	Sailing the seven seas: a multinational comparison of ChatGPT's performance on medical licensing examinations	https://doi.org/10.1007/s10439-023-03338-3
	Evaluating the performance of ChatGPT in a dermatology specialty certificate examination: a comparative analysis between English and Korean language settings	https://doi.org/10.21203/rs.3.rs-3241164/v1
	Better to ask in English: cross-lingual evaluation of large language models for healthcare queries	https://doi.org/10.48550/arxiv.2310.13132

Table 4 (continued)

LLM task	Representative paper	DOI
Text generation	Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers	https://doi.org/10.1038/s41746-023-00819-6
	Artificial intelligence can generate fraudulent but authentic-looking scientific medical articles: Pandora's box has been opened	https://doi.org/10.2196/46924
	Automatic medical report generation via latent space conditioning and transformers	https://doi.org/10.1109/dasc/picom/cbdcom/cy59711.2023.10361320
Education	Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models	https://doi.org/10.1371/journal.pdig.0000198
	The rise of <scp>ChatGPT</scp>: exploring its potential in medical education	https://doi.org/10.1002/ase.2270
	Large language models in medical education: opportunities, challenges, and future directions	https://doi.org/10.2196/48291
Meta-analysis and literature review	Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review	https://doi.org/10.2196/48785
	ChatGPT in healthcare: a taxonomy and systematic review	https://doi.org/10.1101/2023.03.30.23287899
	Chat GPT in diagnostic human pathology: will it be useful to pathologists? a preliminary review with 'query session' and future perspectives	https://doi.org/10.3390/ai4040051
Ethics	Ethics of large language models in medicine and medical research	https://doi.org/10.1016/s2589-7500(23)00083-3
	A medical ethics framework for conversational artificial intelligence	https://doi.org/10.2196/43068
	Exploring the potential utility of AI large language models for medical ethics: an expert panel evaluation of GPT-4	https://doi.org/10.1136/jme-2023-109549

Table 4 (continued)

LLM task	Representative paper	DOI
Image, vision, video, and multimodality	ChatCAD: interactive computer-aided diagnosis on medical images using large language models	https://doi.org/10.48550/arxiv.2302.07257
	Medical image generative pre-trained transformer (MI-GPT): future direction for precision medicine	https://doi.org/10.1007/s00259-023-06450-7
	GPT-4 and medical image analysis: strengths, weaknesses and future directions	https://doi.org/10.21037/jmai-23-94
Scholarship and manuscript writing	Does GPT-3 qualify as a co-author of a scientific paper publishable in peer-reviewed journals according to the ICMJE criteria? A case study	https://doi.org/10.1007/s44163-023-00055-7
	Guiding principles and proposed classification system for the responsible adoption of artificial intelligence in scientific writing in medicine	https://doi.org/10.3389/frai.2023.1283353
	Harnessing large language models in medical research and scientific writing: a closer look to the future	https://doi.org/10.59707/hymrfbya5348
Inference	Pushing back on AI: a dialogue with ChatGPT on causal inference in epidemiology	https://doi.org/10.1007/978-3-031-32013-2_13
	Saama AI research at SemEval-2023 task 7: exploring the capabilities of Flan-T5 for multi-evidence natural language inference in clinical trial data	https://doi.org/10.18653/v1/2023.semeval-1.137
	GPT4MIA: utilizing generative pre-trained transformer (GPT-3) as a plug-and-play transductive model for medical image analysis	https://doi.org/10.1007/978-3-031-47401-9_15
Summarization	Evaluating large language models on medical evidence summarization	https://doi.org/10.1101/2023.04.22.23288967
	Performance analysis of large language models for medical text summarization	https://doi.org/10.31219/osf.io/kn5f2
	SummQA at MEDIQA-Chat 2023: in-context learning with GPT-4 for medical summarization	https://doi.org/10.18653/v1/2023.clinicalnlp-1.51

Table 4 (continued)

LLM task	Representative paper	DOI
Sentiment analysis	Sentiment analysis of COVID-19 survey data: a comparison of ChatGPT and fine-tuned OPT against widely used sentiment analysis tools (preprint)	https://doi.org/10.2196/preprints.50150
	Screening for depression using natural language processing (NLP): a literature review (preprint)	https://doi.org/10.2196/preprints.55067
	Applying BERT and ChatGPT for sentiment analysis of lyme disease in scientific literature	https://doi.org/10.48550/arxiv.2302.06474
Named entity recognition	DeID-GPT: zero-shot medical text de-identification by GPT-4	https://doi.org/10.48550/arxiv.2303.11032
	RIGA at SemEval-2023 task 2: NER enhanced with GPT-3	https://doi.org/10.18653/v1/2023.semeval-1.45
	Identification of ancient Chinese medical prescriptions and case data analysis under artificial intelligence GPT algorithm: a case study of Song Dynasty medical literature	https://doi.org/10.1109/access.2023.3330212

Appendix 5 Specialized and Contextualized Model Evaluation in Disease Categories

Model evaluation represents the largest portion of LLM tasks. Specifically, LLMs have been evaluated in their applications for detecting various diseases, from mental health conditions to infectious diseases (Fig. 9; Table 5). The classification tasks are usually the focus of model evaluation.

Technical literature on the use of LLMs for mental health analysis has examined the performance of LLMs and LLM-based ChatGPT on basic psychopharmacologic tasks [239], explanation generation of analysis results [240], detection of mental diseases and disorders [241], and so on. Such studies usually evaluate the performance of trained LLMs on pre-labeled datasets compared to a baseline model, with a focus on the accuracy of classification tasks and automatic evaluation metrics [157, 241, 242]. For instance, [241] evaluates LLM-based ChatGPT on mental health classification tasks with three publicly available datasets on stress, depression, and suicidality consisting of annotated social media posts with varying numbers of classes. The model achieved higher classification accuracy compared to a baseline model that always predicted the dominant class.

When datasets are not publicly available, researchers come up with classification tasks on their own in specific scenarios [239, 243]. For example, [239] created brief vignettes about the decision to select antidepressant treatment for adults with major depressive disorder, a basic psychopharmacologic task for clinicians. The authors created and validated the vignettes with experienced clinicians, against which the ChatGPT's ratings of the treatment options are compared.

Explanations of decisions are taken into account in understanding the decisions made by LLMs on classification tasks and analysis of health conditions and their explainability [206, 239, 240]. In addition to popular automatic evaluation metrics like perplexity,

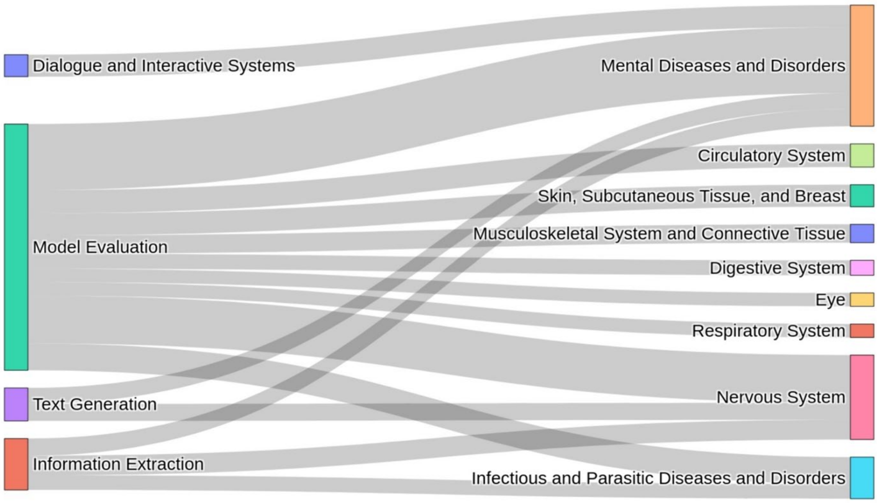

Fig. 9 Sankey diagram of LLM tasks and disease categories (with paper count more than 10)

Table 5 Disease categories, paper counts, and representative publications in model evaluation research category (paper count > 10)

Disease category	Paper count	Example paper	DOI
Mental diseases and disorders	54	Research letter: application of GPT-4 to select next-step antidepressant treatment in major depression	https://doi.org/10.1101/2023.04.14.23288595
		Benefits and harms of large language models in digital mental health	https://doi.org/10.48550/arxiv.2311.14693
Nervous system	39	Predicting seizure recurrence from medical records using large language models	https://doi.org/10.1016/s2589-7500(23)00205-4
		The utility of ChatGPT in the assessment of literature on the prevention of migraine: an observational, qualitative study	https://doi.org/10.3389/fneur.2023.1225223
Infectious and parasitic diseases and disorders	22	Working with AI to persuade: examining a large language model's ability to generate pro-vaccination messages	https://doi.org/10.1145/3579592
		Leveraging large language models and weak supervision for social media data annotation: an evaluation using COVID-19 self-reported vaccination tweets	https://doi.org/10.1007/978-3-031-48044-7_26
Circulatory system	19	Uncovering language disparity of ChatGPT in healthcare: non-english clinical environment for retinal vascular disease classification (preprint)	https://doi.org/10.2196/preprints.51926
		ChatGPT exhibits gender and racial biases in acute coronary syndrome management	https://doi.org/10.1101/2023.11.14.23298525
Skin, subcutaneous tissue, and breast	18	Performance of three large language models on dermatology board examinations	https://doi.org/10.1016/j.jid.2023.06.208
		The chatbots are coming: risks and benefits of consumer-facing artificial intelligence in clinical dermatology	https://doi.org/10.1016/j.jaad.2023.05.088
Musculoskeletal system and connective tissue	15	Search for medical information and treatment options for musculoskeletal disorders through an artificial intelligence chatbot: focusing on shoulder impingement syndrome	https://doi.org/10.1101/2022.12.16.22283512
		Large language models and the future of rheumatology: assessing impact and emerging opportunities	https://doi.org/10.1097/bor.0000000000000981

Table 5 (continued)

Disease category	Paper count	Example paper	DOI
Digestive system	12	Advanced prompting as a catalyst: empowering large language models in the management of gastrointestinal cancers	https://doi.org/10.59717/j.xinn-med.2023.100019
		Large language models for granularized Barrett's esophagus diagnosis classification	https://doi.org/10.48550/arxiv.2308.08660
Respiratory system	11	Natural language processing for COVID-19 consulting system	https://doi.org/10.1016/j.procs.2023.01.112
Eye	11	Chat generative pretrained transformer to optimize accessibility for cataract surgery postoperative management	https://doi.org/10.4103/pajo.pajo_51_23
		Ophtha-LLaMA2: a large language model for ophthalmology	https://doi.org/10.48550/arxiv.2312.04906

BLEU-n, and ROUGE-1 [206, 242], studies also use human annotation for evaluation and for benchmarking automatic evaluation metrics [240, 244]. Additionally, approaches based on prompt engineering are also taken to evaluate the interaction between LLMs and agents by analyzing their mental health referral patterns [245]. Apart from technical literature, other research has also examined and identified the benefits and harms of using LLMs for mental health counseling [246, 247] and the issues of hallucination [244].

Author Contribution H.Y. and L.F. contributed to the ideal conceptualization and formal analysis. H.Y., L.F., L.L., J.Z., Z.M., L.X., W.H., S.H., and M.J. contributed to manuscript writing. H.Y. was responsible for producing all figures and tables. Y.Z., A.G., and X.M. are responsible for result validation and draft editing. All authors reviewed the submitted manuscript.

Data Availability The dataset of the papers analyzed for this manuscript is available from the corresponding author upon request.

Declarations

Ethics Approval and Consent to Participate Not applicable.

Consent for Publication Not applicable.

Competing Interests The authors declare no competing interests.

References

1. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW et al (2023) Large language models encode clinical knowledge. Nature 620:172–180
2. Karabacak M, Margetis K (2023) Embracing large language models for medical applications: opportunities and challenges. Cureus 15:e39305
3. Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt J-N, Laleh NG et al (2023) The future landscape of large language models in medicine. Commun Med. 3:141
4. OpenAI. Introducing ChatGPT. 30 Nov 2022. https://openai.com/blog/chatgpt. Accessed 12 Mar 2024
5. Tseng R, Verberne S, van der Putten P. ChatGPT as a commenter to the news: can LLMs generate human-like opinions? Disinformation in open online media. Springer Nature Switzerland; 2023. pp. 160–174.
6. Ma Y, Liu J, Yi F, Cheng Q, Huang Y, Lu W et al (2023) AI vs. human -- differentiation analysis of scientific content generation. arXiv [cs.CL]. http://arxiv.org/abs/2301.10416. Accessed 12 Feb 2023
7. Danner M, Hadzic B, Gerhardt S, Ludwig S, Uslu I, Shao P, Weber T, Shiban Y, Ratsch M (2023) Advancing mental health diagnostics: GPT-based method for depression detection. 2023 62nd Annual Conference of the Society of Instrument and Control Engineers (SICE). IEEE, Tsu, Japan, pp. 1290–1296. https://doi.org/10.23919/SICE59929.2023.10354236
8. Srivastava R, Srivastava S (2023) Can artificial intelligence aid communication? Considering the possibilities of GPT-3 in palliative care. Indian J Palliat Care 29:418–425
9. Ghim J-L, Ahn S (2023) Transforming clinical trials: the emerging roles of large language models. Transl Clin Pharmacol 31:131–138
10. Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G, Moy L et al (2023) ChatGPT and other large language models are double-edged swords. Radiology 307(2):e230163. https://doi.org/10.1148/radiol.230163
11. Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y et al (2023) A survey of large language models. arXiv [cs.CL]. http://arxiv.org/abs/2303.18223v13. Accessed 9 Apr 2023

12. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN et al (2017) Attention is all you need. Adv Neural Inf Process Syst 30. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf. Accessed 2 Aug 2023
13. Devlin J, Chang M-W, Lee K, Toutanova K (2018) BERT: pre-training of deep bidirectional transformers for language understanding. arXiv [cs.CL]. http://arxiv.org/abs/1810.04805
14. Floridi L, Chiriatti M (2020) GPT-3: its nature, scope, limits, and consequences. Minds Mach 30:681–694
15. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M et al (2019) Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv [cs.LG]. http://arxiv.org/abs/1910.10683. Accessed 19 Sept 2023
16. Sun Y, Wang S, Li Y, Feng S, Tian H, Wu H et al (2020) ERNIE 2.0: a continual pre-training framework for language understanding. AAAI 34: 8968–8975
17. Black S, Biderman S, Hallahan E, Anthony Q, Gao L, Golding L et al (2022) GPT-NeoX-20B: an open-source autoregressive language model. arXiv [cs.CL]. http://arxiv.org/abs/2204.06745. Accessed 1 May 2023
18. Yang J, Jin H, Tang R, Han X, Feng Q, Jiang H et al (2023) Harnessing the power of LLMs in practice: a survey on ChatGPT and beyond. arXiv [cs.CL]. http://arxiv.org/abs/2304.13712. Accessed 1 May 2023
19. Fan L, Hua W, Li L, Ling H, Zhang Y (2023) NPHardEval: dynamic benchmark on reasoning ability of large language models via complexity classes. arXiv [cs.AI]. http://arxiv.org/abs/2312.14890. Accessed 1 May 2023
20. Fan L, Hua W, Li X, Zhu K, Jin M, Li L et al (2024) NPHardEval4V: a dynamic reasoning benchmark of multimodal large language models. arXiv [cs.CL]. http://arxiv.org/abs/2403.01777. Accessed 21 Apr 2024
21. Anthropic (2023) Claude 2. [cited 12 Mar 2024]. https://www.anthropic.com/news/claude-2
22. Google (2023) Introducing Gemini: our largest and most capable AI model. [cited 12 Mar 2024]. Available: https://blog.google/technology/ai/google-gemini-ai/
23. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y et al (2023) Llama 2: open foundation and fine-tuned chat models. arXiv [cs.CL]. http://arxiv.org/abs/2307.09288. Accessed 19 July 2023
24. Li Y, Bubeck S, Eldan R, Del Giorno A, Gunasekar S, Lee YT (2023) Textbooks are all you need II: phi-1.5 technical report. arXiv [cs.CL]. http://arxiv.org/abs/2309.05463. Accessed 1 Nov 2023
25. Wang Y, Wu S, Li D, Mehrabi S, Liu H (2016) A part-of-speech term weighting scheme for biomedical information retrieval. J Biomed Inform 63:379–389
26. Bui Q-C, Sloot PMA, van Mulligen EM, Kors JA (2014) A novel feature-based approach to extract drug-drug interactions from biomedical text. Bioinformatics 30:3365–3371
27. Rink B, Harabagiu S, Roberts K (2011) Automatic extraction of relations between medical concepts in clinical texts. J Am Med Inform Assoc 18:594–600
28. Habibi M, Weber L, Neves M, Wiegandt DL, Leser U (2017) Deep learning with word embeddings improves biomedical named entity recognition. Bioinformatics 33:i37–i48
29. Jiang Z, Li L, Huang D, Jin L (2015) Training word embeddings for deep learning in biomedical text mining tasks. 2015 IEEE international conference on bioinformatics and biomedicine (BIBM). IEEE. pp. 625–628
30. Peng Y, Yan S, Lu Z (2019) Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. arXiv [cs.CL]. http://arxiv.org/abs/1906.05474
31. Yao L, Jin Z, Mao C, Zhang Y, Luo Y (2019) Traditional Chinese medicine clinical records classification with BERT and domain specific corpora. J Am Med Inform Assoc 26:1632–1636
32. Prakash PKS, Chilukuri S, Ranade N, Viswanathan S (2021) RareBERT: transformer architecture for rare disease patient identification using administrative claims. AAAI 35:453–460
33. Kawazoe Y, Shibata D, Shinohara E, Aramaki E, Ohe K (2021) A clinical specific BERT developed using a huge Japanese clinical text corpus. PLoS One. 16:e0259763
34. Yu H, Fan L, Gilliland AJ (2022) Disparities and resilience: analyzing online health information provision, behaviors and needs of LBGTQ + elders during COVID-19. BMC Public Health 22:2338
35. Hakala K, Pyysalo S (2019) Biomedical named entity recognition with multilingual BERT. In: Jin-Dong K, Claire N, Robert B, Louise D, editors. Proceedings of the 5th Workshop on BioNLP Open Shared Tasks. Association for Computational Linguistics, Hong Kong, China, pp 56–61
36. Sun C, Yang Z, Wang L, Zhang Y, Lin H, Wang J (2021) Biomedical named entity recognition using BERT in the machine reading comprehension framework. J Biomed Inform. 118:103799

37. Roy A, Pan S (2021) Incorporating medical knowledge in BERT for clinical relation extraction. In: Moens M-F, Huang X, Specia L, Yih SW-T, editors. Proceedings of the 2021 conference on empirical methods in natural language processing. Online and Punta Cana, Association for Computational Linguistics, Dominican Republic pp 5357–5366
38. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH et al (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36:1234–1240
39. Alsentzer E, Murphy JR, Boag W, Weng W-H, Jin D, Naumann T et al (2019) Publicly available clinical BERT embeddings. arXiv [cs.CL]. http://arxiv.org/abs/1904.03323. Accessed 1 May 2023
40. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW (2023) Large language models in medicine. Nat Med 29:1930–1940
41. Kuroiwa T, Sarcon A, Ibara T, Yamada E, Yamamoto A, Tsukamoto K et al (2023) The potential of ChatGPT as a self-diagnostic tool in common orthopedic diseases: exploratory study. J Med Internet Res 25:e47621
42. Caruccio L, Cirillo S, Polese G, Solimando G, Sundaramurthy S, Tortora G (2024) Can ChatGPT provide intelligent diagnoses? A comparative study between predictive models and ChatGPT to define a new medical diagnostic bot. Expert Syst Appl 235:121186
43. Koga S, Martin NB, Dickson DW (2023) Evaluating the performance of large language models: ChatGPT and google bard in generating differential diagnoses in clinicopathological conferences of neurodegenerative disorders. Brain Pathol e13207
44. Jin M, Yu Q, Shu D, Zhang C, Zhu S, Du M et al (2024) Health-LLM: personalized retrieval-augmented disease prediction system. arXiv [cs.CL]. http://arxiv.org/abs/2402.00746. Accessed 19 Feb 2024
45. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C et al (2022) A large language model for electronic health records. NPJ Digit Med 5:194
46. Al-Ashwal FY, Zawiah M, Gharaibeh L, Abu-Farha R, Bitar AN (2023) Evaluating the sensitivity, specificity, and accuracy of ChatGPT-3.5, ChatGPT-4, Bing AI, and Bard against conventional drug-drug interactions clinical tools. Drug Healthc Patient Saf 15:137–147
47. Gao Z, Li L, Ma S, Wang Q, Hemphill L, Xu R (2023) Examining the potential of ChatGPT on biomedical information retrieval: fact-checking drug-disease associations. Ann Biomed Eng. https://doi.org/10.1007/s10439-023-03385-w
48. Eysenbach G (2023) The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. JMIR Med Educ. 9:e46885
49. Lee H (2023) The rise of ChatGPT: exploring its potential in medical education. Anat Sci Educ. https://doi.org/10.1002/ase.2270
50. Sallam M (2023) ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel) 11. https://doi.org/10.3390/healthcare11060887
51. Li L, Ma Z, Fan L, Lee S, Yu H, Hemphill L (2023) ChatGPT in education: a discourse analysis of worries and concerns on social media. Educ Inf Technol. https://doi.org/10.1007/s10639-023-12256-9
52. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C et al (2023) Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2:e0000198
53. Li J, Dada A, Puladi B, Kleesiek J, Egger J (2024) ChatGPT in healthcare: a taxonomy and systematic review. Comput Methods Programs Biomed 245:108013
54. Tian S, Jin Q, Yeganova L, Lai P-T, Zhu Q, Chen X et al (2023) Opportunities and challenges for ChatGPT and large language models in biomedicine and health. Brief Bioinform 25. https://doi.org/10.1093/bib/bbad493
55. Fan L, Li L, Ma Z, Lee S, Yu H, Hemphill L (2024) A bibliometric review of large language models research from 2017 to 2023. ACM Trans Intell Syst Technol. https://doi.org/10.1145/3664930
56. Li L, Zhou J, Gao Z, Hua W, Fan L, Yu H et al (2024) A scoping review of using large language models (LLMs) to investigate electronic health records (EHRs). arXiv [cs.ET]. https://scholar.google.com/citations?view_op=view_citation&hl=en&user=kO-WycAAAAAJ&cstart=20&pagesize=80&citation_for_view=kO-WycAAAAAJ:iH-uZ7U-co4C. Accessed 20 May 2024
57. Thapa S, Adhikari S (2023) ChatGPT, Bard, and large language models for biomedical research: opportunities and pitfalls. Ann Biomed Eng 51:2647–2651

58. Cheng H, Liu S, Sun W, Sun Q (2023) A neural topic modeling study integrating SBERT and data augmentation. Appl Sci (Basel) 13:4595
59. Hott HR, Silva MO, Oliveira GP, Brandão MA, Lacerda A, Pappa G (2023) Evaluating contextualized embeddings for topic modeling in public bidding domain. Intelligent Systems. Springer Nature Switzerland, Cham, pp 410–426
60. Berlanga R, Soriano M (2024) Explaining semantic text similarity in knowledge graphs. Progress in pattern recognition, image analysis, computer vision, and applications. Springer Nature Switzerland, pp 526–539.
61. Grootendorst M (2022) BERTopic: neural topic modeling with a class-based TF-IDF procedure. arXiv [cs.CL]. http://arxiv.org/abs/2203.05794. Accessed 1 May 2023
62. Guo Z, Zhu L, Han L (2021) Research on short text classification based on RoBERTa-TextRCNN. 2021 International conference on Computer Information Science and Artificial Intelligence (CISAI). IEEE, pp. 845–849. https://doi.org/10.1109/CISAI54367.2021.00171
63. Xu Z (2021) RoBERTa-wwm-ext fine-tuning for Chinese text classification. arXiv [cs.CL]. http://arxiv.org/abs/2103.00492. Accessed 1 May 2023
64. Chang W-C, Yu H-F, Zhong K, Yang Y, Dhillon IS (2020) Taming pretrained transformers for extreme multi-label text classification. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Association for Computing Machinery, New York, NY, USA, pp 3163–3171
65. Yin W, Hay J, Roth D (Available:) Benchmarking zero-shot text classification: datasets, evaluation and entailment approach. arXiv [cs.CL]. http://arxiv.org/abs/1909.00161. Accessed 1 May 2023
66. VOSviewer (2022) VOSviewer - visualizing scientific landscapes. In: VOSviewer [Internet]. https://www.vosviewer.com/. Accessed 12 Mar 2024
67. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA et al (2023) How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 9:e45312
68. Adams LC, Truhn D, Busch F, Kader A, Niehues SM, Makowski MR et al (2023) Leveraging GPT-4 for post hoc transformation of free-text radiology reports into structured reporting: a multilingual feasibility study. Radiology 307:e230725
69. Haver HL, Ambinder EB, Bahl M, Oluyemi ET, Jeudy J, Yi PH (2023) Appropriateness of breast cancer prevention and screening recommendations provided by ChatGPT. Radiology 307:e230424
70. Sun Z, Ong H, Kennedy P, Tang L, Chen S, Elias J et al (2023) Evaluating GPT4 on impressions generation in radiology reports. Radiology 307:e231259
71. Bhayana R, Krishna S, Bleakney RR (2023) Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. Radiology 307:e230582
72. Thurzo A, Strunga M, Urban R, Surovková J, Afrashtehfar KI (2023) Impact of artificial intelligence on dental education: a review and guide for curriculum update. Educ Sci 13:150
73. Huang H, Zheng O, Wang D, Yin J, Wang Z, Ding S et al (2023) ChatGPT for shaping the future of dentistry: the potential of multi-modal large language model. Int J Oral Sci 15:29
74. Surovková J, Haluzová S, Strunga M, Urban R, Lifková M, Thurzo A (2023) The new role of the dental assistant and nurse in the age of advanced artificial intelligence in telehealth orthodontic care with dental monitoring: preliminary report. NATO Adv Sci Inst Ser E Appl Sci 13:5212
75. Cervera MR, Bermejo-Peláez D, Gómez-Álvarez M, Hidalgo Soto M, Mendoza-Martínez A, Oñós Clausell A et al (2023) Assessment of artificial intelligence language models and information retrieval strategies for QA in hematology. Blood 142:7175–7175
76. Agrawal M, Hegselmann S, Lang H, Kim Y, Sontag D (2022) Large language models are few-shot clinical information extractors. In: Goldberg Y, Kozareva Z, Zhang Y (eds) Proceedings of the 2022 conference on empirical methods in natural language processing. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, pp 1998–2022
77. Gero Z, Singh C, Cheng H, Naumann T, Galley M, Gao J et al (2023) Self-verification improves few-shot clinical information extraction. arXiv [cs.CL]. http://arxiv.org/abs/2306.00024. Accessed 1 Jun 2023
78. Goel A, Gueta A, Gilon O, Liu C, Erell S, Nguyen LH et al (2023) LLMs accelerate annotation for medical information extraction. arXiv [cs.CL]. http://arxiv.org/abs/2312.02296. Accessed 20 Dec 2023
79. Hu D, Liu B, Zhu X, Lu X, Wu N (2024) Zero-shot information extraction from radiological reports using ChatGPT. Int J Med Inform 183:105321
80. Shyr C, Hu Y, Bastarache L, Cheng A, Hamid R, Harris P et al (2024) Identifying and extracting rare diseases and their phenotypes with large language models. J Healthc Inform Res 8:438–461

81. Chen J, Chen P, Wu X (2023) Generating Chinese event extraction method based on ChatGPT and prompt learning. NATO Adv Sci Inst Ser E Appl Sci 13:9500
82. Wang L, Ma Y, Bi W, Lv H, Li Y (2023) An entity extraction pipeline for medical text records utilizing large language models: an analytical study. In: JMIR Preprints [Internet]. [cited 12 Mar 2024]. https://preprints.jmir.org/preprint/54580
83. Sousa H, Guimarães N, Jorge A, Campos R (2023) GPT struct me: probing GPT models on narrative entity extraction. arXiv [cs.CL]. http://arxiv.org/abs/2311.14583. Accessed 20 Dec 2023
84. Mohammed S, Fiaidhi J, Shaik H (2023) Empowering transformers for evidence-based medicine. medRxiv 2023.12.25.23300520. https://doi.org/10.1101/2023.12.25.23300520
85. Goenaga I, Atutxa A, Gojenola K, Oronoz M, Agerri R (2023) Explanatory argument extraction of correct answers in resident medical exams. arXiv [cs.CL]. http://arxiv.org/abs/2312.00567. Accessed 20 Dec 2023
86. Jethani N, Jones S, Genes N, Major VJ, Jaffe IS, Cardillo AB, et al (2023) Evaluating ChatGPT in information extraction: a case study of extracting cognitive exam dates and scores. medRxiv. 2023.07.10.23292373. https://doi.org/10.1101/2023.07.10.23292373
87. Bitterman DS, Goldner E, Finan S, Harris D, Durbin EB, Hochheiser H et al (2023) An end-to-end natural language processing system for automatically extracting radiation therapy events from clinical texts. Int J Radiat Oncol Biol Phys 117:262–273
88. Chen S, Guevara M, Ramirez N, Murray A, Warner JL, Aerts HJWL et al (2023) Natural language processing to automatically extract the presence and severity of esophagitis in notes of patients undergoing radiotherapy. JCO Clin Cancer Inform 7:e2300048
89. Mahajan D, Liang JJ, Tsou C-H, Uzuner Ö (2023) Overview of the 2022 n2c2 shared task on contextualized medication event extraction in clinical notes. J Biomed Inform 144:104432
90. Chen A, Yu Z, Yang X, Guo Y, Bian J, Wu Y (2023) Contextualized medication information extraction using transformer-based deep learning architectures. arXiv [cs.CL]. http://arxiv.org/abs/2303.08259
91. Trevena W, Zhong X, Alvarado M, Semenov A, Oktay A, Devlin D et al (2023) Utilizing open-source language models and ChatGPT for zero-shot identification of drug discontinuation events in online forums: development and validation study. In: JMIR Preprints [Internet]. Available: https://preprints.jmir.org/preprint/54601. Accessed 12 Mar 2024
92. Tu H, Han L, Nenadic G (2023) Extraction of medication and temporal relation from clinical text using neural language models. arXiv [cs.CL]. Available: http://arxiv.org/abs/2310.02229. Accessed 20 Dec 2023
93. Abu-Ashour W, Emil S, Poenaru D (2023) Using artificial intelligence to label free-text operative and ultrasound reports for grading pediatric appendicitis. medRxiv 2023.08.30.23294850. https://doi.org/10.1101/2023.08.30.23294850
94. He J, Li F, Li J, Hu X, Nian Y, Xiang Y et al (2024) Prompt tuning in biomedical relation extraction. J Healthc Inform Res 8:206–224
95. Ramachandran GK, Fu Y, Han B, Lybarger K, Dobbins NJ, Uzuner Ö et al (2023) Prompt-based extraction of social determinants of health using few-shot learning. arXiv [cs.CL]. http://arxiv.org/abs/2306.07170
96. Bhate N, Mittal A, He Z, Luo X (2023) Zero-shot learning with minimum instruction to extract social determinants and family history from clinical notes using GPT model. arXiv [cs.CL]. http://arxiv.org/abs/2309.05475. Accessed 20 Dec 2023
97. Chakraborty C, Bhattacharya M, Lee S-S (2024) Need an AI-enabled, next-generation, advanced ChatGPT or large language models (LLMs) for error-free and accurate medical information. Ann Biomed Eng 52:134–135
98. Guevara M, Chen S, Thomas S, Chaunzwa TL, Franco I, Kann BH et al (2024) Large language models to identify social determinants of health in electronic health records. NPJ Digit Med 7:6
99. Derton A, Guevara M, Chen S, Moningi S, Kozono DE, Liu D et al (2023) Natural language processing methods to empirically explore social contexts and needs in cancer patient notes. JCO Clin Cancer Inform 7:e2200196
100. Cox LA Jr (2023) Pushing back on AI: a dialogue with ChatGPT on causal inference in epidemiology. In: Cox LA (ed) AI-ML for decision and risk analysis: challenges and opportunities for normative decision theory. Springer International Publishing, Cham, pp 407–423
101. Kanakarajan KR, Sankarasubbu M (2023) Saama AI research at SemEval-2023 Task 7: exploring the capabilities of Flan-T5 for multi-evidence natural language inference in clinical trial data. In: Ojha AK, Doğruöz AS, Da San Martino G, Tayyar Madabushi H, Kumar R, Sartori E (eds),

Proceedings of the 17th international workshop on semantic evaluation (SemEval-2023). Association for Computational Linguistics, Toronto, Canada, pp 995–1003
102. Zhang Y, Chen DZ (2023) GPT4MIA: utilizing generative pre-trained transformer (GPT-3) as a plug-and-play transductive model for medical image analysis. arXiv [cs.CV]. http://arxiv.org/abs/2302.08722. Accessed 1 May 2023
103. Lossio-Ventura JA, Weger R, Lee AY, Guinee EP, Chung J, Atlas L et al (2024) A comparison of ChatGPT and fine-tuned open pre-trained transformers (OPT) against widely used sentiment analysis tools: sentiment analysis of COVID-19 survey data. JMIR Ment Health 11:e50150
104. De S, Vats S (2023) Decoding concerns: multi-label classification of vaccine sentiments in social media. arXiv [cs.CL]. http://arxiv.org/abs/2312.10626
105. Abramski KE, Citraro S, Lombardi L, Rossetti G, Stella M (2023) Cognitive network science reveals bias in GPT-3, ChatGPT, and GPT-4 mirroring math anxiety in high-school students. https://doi.org/10.31234/osf.io/27u6z
106. Clarke P, Leininger C, Principato C, Staples P, Goodwin GM, Ryslik GA et al (2023) From a large language model to three-dimensional sentiment. https://doi.org/10.31234/osf.io/kaeqy
107. Mittal S, De Choudhury M (2023) Moral framing of mental health discourse and its relationship to stigma: a comparison of social media and news. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, pp. 1–19
108. Szántó Z, Bánáti B, Zombori T (2023) Enhancing Medication Event Classification with Syntax Parsing and Adversarial Learning. In: Maglogiannis I, Iliadis L, MacIntyre J, Dominguez M (eds) Artificial Intelligence Applications and Innovations. AIAI 2023. IFIP Advances in Information and Communication Technology, vol 675. Springer, Cham. https://doi.org/10.1007/978-3-031-34111-3_11
109. Zhang X, Ansah AA (2023) A mobile app for tracking psychological mood changes and providing E-therapy using natural language processing and GPT-3. Artificial Intelligence & Applications. Academy & Industry Research Collaboration Center. https://doi.org/10.5121/csit.2023.131925
110. Gómez-Zaragozá L, Minissi ME, Llanes-Jurado J, Altozano A, Alcañiz Raya M, Marín-Morales J (2023) Linguistic indicators of depressive symptoms in conversations with virtual humans. Collaborative Networks in Digitalization and Society 50. Springer Nature Switzerland, pp. 521–534.
111. Qi H, Zhao Q, Li J, Song C, Zhai W, Dan L et al (2023) Supervised learning and large language model benchmarks on mental health datasets: cognitive distortions and suicidal risks in Chinese social media. [cited 12 Mar 2024]. https://doi.org/10.21203/rs.3.rs-3523508/v1
112. Theophilou E, Koyuturk C, Yavari M, Bursic S, Donabauer G, Telari A, et al. Learning to prompt in the classroom to understand AI limits: a pilot study. arXiv [cs.HC]. 2023. Available: http://arxiv.org/abs/2307.01540. Accessed 20 Dec 2023
113. Forman N, Udvaros J, Avornicului MS (2023) ChatGPT: a new study tool shaping the future for high school students. IJANSER 7:95–102
114. Abouammoh N, Alhasan K, Raina R, Malki KA, Aljamaan F, Tamimi I et al (2023) Exploring perceptions and experiences of ChatGPT in medical education: a qualitative study among medical college faculty and students in Saudi Arabia. bioRxiv. https://doi.org/10.1101/2023.07.13.23292624
115. Ackerman R, Balyan R (2023) Automatic multilingual question generation for health data using LLMs. https://doi.org/10.1007/978-981-99-7587-7_1
116. Gin BC, ten Cate O, O'Sullivan PS, Boscardin CK (2023) Trainee versus supervisor viewpoints of entrustment: using artificial intelligence language models to detect thematic differences and potential biases. https://doi.org/10.21203/rs.3.rs-3223749/v1
117. Perlis Roy H., Jones David S (2023) High-impact medical journals reflect negative sentiment toward psychiatry. NEJM AI 1: AIcs2300066
118. Frei J, Kramer F (2023) Annotated dataset creation through large language models for non-English medical NLP. J Biomed Inform 145:104478
119. Fontaine X, Gaschi F, Rastin P, Toussaint Y (2023) Multilingual Clinical NER: translation or cross-lingual transfer? arXiv [cs.CL]. http://arxiv.org/abs/2306.04384. Accessed 1 Jul 2023
120. Li M, Zheng X (2023) Identification of Ancient Chinese medical prescriptions and case data analysis under artificial intelligence GPT algorithm: a case study of song dynasty medical literature. IEEE Access 11:131453–131464
121. Lee Y-Q, Chen C-T, Chen C-C, Lee C-H, Chen P, Wu C-S et al (2024) Unlocking the secrets behind advanced artificial intelligence language models in deidentifying Chinese-English mixed clinical text: development and validation study. J Med Internet Res 26:e48443

122. Alfertshofer M, Hoch CC, Funk PF, Hollmann K, Wollenberg B, Knoedler S et al (2023) Sailing the seven seas: a multinational comparison of ChatGPT's performance on medical licensing examinations. Ann Biomed Eng. https://doi.org/10.1007/s10439-023-03338-3
123. Zong H, Li J, Wu E, Wu R, Lu J, Shen B (2023) Performance of ChatGPT on Chinese national medical licensing examinations: a five-year examination evaluation study for physicians, pharmacists and nurses. bioRxiv. https://doi.org/10.1101/2023.07.09.23292415
124. Jin Y, Chandra M, Verma G, Hu Y, De Choudhury M, Kumar S (2023) Better to ask in English: cross-lingual evaluation of large language models for healthcare queries. arXiv [cs.CL]. http://arxiv.org/abs/2310.13132. Accessed 1 Nov 2023
125. Khorshidi H, Mohammadi A, Yousem DM, Abolghasemi J, Ansari G, Mirza-Aghazadeh-Attari M et al (2023) Application of ChatGPT in multilingual medical education: how does ChatGPT fare in 2023's Iranian residency entrance examination. Inform Med Unlocked 41:101314
126. Yeo YH, Samaan JS, Ng WH, Ma X, Ting P-S, Kwak M-S et al (2023) GPT-4 outperforms ChatGPT in answering non-English questions related to cirrhosis. bioRxiv. https://doi.org/10.1101/2023.05.04.23289482
127. Fang C, Ling J, Zhou J, Wang Y, Liu X, Jiang Y et al (2023) How does ChatGPT4 preform on Non-English National Medical Licensing Examination? An Evaluation in Chinese Language. bioRxiv. https://doi.org/10.1101/2023.05.03.23289443
128. Türkmen H, Dikenelli O, Eraslan C, Çallı MC, Özbek SS (2023) BioBERTurk: exploring Turkish biomedical language model development strategies in low-resource setting. J Healthc Inform Res 7:433–446
129. Kunitsu Y (2023) The Potential of GPT-4 as a support tool for pharmacists: analytical study using the Japanese National Examination for Pharmacists. JMIR Med Educ 9:e48452
130. Eggmann F, Weiger R, Zitzmann NU, Blatz MB (2023) Implications of large language models such as ChatGPT for dental medicine. J Esthet Restor Dent 35:1098–1102
131. Liao W, Liu Z, Dai H, Xu S, Wu Z, Zhang Y et al (2023) Differentiate ChatGPT-generated and human-written medical texts. arXiv [cs.CL]. http://arxiv.org/abs/2304.11567
132. Li K, Hong S, Fu C, Zhang Y, Liu M (2023) Discriminating human-authored from ChatGPT-generated code via discernable feature analysis. 2023 IEEE 34th International Symposium on Software Reliability Engineering Workshops (ISSREW), pp 120-127
133. Alawida M, Mejri S, Mehmood A, Chikhaoui B, Isaac Abiodun O (2023) A comprehensive study of ChatGPT: advancements, limitations, and ethical considerations in natural language processing and cybersecurity. Information 14:462
134. Wang JTH (2023) Is the laboratory report dead? AI and ChatGPT. Microbiol Aust 144–148.
135. Abuyaman O (2023) Strengths and weaknesses of ChatGPT models for scientific writing about medical vitamin B12: mixed methods study. JMIR Form Res 7:e49459
136. Grigio TR, Timmerman H, Wolff AP (2023) ChatGPT in anaesthesia research: risk of fabrication in literature searches. Br J Anaesth 131:e29–e30
137. Májovský M, Černý M, Kasal M, Komarc M, Netuka D (2023) Artificial intelligence can generate fraudulent but authentic-looking scientific medical articles: Pandora's Box has been opened. J Med Internet Res 25:e46924
138. Gao CA, Howard FM, Markov NS, Dyer EC, Ramesh S, Luo Y et al (2023) Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. NPJ Digit Med 6:75
139. Huespe IA, Echeverri J, Khalid A, Carboni Bisso I, Musso CG, Surani S et al (2023) Clinical research with large language models generated writing-clinical research with AI-assisted writing (CRAW) Study. Crit Care Explor 5:e0975
140. Hamed AA, Wu X (2023) Detection of ChatGPT fake science with the xFakeBibs Learning algorithm. arXiv [cs.CL]. http://arxiv.org/abs/2308.11767. Accessed 1 Sept 2023
141. Katib I, Assiri FY, Abdushkour HA, Hamed D, Ragab M (2023) Differentiating chat generative pretrained transformer from humans: detecting ChatGPT-generated text and human text using machine learning. Sci China Ser A Math 11:3400
142. Leung TI, de Azevedo Cardoso T, Mavragani A, Eysenbach G (2023) Best practices for using AI tools as an author, peer reviewer, or editor. J Med Internet Res 25:e51584
143. Waisberg E, Ong J, Masalkhi M, Zaman N, Tavakkoli A (2023) Chat generative pretrained transformer to optimize accessibility for cataract surgery postoperative management. The Pan-Am J Ophthalmol 5. https://doi.org/10.4103/pajo.pajo_51_23

144. Lim S, Schmälzle R (2023) Artificial intelligence for health message generation: an empirical study using a large language model (LLM) and prompt engineering. Front Commun 8. https://doi.org/10.3389/fcomm.2023.1129082
145. Xie Y, Seth I, Hunter-Smith DJ, Rozen WM, Ross R, Lee M (2023) Aesthetic surgery advice and counseling from artificial intelligence: a rhinoplasty consultation with ChatGPT. Aesthetic Plast Surg 47:1985–1993
146. Karinshak E, Liu SX, Park JS, Hancock JT (2023) Working with AI to persuade: examining a large language model's ability to generate pro-vaccination messages. Proc ACM Hum-Comput Interact 7:1–29
147. Meskó B (2023) The impact of multimodal large language models on health care's future. J Med Internet Res 25:e52865
148. Temsah R, Altamimi I, Alhasan K, Temsah M-H, Jamal A (2023) Healthcare's new horizon with ChatGPT's voice and vision capabilities: a leap beyond text. Cureus 15:e47469
149. Waisberg E, Ong J, Masalkhi M, Zaman N, Sarker P, Lee AG et al (2023) GPT-4 and medical image analysis: strengths, weaknesses and future directions. J Med Artif Intell 6:29–29
150. Li X, Zhang I, Wu Z, Liu Z, Zhao l, Yuan Y et al (2023) artificial general intelligence for medical imaging. arXiv [cs.AI]. http://arxiv.org/abs/2306.05480. Accessed 1 Sept 2023
151. Hu M, Pan S, Li Y, Yang X (2023) Advancing medical imaging with language models: a journey from N-grams to ChatGPT. arXiv [cs.CV]. http://arxiv.org/abs/2304.04920. Accessed 1 May 2023
152. Liu Z, Jiang H, Zhong T, Wu Z, Ma C, Li Y et al (2023) Holistic evaluation of GPT-4V for biomedical imaging. [cited 13 Mar 2024]. Available: https://paperswithcode.com/paper/holistic-evaluation-of-gpt-4v-for-biomedical. Accessed 3 Dec 2023
153. Sim JZT, Bhanu Prakash KN, Huang WM, Tan CH (2023) Harnessing artificial intelligence in radiology to augment population health. Front Med Technol 5:1281500
154. Daungsupawong H, Wiwanitkit V (2024) Transforming radiology with ai visual chatbot. J Am Coll Radiol 21:3
155. Davies NM (2023) Adapting artificial intelligence into the evolution of pharmaceutical sciences and publishing: Technological Darwinism. J Pharm Pharm Sci 26:11349. Accessed 1 May 2023
156. Awan A, Gonzalez A, Sharma M (2023) A Neoteric approach toward social media in public health informatics: a narrative review of current trends and future directions. https://doi.org/10.20944/preprints202312.2102.v1
157. Chen Q, Hu X, Wang Z, Hong Y (2023) MedBLIP: bootstrapping language-image pre-training from 3D medical images and texts. arXiv [cs.CV]. http://arxiv.org/abs/2305.10799. Accessed 1 June 2023
158. Liu J, Wang Z, Ye Q, Chong D, Zhou P, Hua Y (2023) Qilin-Med-VL: towards Chinese large vision-language model for general healthcare. arXiv [csCV]. https://arxiv.org/abs/2310.17956. Accessed 1 Dec 2023
159. Selivanov A, Rogov OY, Chesakov D, Shelmanov A, Fedulova I, Dylov DV (2022) Medical image captioning via generative pretrained transformers. arXiv [cs.CV]. http://arxiv.org/abs/2209.13983. Accessed 1 May 2023
160. Zhu T, Wu X, Yang B, You C, Wang C, Lu L et al (2023) A large language modelling deep learning framework for the next pandemic. [cited 13 Mar 2024]. https://doi.org/10.21203/rs.3.rs-2777372/v1
161. Zhang Z, Wang B, Liang W, Li Y, Guo X, Wang G et al (2023) SAM-guided enhanced fine-grained encoding with mixed semantic learning for medical image captioning. arXiv [cs.CV]. http://arxiv.org/abs/2311.01004. Accessed 15 Nov 2023
162. Li Q, Yang X, Wang H, Wang Q, Liu L, Wang J et al (2023) From beginner to expert: modeling medical knowledge into general LLMs. arXiv [cs.CL]. http://arxiv.org/abs/2312.01040. Accessed 10 Dec 2023
163. Wang R, Yao Q, Lai H, He Z, Tao X, Jiang Z et al (2023) ECAMP: Entity-centered context-aware medical vision language pre-training. arXiv [cs.CV]. http://arxiv.org/abs/2312.13316. Accessed 20 Mar 2024
164. Wu S, Yang B, Ye Z, Wang H, Zheng H, Zhang T (2023) Improving medical report generation with adapter tuning and knowledge enhancement in vision-language foundation models. arXiv [cs.CV]. http://arxiv.org/abs/2312.03970. Accessed 20 Mar 2024
165. Zhang X, Wu C, Zhao Z, Lin W, Zhang Y, Wang Y et al (2023) PMC-VQA: visual instruction tuning for medical visual question answering. arXiv [cs.CV]. http://arxiv.org/abs/2305.10415. Accessed 1 May 2023

166. Gu Y, Yang J, Usuyama N, Li C, Zhang S, Lungren MP et al (2023) BiomedJourney: counterfactual biomedical image generation by instruction-learning from multimodal patient journeys. arXiv [cs.CV]. http://arxiv.org/abs/2310.10765. Accessed 20 Oct 2023
167. Nicolson A, Dowling J, Koopman B (2022) Improving chest X-ray report generation by leveraging warm starting. arXiv [cs.CV]. http://arxiv.org/abs/2201.09405. Accessed 1 May 2023
168. Yang X, Xu L, Li H, Zhang S (2023) ViLaM: a vision-language model with enhanced visual grounding and generalization capability. arXiv [cs.CV]. http://arxiv.org/abs/2311.12327. Accessed 20 Jan 2024
169. Kim J, Yoon S, Choi T, Sull S (2023) Unsupervised video anomaly detection based on similarity with predefined text descriptions. Sensors 23. https://doi.org/10.3390/s23146256
170. Thawakar O, Shaker AM, Mullappilly SS, Cholakkal H, Anwer R, Khan SS et al (2023) XrayGPT: chest radiographs summarization using medical vision-language models. ArXiv abs/2306.07971. https://doi.org/10.48550/arXiv.2306.07971
171. Mehboob F, Malik KM, Saudagar AKJ, Rauf A, AlTameem A (2023) Medical report generation and Chatbot for COVID_19 diagnosis using open-AI. https://doi.org/10.21203/rs.3.rs-2563448/v1
172. Yang L, Wang Z, Zhou L (2023) MedXChat: Bridging CXR modalities with a unified multimodal large model. arXiv [cs.CV]. http://arxiv.org/abs/2312.02233. Accessed 1 May 2024
173. Sai SVC, Nikhil ET, Ponraj RKK (2023) Comprehensive strategy for analyzing dementia brain images and generating textual reports through ViT,. 2023 First International Conference on Advances in Electrical, Electronics and Computational Intelligence (ICAEECI). unknown. pp 1–10
174. Kim G-Y, Oh B-D, Kim C, Kim Y-S (2023) Convolutional neural network and language model-based sequential CT image captioning for intracerebral hemorrhage. NATO Adv Sci Inst Ser E Appl Sci 13:9665
175. Lei N, Cai J, Qian Y, Zheng Z, Han C, Liu Z, Huang Q (2023) A two-stage Chinese medical video retrieval framework with LLM. In Natural Language Processing and Chinese Computing. 12th National CCF Conference, NLPCC 2023. Proceedings, Part III. Springer-Verlag, Berlin, Heidelberg. Springer Nature Switzerland, pp 211–220. https://doi.org/10.1007/978-3-031-44699-3_19
176. Chen Z, Lu Y, Wang WY (2023) Empowering psychotherapy with large language models: cognitive distortion detection through diagnosis of thought prompting. arXiv [cs.CL]. http://arxiv.org/abs/2310.07146
177. Soylemez O, Cordero P (2022) Protein language model rescue mutations highlight variant effects and structure in clinically relevant genes. arXiv [cs.LG]. http://arxiv.org/abs/2211.10000. Accessed 1 Dec 2023
178. Jo E, Epstein DA, Jung H, Kim Y-H (2023) Understanding the benefits and challenges of deploying conversational AI leveraging large language models for public health intervention. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, pp. 1–16.
179. Akilesh S, Sheik AA, Abinaya R, Dhanushkodi S, Sekar R (2023) A novel AI-based chatbot application for personalized medical diagnosis and review using large language models. 2023 International conference on research methodologies in knowledge management, artificial intelligence and telecommunication engineering (RMKMATE). IEEE, pp. 1–5
180. Chen S, Kann BH, Foote MB, Aerts HJWL, Savova GK, Mak RH, et al (2023) Use of artificial intelligence chatbots for cancer treatment information. JAMA Oncology 1459–1462
181. Chen S, Guevara M, Moningi S, Hoebers F, Elhalawani H, Kann BH et al (2023) The impact of responding to patient messages with large language model assistance. arXiv e-prints. arXiv:2310.17703
182. Laker B, Currell E (2023) ChatGPT: a novel AI assistant for healthcare messaging-a commentary on its potential in addressing patient queries and reducing clinician burnout. BMJ Lead. https://doi.org/10.1136/leader-2023-000844
183. Heston TF (2023) Safety of large language models in addressing depression. Cureus 15:e50729
184. Khalifa M, Albadawy M (2024) Using artificial intelligence in academic writing and research: an essential productivity tool. Comput Methods Programs Biomed Update 5:100145
185. Osmanovic-Thunström A, Steingrimsson S (2023) Does GPT-3 qualify as a co-author of a scientific paper publishable in peer-review journals according to the ICMJE criteria? A case study. Discover Artificial Intelligence 3:12
186. Hryciw BN, Seely AJE, Kyeremanteng K (2023) Guiding principles and proposed classification system for the responsible adoption of artificial intelligence in scientific writing in medicine. Front Artif Intell 6:1283353
187. Abu-Jeyyab M, Alrosan S, Alkhawaldeh I (2023) Harnessing large language models in medical research and scientific writing: a closer look to the future: LLMs in medical research and scientific writing. HYMR 1. https://doi.org/10.59707/hymrFBYA5348

188. Schubert MC, Wick W, Venkataramani V (2023) Performance of large language models on a neurology board-style examination. JAMA Netw Open 6:e2346721
189. Abd-Alrazaq A, AlSaad R, Alhuwail D, Ahmed A, Healy PM, Latifi S et al (2023) Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ 9:e48291
190. Reddy S (2023) Evaluating large language models for use in healthcare: a framework for translational value assessment. Inform Med Unlocked 41:101304
191. Jin H, Chen S, Wu M, Zhu KQ (2023) PsyEval: A comprehensive large language model evaluation benchmark for mental health. arXiv [cs.CL]. http://arxiv.org/abs/2311.09189. Accessed 20 Jan 2024
192. He Z, Wang Y, Yan A, Liu Y, Chang E, Gentili A et al (2023) MedEval: a multi-level, multi-task, and multi-domain medical benchmark for language model evaluation. In: Bouamor H, Pino J, Bali K (eds.). Proceedings of the 2023 conference on empirical methods in natural language processing. Association for Computational Linguistics, Singapore, pp. 8725–8744
193. Liu Z, Zhong T, Li Y, Zhang Y, Pan Y, Zhao Z et al (2023) RadLLM: a comprehensive healthcare benchmark of large language models for radiology. arXiv [cs.CL]. http://arxiv.org/abs/2307.13693
194. Lin C-Y (2004) ROUGE: a package for automatic evaluation of summaries. Text summarization branches out. Association for Computational Linguistics, Barcelona, Spain, pp. 74–81.
195. Tang L, Sun Z, Idnay B, Nestor JG, Soroush A, Elias PA et al (2023) Evaluating large language models on medical evidence summarization. NPJ Digit Med 6:158
196. Yao X, Mikhelson M, Craig Watkins S, Choi E, Thomaz E, de Barbaro K (2023) Development and evaluation of three chatbots for postpartum mood and anxiety disorders. arXiv [cs.CL]. https://doi.org/10.1145/nnnnnnn.nnnnnnn
197. Duong D, Solomon BD (2023) Analysis of large-language model versus human performance for genetics questions. medRxiv. https://doi.org/10.1101/2023.01.27.23285115
198. Fournier-Tombs E, McHardy J (2023) A medical ethics framework for conversational artificial intelligence. J Med Internet Res 25:e43068
199. Perni S, Lehmann LS, Bitterman DS (2023) Patients should be informed when AI systems are used in clinical trials. Nat Med 29:1890–1891
200. Valiña LG, Mastroleo I (2023) The ethical and scientific challenges of ChatGPT in health: utopianism, technophobia and pragmatism. https://doi.org/10.31219/osf.io/kvj45
201. Cohen IG (2023) What should ChatGPT mean for bioethics? Am J Bioeth 23:8–16
202. Li H, Moon JT, Purkayastha S, Celi LA, Trivedi H, Gichoya JW (2023) Ethics of large language models in medicine and medical research. Lancet Digit Health 5:e333–e335
203. Doyal AS, Sender D, Nanda M, Serrano RA (2023) ChatGPT and artificial intelligence in medical writing: concerns and ethical considerations. Cureus 15:e43292
204. Piñeiro-Martín A, Garcia-Mateo C, Docío-Fernández L, López Pérez M del C (2023) Ethical challenges in the development of virtual assistants powered by large language models. Preprints. https://doi.org/10.20944/preprints202306.0196.v1
205. D'Souza R, Sousa A (2023) Ethics in managing big data: ensuring privacy and data security while using ChatGPT in healthcare. Glob Bioeth Enq J. https://doi.org/10.38020/gbe.11.1.2023.1-4
206. Mazumdar H, Chakraborty C, Sathvik M, Mukhopadhyay S, Panigrahi PK (2023) GPTFX: a novel GPT-3 based framework for mental health detection and explanations. IEEE J Biomed Health Inform. https://doi.org/10.1109/JBHI.2023.3328350
207. Fu G, Zhao Q, Li J, Luo D, Song C, Zhai W et al (2023) Enhancing psychological counseling with large language model: a multifaceted decision-support system for non-professionals. arXiv [cs.AI]. http://arxiv.org/abs/2308.15192
208. He Y, Yang L, Qian C, Li T, Su Z, Zhang Q et al (2023) Conversational agent interventions for mental health problems: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 25:e43862
209. Balan R, Dobrean A, Poetar CR (2024) Use of automated conversational agents in improving young population mental health: a scoping review. NPJ Digit Med 7:75
210. Li H, Zhang R, Lee Y-C, Kraut RE, Mohr DC (2023) Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. NPJ Digit Med 6:236
211. Lv X, Zhang X, Li Y, Ding X, Lai H, Shi J (2024) Leveraging large language models for improved patient access and self-management: assessor-blinded comparison between expert- and AI-generated content. J Med Internet Res 26:e55847
212. Agbavor F, Liang H (2022) Predicting dementia from spontaneous speech using large language models. PLoS Digit Health 1:e0000168

213. Cai H, Huang X, Liu Z, Liao W, Dai H, Wu Z, Zhu D, Ren H, Li Q, Liu T, Li X (2023) Multimodal approaches for Alzheimer's detection using patients' speech and transcript. In Brain Informatics: 16th International Conference, BI 2023, Hoboken, NJ, USA, August 1–3, 2023, Proceedings. Springer-Verlag, Berlin, Heidelberg, pp. 395–406. https://doi.org/10.1007/978-3-031-43075-6_34
214. Liu X, Xu P, Wu J, Yuan J, Yang Y, Zhou Y et al (2024) Large language models and causal inference in collaboration: a comprehensive survey. arXiv [cs.CL]. http://arxiv.org/abs/2403.09606. Accessed 1 May 2024
215. Nashwan AJ (2023) Leveraging large language models to improve triage accuracy in emergency departments. J Emerg Nurs 49:651–653
216. Savage T, Nayak A, Gallo R, Rangan E, Chen JH (2024) Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. NPJ Digital Medicine 7. https://doi.org/10.1038/s41746-024-01010-1
217. Benary M, Wang XD, Schmidt M, Soll D, Hilfenhaus G, Nassir M et al (2023) Leveraging large language models for decision support in personalized oncology. JAMA Netw Open 6:e2343689
218. Gu Y, Zhang S, Usuyama N, Woldesenbet Y, Wong C, Sanapathi P et al (2023) Distilling large language models for biomedical knowledge extraction: a case study on adverse drug events. arXiv [cs.CL]. http://arxiv.org/abs/2307.06439. Accessed 5 Aug 2023
219. Schwartz IS, Link KE, Daneshjou R, Cortés-Penfield N (2024) Black box warning: large language models and the future of infectious diseases consultation. Clin Infect Dis 78:860–866
220. Ravi A, Neinstein A, Murray SG (2023) Large language models and medical education: preparing for a rapid transformation in how trainees will learn to be doctors. ATS Sch 4:282–292
221. Bak M, Chin J (2024) The potential and limitations of large language models in identification of the states of motivations for facilitating health behavior change. J Am Med Inform Assoc. https://doi.org/10.1093/jamia/ocae057
222. Lin J, Yu Y, Zhou Y, Zhou Z, Shi X (2020) How many preprints have actually been printed and why: a case study of computer science preprints on arXiv. Scientometrics 124:555–574
223. Lawson McLean A (2023) Artificial intelligence in surgical documentation: a critical review of the role of large language models. Ann Biomed Eng 51:2641–2642. Accessed 1 May 2023
224. Miao H, Li C, Wang J (2023) A future of smarter digital health empowered by generative pretrained transformer. J Med Internet Res 25:e49963
225. Sanii RY, Kasto JK, Wines WB, Mahylis JM, Muh SJ (2023) Utility of artificial intelligence in orthopedic surgery literature review: a comparative pilot study. Orthopedics 47(3):e125–e130. https://doi.org/10.3928/01477447-20231220-02
226. Liu F, Zhu T, Wu X, Yang B, You C, Wang C et al (2023) A medical multimodal large language model for future pandemics. NPJ Digit Med 6:226
227. Abi-Rafeh J, Xu HH, Kazan R, Tevlin R, Furnas H (2024) Large language models and artificial intelligence: a primer for plastic surgeons on the demonstrated and potential applications, promises, and limitations of ChatGPT. Aesthet Surg J 44:329–343
228. Dossantos J, An J, Javan R (2023) Eyes on AI: ChatGPT's transformative potential impact on ophthalmology. Cureus 15:e40765
229. Rammohan R, Joy M, Natt D, Magam SG, Patel A, Saggar T, et al (2023) S1718 understanding the landscape: the emergence of AI, ChatGPT, and Google BARD in gastroenterology. Off J Am College of Gastroenterol | ACG 118:S1281
230. Sohail SS (2023) A promising start and not a Panacea: ChatGPT's early impact and potential in medical science and biomedical engineering research. Ann Biomed Eng. https://doi.org/10.1007/s10439-023-03335-6
231. Nasarian E, Alizadehsani R, Acharya UR, Tsui K-L (2024) Designing interpretable ML system to enhance trust in healthcare: a systematic review to proposed responsible clinician-AI-collaboration framework. Inf Fusion 108:102412
232. Tanaka Y, Nakata T, Aiga K, Etani T, Muramatsu R, Katagiri S, et al (2023) Performance of generative pretrained transformer on the national medical licensing examination in Japan. medRxiv. 2023.04.17.23288603. https://doi.org/10.1101/2023.04.17.23288603
233. Liu Z, Zhong A, Li Y, Yang L, Ju C, Wu Z et al (2024) Tailoring large language models to radiology: a preliminary approach to LLM adaptation for a highly specialized domain. In Machine learning in medical imaging. Springer Nature Switzerland, pp. 464–473

234. Lun W, Luo C, Liu Y, Chen HW, Li G (2023) Diagnostic accuracy of ChatGPT and physicians in patients with abdominal pain: a cohort study. In: JMIR Preprints. [cited 13 Mar 2024]. https://preprints.jmir.org/preprint/48540. Accessed 10 Jan 2024
235. Cazzato G, Capuzzolo M, Parente P, Arezzo F, Loizzi V, Macorano E et al (2023) Chat GPT in diagnostic human pathology: will it be useful to pathologists? A preliminary review with “query session” and future perspectives. AI 4:1010–1022
236. Schukow C, Smith SC, Landgrebe E, Parasuraman S, Folaranmi OO, Paner GP et al (2024) Application of ChatGPT in routine diagnostic pathology: promises, pitfalls, and potential future directions. Adv Anat Pathol 31:15–21
237. Suppadungsuk S, Thongprayoon C, Krisanapan P, Tangpanithandee S, Garcia Valencia O, Miao J et al (2023) Examining the validity of ChatGPT in identifying relevant nephrology literature: findings and implications. J Clin Med Res 12. https://doi.org/10.3390/jcm12175550
238. Gödde D, Nöhl S, Wolf C, Rupert Y, Rimkus L, Ehlers J et al (2023) ChatGPT in medical literature – a concise review and SWOT analysis. medRxiv. 2023.05.06.23289608. https://doi.org/10.1101/2023.05.06.23289608
239. Perlis RH (2023) Research letter: application of GPT-4 to select next-step antidepressant treatment in major depression. medRxiv. https://doi.org/10.1101/2023.04.14.23288595
240. Yang K, Ji S, Zhang T, Xie Q, Kuang Z, Ananiadou S (2023) Towards interpretable mental health analysis with large language models. arXiv [cs.CL]. http://arxiv.org/abs/2304.03347. Accessed 1 May 2023
241. Lamichhane B (2023) Evaluation of ChatGPT for NLP-based mental health applications. arXiv [cs.CL]. http://arxiv.org/abs/2303.15727. Accessed 1 May 2023
242. Tripathy S, Singh R, Ray M (2023) Natural language processing for COVID-19 consulting system. Procedia Comput Sci 218:1335–1341
243. Zhang L, Tashiro S, Mukaino M, Yamada S (2023) Use of artificial intelligence large language models as a clinical tool in rehabilitation medicine: a comparative test case. J Rehabil Med 55:jrm13373. Accessed 1 May 2023
244. Ahmad MA, Yaramis I, Roy TD (2023) Creating trustworthy LLMs: dealing with hallucinations in healthcare AI. arXiv [cs.CL]. http://arxiv.org/abs/2311.01463. Accessed 1 May 2023
245. Heston TF (2023) Evaluating risk progression in mental health chatbots using escalating prompts. bioRxiv. https://doi.org/10.1101/2023.09.10.23295321
246. Chung NC, Dyer G, Brocki L (2023) Challenges of large language models for mental health counseling. arXiv [cs.CL]. http://arxiv.org/abs/2311.13857. Accessed 15 Dec 2023
247. De Choudhury M, Pendse SR, Kumar N (2023) Benefits and harms of large language models in digital mental health. arXiv [cs.CL]. http://arxiv.org/abs/2311.14693. Accessed 15 Dec 2023

Authors and Affiliations

Huizi Yu[1] · Lizhou Fan[1] · Lingyao Li[1] · Jiayan Zhou[2] · Zihui Ma[3] · Lu Xian[1] · Wenyue Hua[4] · Sijia He[1] · Mingyu Jin[4] · Yongfeng Zhang[4] · Ashvin Gandhi[5] · Xin Ma[6]

✉ Xin Ma
maxin@sdu.edu.cn

1 University of Michigan, Ann Arbor, MI, USA

2 Stanford University, Stanford, CA, USA

3 University of Maryland, College Park, MD, USA

4 Rutgers University, Newark, NJ, USA

5 University of California, Los Angeles, Los Angeles, CA, USA

6 Shandong University, Jinan, Shandong, China