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

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]

[1] University of Michigan, [2] Stanford University, [3] University of Maryland, [4] Rutgers University, [5] University of California, Los Angeles, [6] Shandong University

* Correspondence to huiziy@umich.edu

Abstract

Large Language Models (LLMs) have rapidly become important tools in Biomedical and Health Informatics (BHI), enabling new ways to analyze data, treat patients, and conduct research. This bibliometric review aims to provide a panoramic view of how LLMs have been used in BHI by examining research articles and collaboration networks from 2022 to 2023. It further explores how LLMs can improve Natural Language Processing (NLP) applications in various BHI areas like medical diagnosis, patient engagement, electronic health record management, and personalized medicine. To do this, our bibliometric review identifies key trends, maps out research networks, and highlights major developments in this fast-moving field. Lastly, it discusses the ethical concerns and practical challenges of using LLMs in BHI, such as data privacy and reliable medical recommendations. Looking ahead, we consider how LLMs could further transform biomedical research as well as healthcare delivery and patient outcomes. This bibliometric review 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.

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], [2], [3]. The advent of groundbreaking models like OpenAI's Generative Pre-trained Transformer (GPT) [4] has demonstrated the 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 can 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 applications include clinical decision support, patient engagement enhancement, and medical literature analysis [7], [8], [9]. These developments have not only augmented traditional methodologies but also paved the way for novel approaches to addressing complex challenges in the healthcare sector.

Our bibliometric review uniquely contributes to the discourse by offering a comprehensive analysis of LLM applications in BHI from 2022 to 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 on 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 bibliometric review aims to highlight key trends and gaps in current research and further point out the opportunities. Our findings aim to provide a foundation for future research, providing stakeholders with important insights to understand and contribute to this rapidly changing field. This review not only shows how LLMs could improve healthcare outcomes but also emphasizes the need to consider ethics and address practical challenges when 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, from the evolution of LLMs to their applications in BHI, as well as the synthesized knowledge of LLMs in BHI. Then the methods section outlines our approach, including data collection and description, topic classification for content analysis, and the network analysis algorithm and visualization techniques employed.

The result sections are organized in an overall-to-specific manner. First, we provide a two-fold overview using content analysis, focusing on research themes and topics, and 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*, (2) *the diverse outcomes of LLM applications*, and (3) *specific disease categories* where LLMs have shown promise. Finally, the conclusions and discussion section summarizes our key findings, addresses limitations, and provides recommendations for future work in this rapidly evolving field.

2. Backgrounds

The intersection of LLMs and BHI represents a frontier of innovation. To better understand the applications 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 The evolution of Large Language Models (LLMs)

LLMs represent a sophisticated category of language models that utilize neural networks with multi-billion parameter architectures. These models are trained on vast unlabelled 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, the introduction of Bidirectional

Encoder Representations from Transformers (BERT) in 2018 was another milestone that revolutionized the way machines understand human language [13].

Later, the evolution of LLMs witnessed a significant moment with the release of OpenAI's GPT-3 in 2020, 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 like 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 open-source and closed-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].

2.2 Applications of LLMs in the domain of BHI

Early 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 application of 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], [32], [33], [34]. The BERT model has been extensively applied in name entity recognition (NER) and relation extraction within the BHI domain [35], [36], [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 in interpreting complex patient data and suggesting medical diagnoses [41], [42], [43], [44], [45]. This capability could be 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], [49], [50], [51]. These models can 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, [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. [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, although highlighting some limitations.

2.3 Synthesized Knowledge of LLMs in BHI

Due to the aforementioned applications of LLMs in BHI, several review papers have appeared [40], [53], [54], [55], [56], [57]. For example, [40] outlined how LLM applications were developed and leveraged in clinical settings. They discussed the strengths and limitations of using LLMs for clinical and educational purposes in the biomedical context. [56] 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 applications.

Our investigation into existing review papers highlights a research gap in the literature: there remains a need for a comprehensive bibliometric review 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 bibliometric analysis of the latest LLM applications, providing a perspective on the evolving trends and challenges within this field. Secondly, the data-driven nature of our bibliometric review allows for a deeper understanding of the interdisciplinary connections within the published literature and assists in locating 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.

3. Methods

3.1 Data Collection and Analytics Workflow

Figure 1. Paper Selection and Analytic Workflow

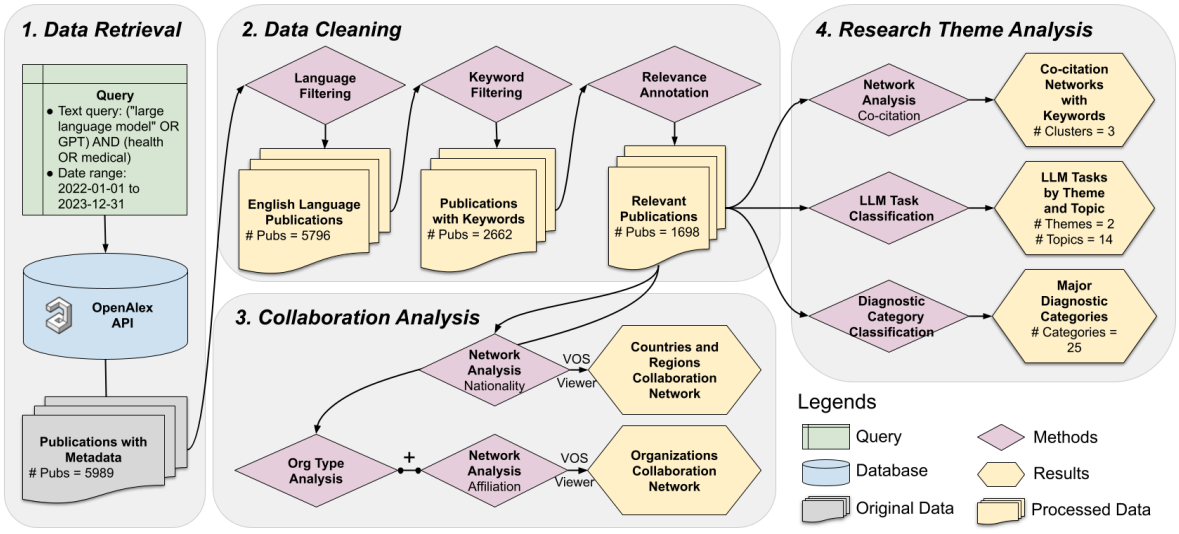


Figure 1 shows the data retrieval, cleaning, and analysis workflow. In this study, the primary data source is OpenAlex, a comprehensive database known for its extensive collection of academic publications. OpenAlex includes both published papers preprints on platforms like arXiv and medRxiv. This feature 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 employed to extract relevant data was: `("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 large language models, 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[1]. Importantly, the decision to avoid explicitly including the model names such as "llama" in the query was deliberate. While "llama" is a term associated with certain models [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 OpenAlex help page, when searching for works, the API scans through titles, abstracts, and full texts of documents. However, it employs techniques like the removal of stop words and the use of stemming (specifically, the Kstem token filter) to enhance search results. While these techniques are generally effective, they can sometimes lead to the inclusion of non-relevant documents, particularly after the stemming process. To counteract this, 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 advanced algorithmic filtering, some false positives—particularly non-health and non-medical articles—may be retained. To address this, 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 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. This model was chosen for its high performance on the Multi-Genre Natural Language Inference (MNLI) benchmark, which makes it well-suited for zero-shot learning tasks. This

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

model is especially adept at categorizing LLM research papers, which may encompass multiple topics within a single document [58], [59], [60], [61].

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. Using the `'roberta-large-mnli'` model, each title and abstract was classified into one or more of the 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[2]. To select the most likely set of predefined topics, we restrict the likelihood to be above 0.1.

3.2.2 Major Diagnostic Categories

To evaluate the medical domains and applications of LLMs, 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.

Specifically, we employed a multi-step approach to categorizing diseases mentioned in abstracts, ensuring accuracy and reliability through 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. Then to ensure the reliability and consistency of the categorization process, an intercoder reliability check is performed. The researchers then meet and resolve any discrepancies in data labeling.

3.3 Network Analysis Algorithm and Visualization

To construct the bibliometric networks, we employed the VOSviewer [62] software. These networks can include organizations, researchers, or individual publications, and are based on co-citation, bibliographic coupling, or co-authorship relations. VOSviewer utilizes a clustering algorithm based on the Visualization of Similarities (VOS) technique, designed to effectively map and visualize 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 scientific fields.

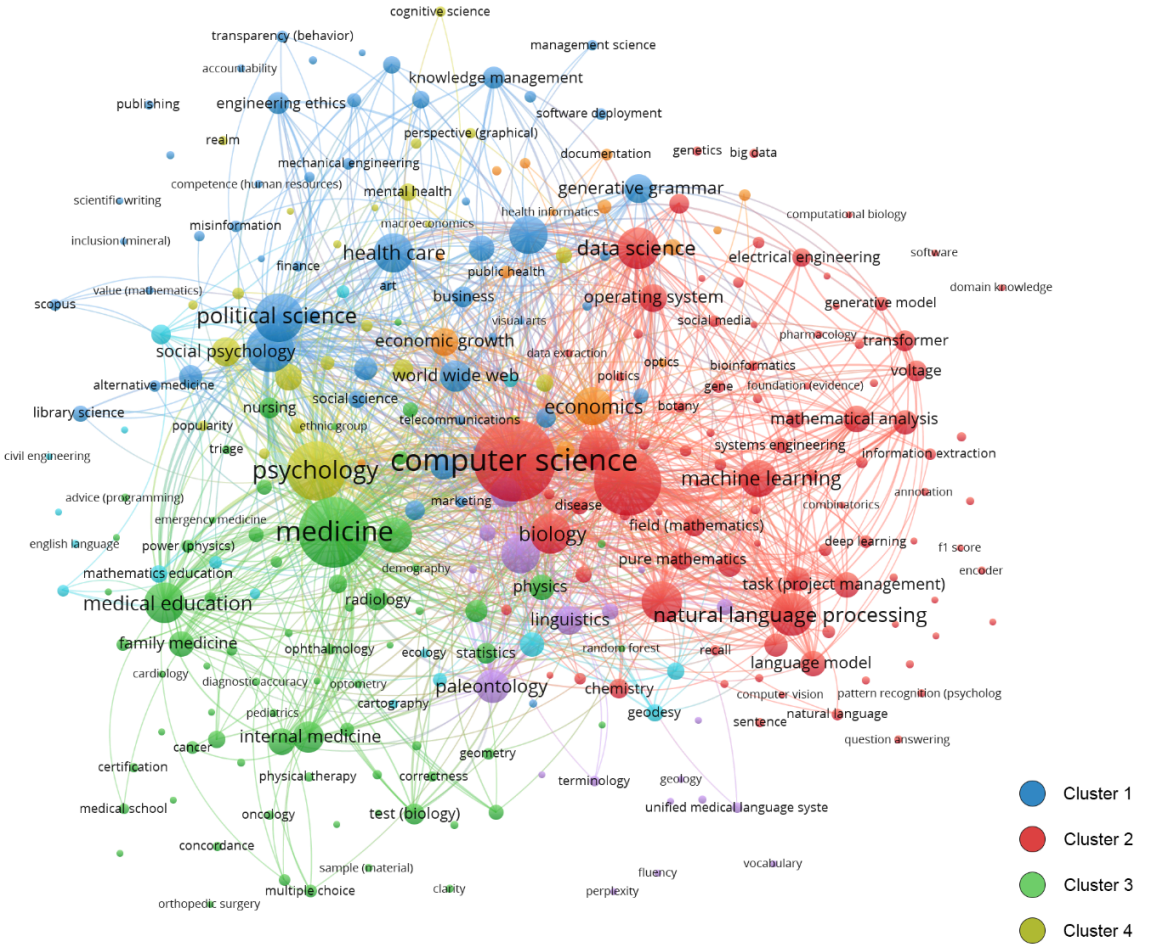
[2] In our analysis, the hypothesis is “The topic of this paper is {}.” The classification did not require any fine-tuning or training on a labeled dataset, as the model leveraged its pre-trained knowledge to make inferences about the unseen topics.

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

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 **Figure 2** indicates, the keyword co-occurrence network adeptly represents the diverse research themes and topics that converge in this multidisciplinary domain[3]. 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.

Figure 2. Keyword Co-occurrence Network



Cluster 2 in the network graph is strongly associated with the core technical disciplines of LLMs, including fields like computer science and mathematics. This cluster's prominence underscores

[3] **Appendix I** shows the top 50 keywords in the network ranked by total link strength in descending order.

a significant research focus on the theoretical and computational foundations 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 practical medical applications of LLMs and encompasses various medical specialties and fields like internal medicine and medical education. This cluster signifies the role of LLMs in clinical practice, medical training, and patient care. Cluster 1 highlights the social implications of deploying LLMs in the biomedical and health sciences, highlighting cluster includes 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 4 shows concepts at the crossroads of psychological science and its application 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 trends, and the societal impact of health interventions.

Overall, this keyword network provides an overview of the current state of research in the area of LLMs applied to 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 more depth, explaining their contributions to the field and highlighting the interconnected research efforts that can drive the continued advancement of BHI.

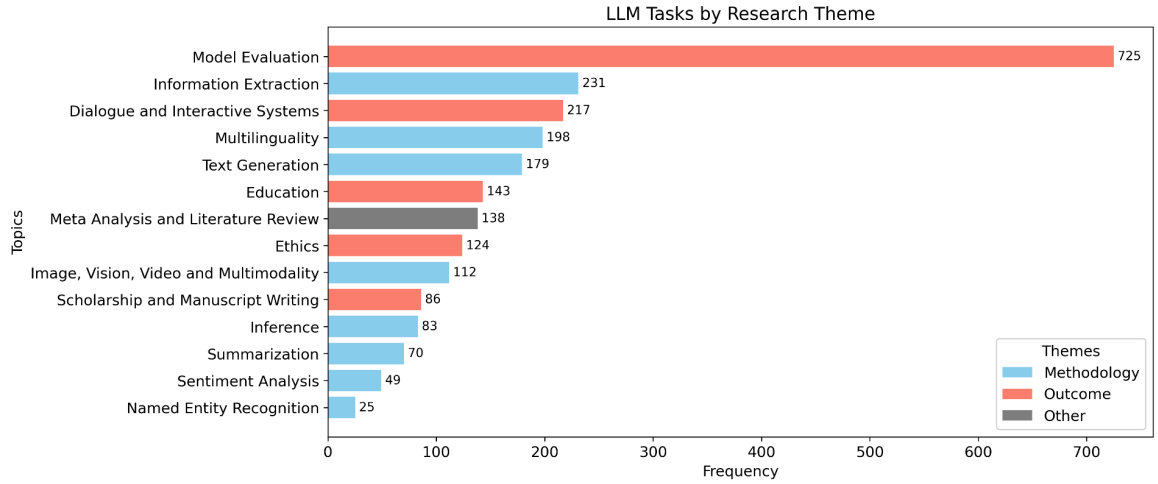
4.1.1 LLM Research Themes

The categorization of tasks associated with LLMs in the context of BHI into Methodology and Outcome is a strategic way of organizing the research focus areas[4]. **Figure 3** shows the highest number of papers centered on the Model Evaluation category under the theme of Outcome, which suggests that there is a significant emphasis on validating and testing the effectiveness and reliability of LLMs within the biomedical field. This 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 the topics of Dialogue and Interactive System, Education, and Scholarship and Manuscript Drafting, representing the substantial interest in using LLMs to distill medical information from various data sources to enhance patient interaction, medical education, and research. Lastly, 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.

In terms of Methodology, the topic of Multilinguality and the topic of Text Generation are 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, a step towards comprehensive analytics in diagnostics and patient care. Furthermore, LLM topics such as Information Extraction, Inference, Summarization, Sentiment Analysis, and Named Entity Recognition show nuanced capabilities of LLMs in processing and analyzing textual data, which can support various aspects of clinical and research activities in the biomedical sector.

[4] In **Appendix II**, we present the representative papers for each LLM task.

Figure 3: LLM Tasks by Research Theme[5]



4.1.2 Major Diagnostic Categories

Table 1 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. The exploration into skin, subcutaneous tissue, and breast disorders, including cancers, demonstrates a deep investment in leveraging LLMs for early detection and enhancing patient education. Furthermore, research on the circulatory system targets widespread conditions like 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. These investigations highlight LLMs' potential in diagnostics, enhancing treatment efficacy, patient monitoring, and even predictive modeling of disease progression. This breadth of application showcases LLMs' adaptability to meet the complex needs of different medical fields and patient care scenarios.

[5] We include "Meta Analysis and Literature Review" to better 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 III**.

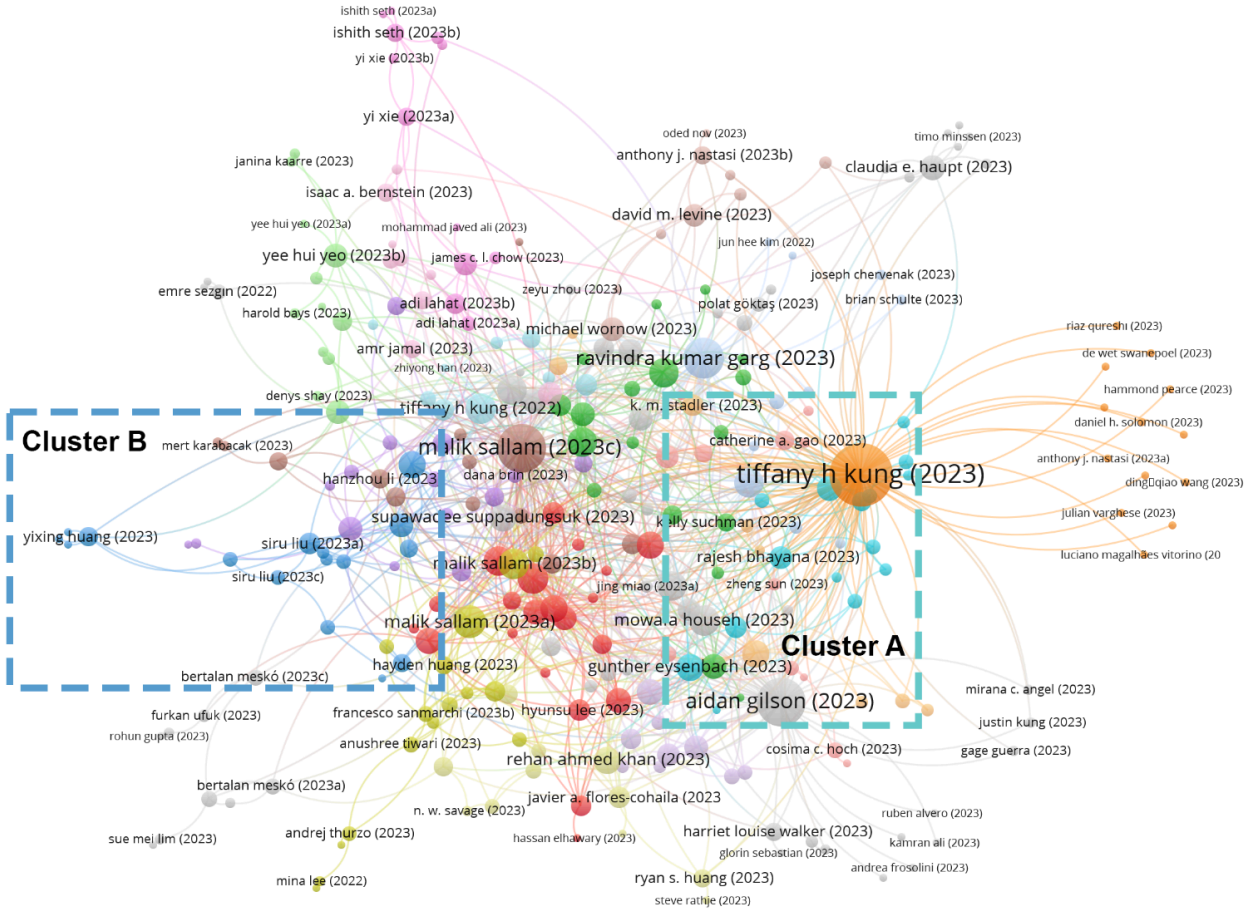
Table 1: Major Diagnostic Categories Count with Example		
Category	**Count**	**Examples**
Mental Diseases and Disorders	89	depression, post-traumatic stress disorder (ptsd), attention-deficit/hyperactivity disorder (adhd)
Nervous System	87	epilepsy, vestibular schwannoma, carpal tunnel syndrome (cts), 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 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 Influencing Health Status	9	drug-drug interaction (ddi)
Pregnancy, Childbirth, and Puerperium	7	postpartum hemorrhage (pph)
Injuries, Poison, and Toxic Effect of Drugs	5	acute organophosphate poisoning
Multiple Significant Trauma	5	joint contractures, internal organ dysfunction
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

4.2 Scholarly networks and partnerships

The visualization of the citation network in **Figure 4** offers a detailed perspective on the emergent field of LLMs in healthcare. The network includes 312 papers, each with at least five citations, which ensures that the visualization emphasizes the more influential and recognized studies within the field. This selection criteria aids in better visualization and interpretation of the network. The structure of the network indicates a close connection between studies, with certain seminal papers emerging as central nodes due to their high citations. Papers by Tiffany H. Kuang, Aidan Gilson, and Malik Sallam [50], [52], [63] are particularly prominent, suggesting their work on model performance evaluation and systematic literature reviews has been widely recognized and influential 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 highlights the focus on radiology reports [64], [65], [66], [67], whereas Cluster B is dedicated to the educational applications within medical specialties, such as dentistry [68], [69], [70].

Figure 4: Paper Co-Citation Network

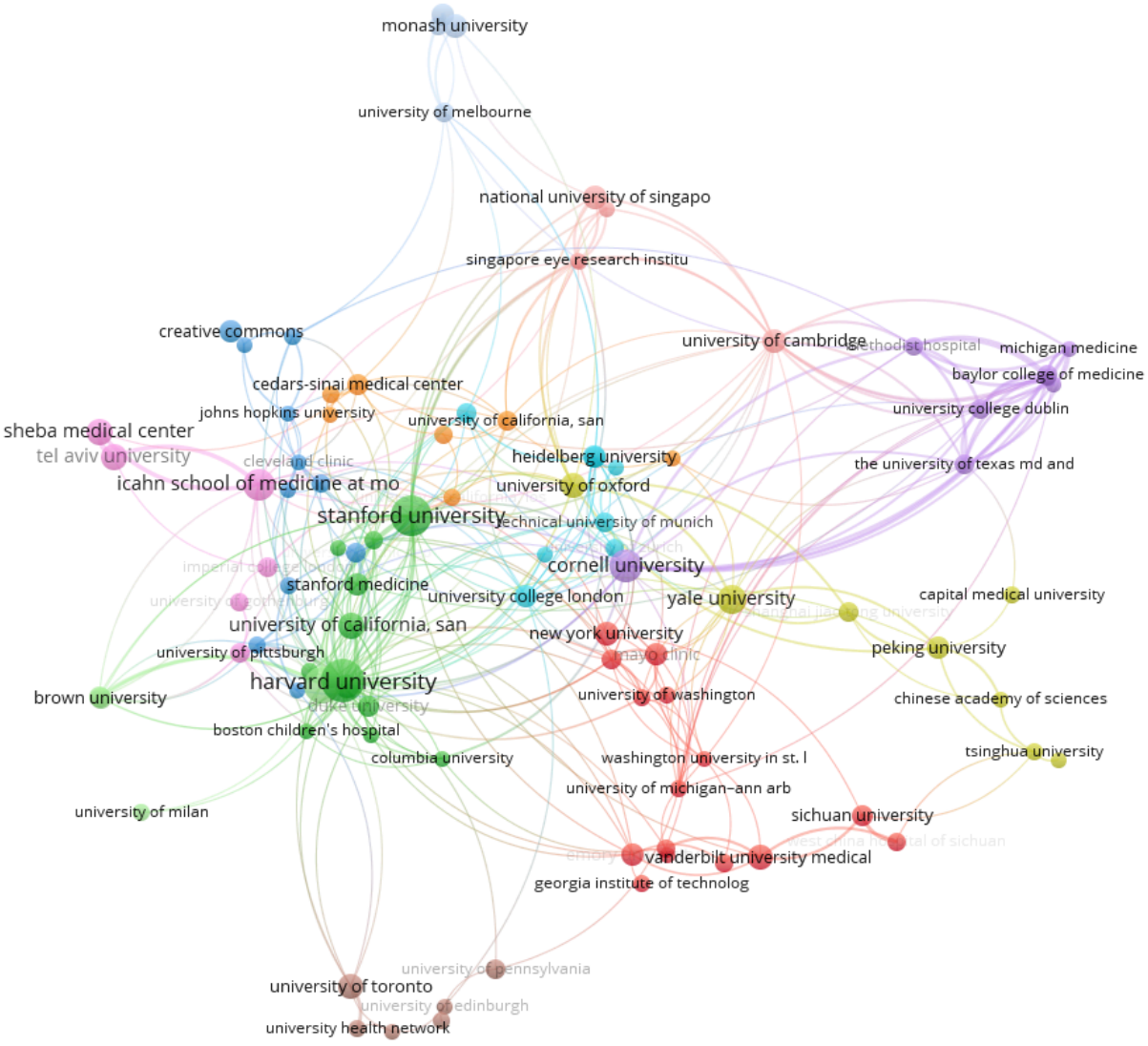


The dynamic and collaborative nature of the citation network indicates ongoing development within this field of study. New theories and methodologies are continuously being integrated. This dynamic is typical for an emerging field, where the foundational work is still being established and where there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

4.2.1 Organization Collaboration Network

The network map in **Figure 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 approach where applied 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.

Figure 5: Organization Collaboration Network



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 suggests a high

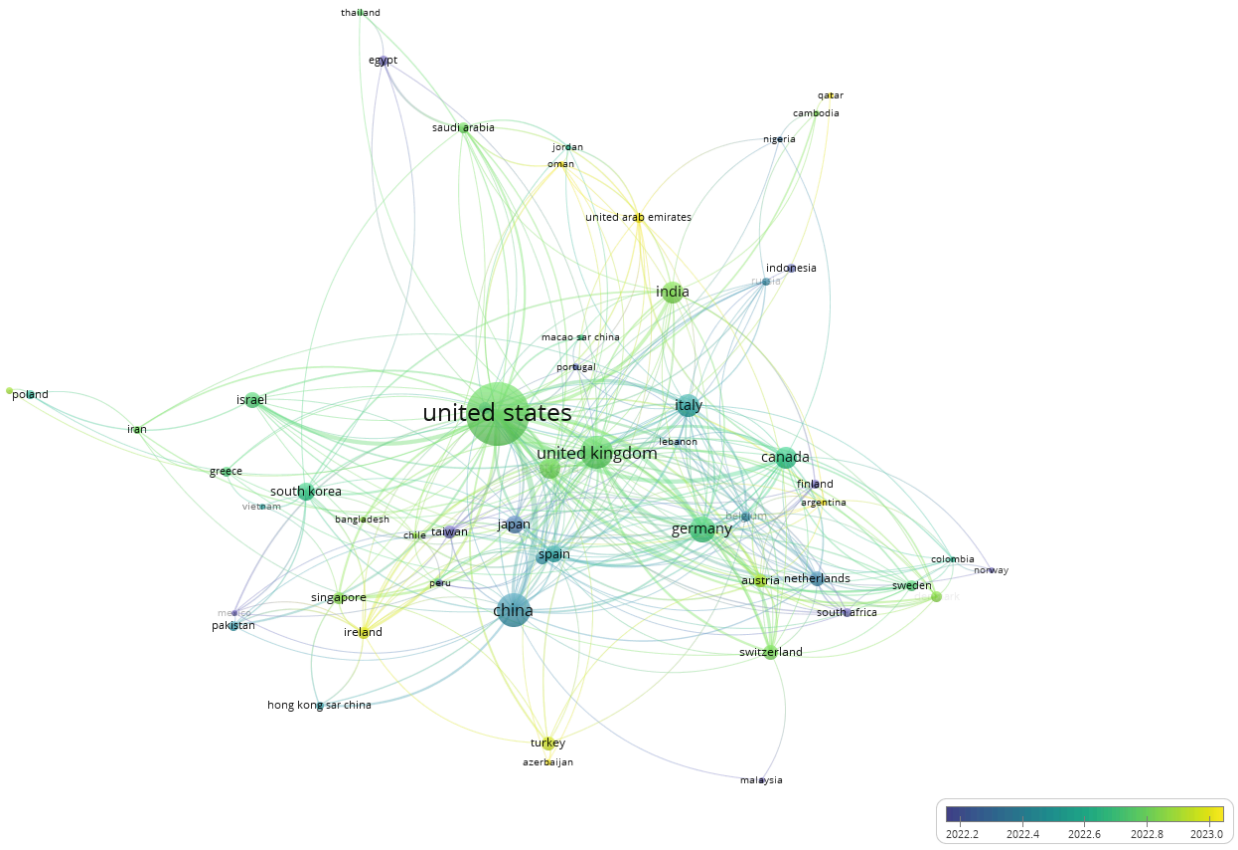
degree of collaborative engagement, which is often a reflection of the institution's broad research portfolio and its pivotal role in facilitating multidisciplinary studies.

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 Country Collaboration Network

Figure 6 provides a visual representation of a collaboration network among various 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 between 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.

Figure 6: Country Collaboration Network



Early Pioneers. Countries like Japan and the Netherlands are shown to have begun research on LLMs early on, making them pioneers in this field. This early start could suggest that these countries have established a strong foundation in LLM research, contributing significantly to the early development and understanding of these technologies.

Major Collaborators. The United States and the United Kingdom are depicted with a large total link strength, 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 in 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. This 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 and importance of LLM research.

5. Navigating the Spectrum: the Distributed LLM Methodologies in BHI

LLMs have introduced a distributed spectrum of research methodologies in BHI. A profound impact of LLMs in BHI is evident in the enhancement of NLP capabilities. These advancements have led to more accurate patient diagnosis, personalized treatment plans, and a better understanding of patient needs and responses. LLM-based tools have allowed researchers to delve deeper into large datasets, uncovering patterns and correlations that were previously undetectable. This has not only accelerated research processes but also enhanced the quality of medical diagnoses, attributable to the integration of LLMs.

5.1 Information Extraction

The utilization of LLMs is rapidly reshaping the BHI research landscape, notably in the domain of information extraction. Recent literature underscores their transformative impact in a multitude of applications. For example, some studies demonstrate LLMs' proficiency in enhancing diagnostic accuracy in hematology [71], extracting structured information from vast textual data (e.g., clinical notes, EMR notes, and radiological reports) in various languages [72], [73], [74], [75], and identifying narrative entities in the news domain [76], [77], [78]. 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 clinical information from clinical documentation, providing support for medical practitioners' decision-making [79], [80], [81], [82], [83]. The selected literature further indicated that LLMs are instrumental in extracting medication mentions [84], classifying events and contexts in clinical notes [85], and improving the understanding of medication adherence through the detection of drug discontinuation events from social media data [86]. They also excel in generating structured outputs on medications and temporal relations, further aiding in disease prediction and clinical decision support [87], [88]. These advancements, coupled with self-verification techniques [73] and the extraction of demographics and social determinants of health from Electronic Health Records (EHR) [89], [90], [91], [92], [93] illustrate LLMs' capacity to integrate and analyze healthcare data effectively.

5.2 Multilinguality

Research on multilingual LLMs in healthcare has primarily concentrated on three areas: (1) evaluating LLM performance across various languages [94], [95], [96], [97], [98], [99]; (2)

utilizing multilingual LLMs to generate datasets in different languages [100]; and (3) applying multilingual LLMs in further research [101], [102].

Studies have explored the performance discrepancies of LLMs, notably ChatGPT and GPT-4, across a diverse range of languages including English, Korean, Spanish, Chinese, Hindi, Italian, French, UK, Indian, and Iranian, revealing mixed results. While some papers find significant variance in LLMs' test accuracy on English against Korean and English against Spanish, Chinese, and Hindi [95], [96], and some find statistical significance of test accuracy on different countries [94], no statistical differences were observed when comparing the performance of ChatGPT on the English and Iranian versions of the Iranian Medical Residency Examination [97], as well as between the English and Chinese versions of the Chinese National Medical Licensing Examination [99]. In the second domain, papers have explored how to use multilingual LLM to generate multilingual datasets [100], [103], [104] for NER and Question Generation. Finally, Multilingual LLMs are also leveraged to identify personal health information in Chinese-English code-mixed clinical text and ancient Chinese medical prescriptions in Song Dynasty [101], [102]. These studies demonstrate the versatility and potential of multilingual LLMs in enhancing healthcare research, data generation, and the analysis of historical medical texts, underscoring the significant impact of language technology advancements in the medical domain.

5.3 Text Generation

Works about LLMs' writing abilities mostly focus on their text generation ability applied to scientific writing [105], [106], [107], [108] and clinical and patient-facing writing.

In the medical scientific domain, most of the research focuses on (1) the potential utility of LLMs, with particular emphasis on GPT-4, as authoring entities for diverse scientific publications, and (2) the efficacy of discerning LLM-generated texts through the means of human evaluation or AI-driven Output Detection mechanisms. Among these works, some papers focus on specific parts of a paper such as the abstract [109], [110] and the background writing. In general, they discovered that it is relatively easy for Output Detector to distinguish AI-written abstracts and original abstracts except for radiology abstract writing [111] where human reviewers and Output Detectors fail to distinguish GPT-generated abstracts from original ones. [110] claim it is hard to distinguish AI-written background from human-written background. More papers focus on overall scientific writing. In general, they find out that texts written by humans are more concrete, more diverse, and typically contained more useful information, while medical texts generated by GPT-4 paid more attention to fluency and logic and usually expressed general terminologies rather than effective information specific to the context of the problem [112], [113]. AI-written texts also face the problem of failing to include timely recent literature, the inclusion of inaccurate information, and fabricated references [114], [115], [116]. There are also works designing stronger Output Detectors [117], [118] to distinguish AI-generated text and human-generated text. In general, scholars and researchers advocate for the role of chatbots to be that of assistants rather than authors in scholarly work. Furthermore, it is widely emphasized that transparency is essential when chatbots are involved in generating academic content [119].

In both the clinical domain and patient-facing writing domain, research endeavors have been directed towards evaluating the feasibility of employing GPT-4 for the generation of case reports and responses to a variety of patient inquiries about surgical procedures and health-related matters, such as responding to postoperative questions [120], generating health message [121], aesthetic surgery advise [122], pro-vaccination message generation [123], communication in

palliative care [8], etc. Most of the studies show positive results on GPT-4's ability to generate coherent, easily comprehensible answers. [121] even shows that AI-generated messages are on par with human-generated ones in terms of sentiment, reading ease, and semantic content. and suggestions, but its accuracy, completeness, and extent of personalization [122] are still lacking. In general, they cannot replace a human agent just yet [8].

5.4 Image, Vision, Video and Multimodality

Multimodal models have shown promising potential across a wide range of applications within the healthcare sector. The body of literature concerning these models can be broadly classified into three distinct domains: (1) surveys that provide comprehensive overviews of the field; (2) proposals of new models that introduce novel methodologies or improvements; and (3) evaluations that assess the efficacy and performance of existing models. Additionally, two studies [124], [125] specifically focus on enhancing the explainability of multimodal models, aiming to make their decision-making processes more transparent and understandable.

There are more than 10 surveys on multimodal LLM on various aspects of healthcare: healthcare in general [126], [127], medical image analysis [128], [129], [130], [131], radiology [132], [133], pharmaceutical sciences [134], dentistry [69], public health informatics [135], and medical question answering [136], where all papers are having great hope that multimodal AI can be a helpful assistant for human professionals, integrating them into everyday workflow.

As healthcare and medicine is a highly specialized field, more than 30 papers have proposed domain specific multimodal models: general medical field [137], [138], [139], [140], ophthalmology [141], medical question answering [142], [143], [144], pandemic-focusing analysis [145], [146], [147], image-to-text medical report generation [147], [148] including Chest X-ray medical report generation [149], [150], chest radiogarphs summarization [151], [152], and dementia brain image report generation [153], lung cancer diagnosis [154], image generation [155], medical image captioning [156], [157], [158], medical image analysis [159], medical video retrieval [160], video anomaly detection [161]. Methods used in these papers can be crudely classified into pretrain-from-scratch [137], [138], [140], [146], [156], [158], [162] as well as finetuning based on existent pretrained or instruction-tuned models [139], [144], [148], [150], [151], [155], [161] such as vicuna, SAM, BLIP, Llama, OpenLlama, etc. [163] is the first paper that proposes MedAGI which leverages a group of small domain-specific models and automatically selects appropriate medical models by analyzing users' questions with our novel adaptive expert selection algorithm. These papers demonstrate effectiveness in various aspects of medical and healthcare domains.

Various papers have evaluated current multimodal models on various benchmarks and tasks. As GPT-4V is the most effective multimodal model for now, most of the evaluation papers focus on probing the ability of GPT-4V. Some papers test GPT-4V on standard examinations, such as USMLE questions [164], [165] where images and explanations are shown to be helpful, Japanese National Medical Licensing Examination [166] where images are helpful for analysis, Chinese Registered Dietitian Examination [167] where GPT-4V suggestions are mostly aligned with best practices, Nephrology Test Questions [168] where GPT-4V has exhibited limited accuracy and high variations in nephrology-related questions), and North American Licensing Examination [169] where GPT-4V passes the exam with 89% accuracy. Various papers have done evaluations on medical imaging analysis [170], [171], [172], [173] including general medical image analysis [173] and specialized field medical image analysis such as ophthalmic image analysis [174], pathology, dermatology, and radiology generic anomaly detection task, pandemic analysis. Several papers present GPT-4V's diagnosis performance [175], [176] such

as clinical diagnosis and Alzheimer's disease detection[177]. have shown that GPT-4V model has better-than-human abilities: GPT-4V has been tested to have significantly higher accuracy compared to the neurologists. However, negative results have also been reported: [178] have shown that although GPT-4V can identify and explain medical images, it has low diagnostic accuracy and clinical decision-making abilities.

5.5 Inferences

Other than summarizing the literature, LLMs have been developed and utilized in the association and causal inference analyses. For example, a Socratic dialogue with ChatGPT explores the causal relationship between PM2.5 and human mortality risks. A causal association was identified using LLMs by refining human reasoning patterns after performing substantial fine-tuning and acknowledging uncertainty from the confounders [179]. Moreover, an LLM-based Natural Language Inference system was developed for Clinical Trial Reports which focuses on obtaining and interpreting medical evidence [180]. In addition to the text input data, the Generative Pre-trained Transformer (GPT) was also established for medical image analysis. This analysis illustrated the possibility of using the GPT as a plug-and-play transductive inference tool with an application using a concrete case, which demonstrates its effectiveness in detecting prediction errors and improving accuracy, suggesting opportunities for broader applications in medical image analysis [159]. These studies show the extensive evolution and impactful role of LLMs in various domains, from refining reasoning patterns to enhancing the prediction of causal association, aiding in clinical trial report analysis, and advancing medical image analysis applications.

5.6 Sentiment Analysis

As advanced AI tools, LLMs have become foundational in quantifying human language sentiment. Within BHI, the refinement of sentiment analysis models has enhanced the accurate processing of extensive unstructured text by LLMs [181], [182], [183]. For instance, a model utilizing weights from a publicly available zero-shot classifier, built from the BART LLM and fine-tuned on the MNLI dataset, has been employed to evaluate linguistic nuances during psychological therapy sessions [184]. The literature indicates that LLMs are being used to analyze patient feedback, clinical notes, and public health discussions, thereby gauging public sentiment on health-related matters [185], understanding patient experiences [186], monitoring mental health trends [187], [188], and identifying cognitive distortions or suicidal tendencies [189]. Additionally, LLMs in sentiment analysis facilitate medical education [183], [190], [191], [192] by fostering interactions between medical trainees and educators, detecting thematic differences and potential biases, and revealing how feedback language may reflect varying attitudes towards learning and improvement [193]. They also contribute to sentiment analysis in research articles and medical journals, offering insights into the research community's responses to novel findings or treatments [194], [195].

5.7 Named Entity Recognition

Medical named entity recognition is one of the essential tasks working with medical data. LLMs have been applied to improve the efficiency and performance of this task, as existing supervised medical NER models necessitate human-annotated data that are often unavailable [196]. In particular, LLMs have helped identify ancient Chinese medical prescriptions from the Song Dynasty [101], [197].

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 Dialogue and Interactive Systems

The LLMs have been implemented in the newly developed chat-box as an AI-assistant for a healthcare conversion including personalized health diagnosis and intervention. 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 the dialog [198], [199], [200], [201], the estimation and the evaluation of the health status [198], and the generation of the high-quality responses [198], [199] 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 [198] 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 the health metrics and emergency alerts, and displays the reports for social works. Another newly developed application powered by the ChatGPT-3.5 model [199] allows advising the callers with up-to-date personalized medical suggestions based on the conversation. In addition, a prospective of using ChatGPT within healthcare, especially during the pandemic period, was proposed which helps with answering the patients [202]. 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 [202], [203]. More rigorous tests are needed to guarantee the safety of using the LLM in clinics [203]. Using LLM models in the dialog system [204] responsibly can bring positive changes to healthcare.

6.2 Scholarship and Manuscript Writing

As more manuscripts are published with ChatGPT as the co-author, the discussions around the use of LLMs in scientific writing have been emphasized, accompanied by a rise in various concerns. The LLMs helped to generate the summaries of current published manuscripts and to improve the writing rather than performing the scientific experiments. However, it faced challenges in accurate referencing [205], unintentional plagiarism, and data biases [206]. Establishing careful regulations and guidelines for the use of LLMs in scientific writing is crucial for assessing both effectiveness and ethical considerations [48], [207]. There is a cautious recommendation regarding its adoptions even LLMs show promise in transforming health care practices. They suggested that using LLMs might fall short of meeting authorship qualifications scientifically [50].

6.3 Education

In contrast to the investigation of LLMs in medical research, their efficacy in enhancing medical education has been assessed, yielding a potential to improve the current education and decision-making. LLMs exhibited comparable performance in comparison to human achievement without specialized training on both neurology board-style examinations [208] and the United States Medical Licensing Exam (USMLE) [52]. The emphasis was also placed on the importance of trust and explainability when implementing LLMs in medical training, aligning with expert-level knowledge [52]. The proposition remains about their potential to enhance student engagement and learning experiences [49], especially personalized curriculum development and study plans [209], albeit with considerations of ethical challenges [49], [50], algorithmic bias, and plagiarism [50], [209]. 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 [209]. These perspectives on using the LLMs highlight both the potential benefits and ethical considerations surrounding the integration of LLMs in medical education and practice.

6.4 Model Evaluation

The application of LLM in analyzing BHI data calls for new evaluation frameworks of those models' NLP performance. Research has harvested the potential of LLMs in numerous applications including serving as virtual doctors, providing mental health support [210], [211], supplying urology information for users [212], [213], and analyzing radiology [214] and EHR data [45]. Evaluating the performance of LLMs requires both automatic and human efforts. Automatic evaluation approaches often rely on metrics like ROUGE-L that compare the output of LLMs to reference outputs. Yet existing evaluation metrics, such as perplexity and the Bilingual Evaluation Understudy (BLEU) score, can't fully capture the utility of LLMs in healthcare. Frameworks like the TEHAI (Translational Evaluation of Healthcare AI) framework have been proposed by research teams to evaluate the capability, utility, and adoption of such systems in healthcare [215]. Other quantitative ways to evaluate the performance of LLMs include testing the performance of selective models comparatively [43], [212] on relevant datasets of a task, for example, MIMIC and OpenI for radiographs [216]. Existing literature has developed several benchmarks for this purpose [189], [217], [218]. As another example where LLMs' task is to provide patient consultation and diagnosis, automatic evaluation involves simulating user agents with LLMs to interact with ChatGPT and collecting such interaction data [219]. Human evaluations often rely on qualitative coding of LLM outputs or variations thereof. For example, for LLMs applied to summarize medical evidence [220], human efforts of evaluating the model-generated summaries involve 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 [221], [222].

In **Appendix IV**, 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-state contexts and center around issues of misinformation, bias and inequalities, privacy, transparency, and so on

[223], [224]. The use of LLMs as a clinical decision support tool as well as service-providing tool through chatbots can potentially harm patients when they make false recommendations, diagnoses, or prescriptions [223], [225]. Such harms, while unintended, are rooted in the corpus of training data embedded in unequal social processes [226]. Moreover, those negative consequences can also be compounded when human health professionals' judgments and decision-making processes are influenced by such biased diagnoses [223]. In particular, the use of AI-generated texts or conversational chatbots in medical contexts often involves patient-specific medical information [107], [223], [227]. This might introduce additional privacy harms to patients, since these technologies often require access to patents' sensitive information and medical record data [228]. 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. Besides, the lack of consent sharing poses another concern around data privacy and security in healthcare [229].

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

This section provides a detailed exploration of the transformative impact LLMs have across various disease categories, focusing particularly on mental health, nervous system disorders, and broad potential applications in other medical areas.

7.1 Mental Health

LLMs are poised to revolutionize mental health care 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.

Diagnosis. The integration of AI in mental health diagnostics is rapidly advancing with tools like GPTFX [230], 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, a crucial aspect for clinical applications. The study "Advancing Mental Health Diagnostics: GPT-Based Method for Depression Detection" [7] leverages transformer networks like BERT, GPT-3.5, and ChatGPT-4 to analyze clinical interviews, and show abilities to understand complex linguistic patterns and contextual cues.

These pioneering studies indicate that LLMs can be instrumental in mental health care by providing nuanced, scalable, and efficient tools for diagnosis. By analyzing language with unprecedented depth and breadth, LLMs can 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.

Intervention. The field of mental health intervention has benefited through the integration of LLMs and digital health technologies. In a paper by Zhang et al [187], 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 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" [231] 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 professional counselors.

These pioneering applications of LLMs in mental health interventions demonstrate their immense potential in both personal and community settings. By providing nuanced, user-friendly, and scalable solutions, LLMs are reshaping 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.

Promotion. Healthcare promotion, particularly in the realm of mental health and well-being, is undergoing a significant transformation with the advent of AI-based conversational agents (CAs). The integration of these advanced technologies is not only reshaping therapeutic approaches but also expanding 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" [232]. The study underscores that the quality of human-AI therapeutic relationships, content engagement, and effective communication significantly shape the user experience. This implies that while AI-based CAs can 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 by not only raising overall awareness but also by offering patient-centric recommendations. 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 within the nervous system.

The study, “Predicting dementia from spontaneous speech using large language models,” [233] 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. This work demonstrates that the nuanced understanding captured from Electronic Medical Records can significantly augment the predictive accuracy of seizure recurrence. Another paper “Multimodal Approaches for Alzheimer’s Detection Using Patients’ Speech and Transcript” [234] ventures into the domain of Alzheimer's disease (AD) detection by employing a multimodal strategy that integrates patients' speech and transcript data. This study underscores the immense potential of multimodal data in enhancing AD 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.

Together, these studies underscore the significant advancements made in the field of medical AI, 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 can 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 the medical sector holds promising potential to revolutionize disease diagnosis, prediction, and intervention, other than the mental health and brain disorders that have been extensively researched. Though their use in fields is still in the beginning stages, there exist several opportunities for LLMs to significantly enhance patient care and disease prognosis, particularly in areas like adverse drug reactions, significant trauma, and infectious diseases.

In clinical settings, LLMs can be instrumental in identifying correlations or even casual relationships by referencing vast datasets such as clinical notes, emergency care reports, and poison control center data. This could lead to the development of more effective triage systems in emergency departments and quicker, more accurate diagnoses. Ultimately, this 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 to mitigate adverse drug reactions or the risk of complications from specific toxic exposures.

In the domain of self-limiting and autoimmune disease research, LLMs can offer substantial contributions by enabling the analysis of unstructured data from pathology reports, clinical notes, and research articles. These diseases often present diagnostic challenges at the early stages due to their poorly differentiated nature. LLMs can assist in the identification of subtle linguistic patterns or terminologies that correlate with specific genetic markers or biospecimen measurements, thus aiding in the accurate and differential diagnosis. Moreover, by aggregating and synthesizing vast amounts of research and clinical data, LLMs can help identify potential therapeutic targets and inform personalized treatment plans, enhancing the precision medicine approach, especially in oncology.

In the management of infectious diseases with and 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. By analyzing patient interactions, social media, and support

group communications, LLMs could identify language indicative of treatment fatigue or social determinants affecting adherence. 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. This could lead to improved health outcomes and quality of life for individuals affected by diseases that currently have no cure.

Lastly, LLMs can also extend their contributions beyond disease settings. For example, it could be used to transform neonatal care by analyzing data from nursing notes and parental reports to identify early signs of distress for both mothers and newborns, and potential congenital anomalies or developmental issues for newborns and other neonates during the perinatal period. By understanding the nuances and complexities of neonatal care language, LLMs could predict which newborns are at risk of developing certain conditions, such as discomfort or stress, enabling preemptive care strategies and interventions. This could improve maternal and neonatal outcomes by ensuring timely and appropriate care during the critical perinatal period.

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.

8. Conclusions and Discussion

Our review has shown important trends and developments in using Large Language Models (LLMs) in biomedical and health informatics (BHI). Applying LLMs is changing the methods and outcomes in the healthcare sector. Particularly from 2022 to 2023, there has been a big increase in the number of research articles, showing rapid progress in this field. These applications include better diagnostic tools, improved patient engagement, more efficient management of Electronic Health Records (EHRs), and the emerging field of personalized medicine.

The use of LLMs in BHI has captured advanced natural language processing capabilities, greatly 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 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 bibliometric review is subject to certain limitations. First, our classification methodology, while able to conduct multi-label classification, primarily focuses on identifying the most relevant topic 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-cleaning 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 abstract.

These limitations present several opportunities for future work to refine the bibliometric 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 the 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 maintain its relevance and impact. Our future work will also explore foundation models beyond LLMs, acknowledging the growing importance of multi-modal 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.

References

[1] K. Singhal *et al.*, “Large language models encode clinical knowledge,” *Nature*, vol. 620, no. 7972, pp. 172–180, Aug. 2023, doi: 10.1038/s41586-023-06291-2.

[2] M. Karabacak and K. Margetis, “Embracing Large Language Models for Medical Applications: Opportunities and Challenges,” *Cureus*, vol. 15, no. 5, p. e39305, May 2023, doi: 10.7759/cureus.39305.

[3] J. Clusmann *et al.*, “The future landscape of large language models in medicine,” *Commun. Med.*, vol. 3, no. 1, p. 141, Oct. 2023, doi: 10.1038/s43856-023-00370-1.

[4] OpenAI, “Introducing ChatGPT.” Accessed: Mar. 12, 2024. [Online]. Available: https://openai.com/blog/chatgpt

[5] R. Tseng, S. Verberne, and P. van der Putten, “ChatGPT as a Commenter to the News: Can LLMs Generate Human-Like Opinions?,” in *Disinformation in Open Online Media*, Springer Nature Switzerland, 2023, pp. 160–174. doi: 10.1007/978-3-031-47896-3_12.

[6] Y. Ma *et al.*, “AI vs. Human -- Differentiation Analysis of Scientific Content Generation,” *arXiv [cs.CL]*, Jan. 24, 2023. [Online]. Available: http://arxiv.org/abs/2301.10416

[7] M. Danner *et al.*, “Advancing Mental Health Diagnostics: GPT-Based Method for Depression Detection,” in *2023 62nd Annual Conference of the Society of Instrument and Control Engineers (SICE)*, IEEE, Sep. 2023, pp. 1290–1296. doi: 10.23919/SICE59929.2023.10354236.

[8] R. Srivastava and S. Srivastava, “Can Artificial Intelligence aid communication? Considering the possibilities of GPT-3 in Palliative care,” *Indian J. Palliat. Care*, vol. 29, no. 4, pp. 418–425, Oct. 2023, doi: 10.25259/IJPC_155_2023.

[9] J.-L. Ghim and S. Ahn, “Transforming clinical trials: the emerging roles of large language models,” *Transl Clin Pharmacol*, vol. 31, no. 3, pp. 131–138, Sep. 2023, doi: 10.12793/tcp.2023.31.e16.

[10] Y. Shen *et al.*, “ChatGPT and Other Large Language Models Are Double-edged Swords,” *Radiology*, vol. 307, no. 2. p. e230163, Apr. 2023. doi: 10.1148/radiol.230163.

[11] W. X. Zhao *et al.*, “A Survey of Large Language Models,” *arXiv [cs.CL]*, Mar. 31, 2023. [Online]. Available: http://arxiv.org/abs/2303.18223v13

[12] A. Vaswani *et al.*, “Attention is All you Need,” *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, Accessed: Mar. 12, 2024. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *arXiv [cs.CL]*, Oct. 11, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[14] L. Floridi and M. Chiriatti, “GPT-3: Its Nature, Scope, Limits, and Consequences,” *Minds Mach.*, vol. 30, no. 4, pp. 681–694, Dec. 2020, doi: 10.1007/s11023-020-09548-1.

[15] C. Raffel *et al.*, “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer,” *arXiv [cs.LG]*, Oct. 23, 2019. [Online]. Available: http://arxiv.org/abs/1910.10683

[16] Y. Sun *et al.*, “ERNIE 2.0: A Continual Pre-Training Framework for Language Understanding,” *AAAI*, vol. 34, no. 05, pp. 8968–8975, Apr. 2020, doi: 10.1609/aaai.v34i05.6428.

[17] S. Black *et al.*, “GPT-NeoX-20B: An Open-Source Autoregressive Language Model,” *arXiv [cs.CL]*, Apr. 14, 2022. [Online]. Available: http://arxiv.org/abs/2204.06745

[18] J. Yang *et al.*, “Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond,” *arXiv [cs.CL]*, Apr. 26, 2023. [Online]. Available: http://arxiv.org/abs/2304.13712

[19] L. Fan, W. Hua, L. Li, H. Ling, and Y. Zhang, “NPHardEval: Dynamic Benchmark on

Reasoning Ability of Large Language Models via Complexity Classes," *arXiv [cs.AI]*, Dec. 22, 2023. [Online]. Available: http://arxiv.org/abs/2312.14890
[20] L. Fan *et al.*, "NPHardEval4V: A Dynamic Reasoning Benchmark of Multimodal Large Language Models," *arXiv [cs.CL]*, Mar. 04, 2024. [Online]. Available: http://arxiv.org/abs/2403.01777
[21] Anthropic, "Claude 2." Accessed: Mar. 12, 2024. [Online]. Available: https://www.anthropic.com/news/claude-2
[22] Google, "Introducing Gemini: our largest and most capable AI model." Accessed: Mar. 12, 2024. [Online]. Available: https://blog.google/technology/ai/google-gemini-ai/
[23] H. Touvron *et al.*, "Llama 2: Open Foundation and Fine-Tuned Chat Models," *arXiv [cs.CL]*, Jul. 18, 2023. [Online]. Available: http://arxiv.org/abs/2307.09288
[24] Y. Li, S. Bubeck, R. Eldan, A. Del Giorno, S. Gunasekar, and Y. T. Lee, "Textbooks Are All You Need II: phi-1.5 technical report," *arXiv [cs.CL]*, Sep. 11, 2023. [Online]. Available: http://arxiv.org/abs/2309.05463
[25] Y. Wang, S. Wu, D. Li, S. Mehrabi, and H. Liu, "A Part-Of-Speech term weighting scheme for biomedical information retrieval," *J. Biomed. Inform.*, vol. 63, pp. 379–389, Oct. 2016, doi: 10.1016/j.jbi.2016.08.026.
[26] Q.-C. Bui, P. M. A. Sloot, E. M. van Mulligen, and J. A. Kors, "A novel feature-based approach to extract drug-drug interactions from biomedical text," *Bioinformatics*, vol. 30, no. 23, pp. 3365–3371, Dec. 2014, doi: 10.1093/bioinformatics/btu557.
[27] B. Rink, S. Harabagiu, and K. Roberts, "Automatic extraction of relations between medical concepts in clinical texts," *J. Am. Med. Inform. Assoc.*, vol. 18, no. 5, pp. 594–600, Sep-Oct 2011, doi: 10.1136/amiajnl-2011-000153.
[28] M. Habibi, L. Weber, M. Neves, D. L. Wiegandt, and U. Leser, "Deep learning with word embeddings improves biomedical named entity recognition," *Bioinformatics*, vol. 33, no. 14, pp. i37–i48, Jul. 2017, doi: 10.1093/bioinformatics/btx228.
[29] Z. Jiang, L. Li, D. Huang, and L. Jin, "Training word embeddings for deep learning in biomedical text mining tasks," in *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, Nov. 2015, pp. 625–628. doi: 10.1109/BIBM.2015.7359756.
[30] Y. Peng, S. Yan, and Z. Lu, "Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets," *arXiv [cs.CL]*, Jun. 13, 2019. [Online]. Available: http://arxiv.org/abs/1906.05474
[31] L. Yao, Z. Jin, C. Mao, Y. Zhang, and Y. Luo, "Traditional Chinese medicine clinical records classification with BERT and domain specific corpora," *J. Am. Med. Inform. Assoc.*, vol. 26, no. 12, pp. 1632–1636, Dec. 2019, doi: 10.1093/jamia/ocz164.
[32] P. K. S. Prakash, S. Chilukuri, N. Ranade, and S. Viswanathan, "RareBERT: Transformer Architecture for Rare Disease Patient Identification using Administrative Claims," *AAAI*, vol. 35, no. 1, pp. 453–460, May 2021, doi: 10.1609/aaai.v35i1.16122.
[33] Y. Kawazoe, D. Shibata, E. Shinohara, E. Aramaki, and K. Ohe, "A clinical specific BERT developed using a huge Japanese clinical text corpus," *PLoS One*, vol. 16, no. 11, p. e0259763, Nov. 2021, doi: 10.1371/journal.pone.0259763.
[34] H. Yu, L. Fan, and A. J. Gilliland, "Disparities and resilience: analyzing online Health information provision, behaviors and needs of LBGTQ + elders during COVID-19," *BMC Public Health*, vol. 22, no. 1, p. 2338, Dec. 2022, doi: 10.1186/s12889-022-14783-5.
[35] K. Hakala and S. Pyysalo, "Biomedical Named Entity Recognition with Multilingual BERT," in *Proceedings of the 5th Workshop on BioNLP Open Shared Tasks*, K. Jin-Dong, N. Claire, B. Robert, and D. Louise, Eds., Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 56–61. doi: 10.18653/v1/D19-5709.
[36] C. Sun, Z. Yang, L. Wang, Y. Zhang, H. Lin, and J. Wang, "Biomedical named entity recognition using BERT in the machine reading comprehension framework," *J. Biomed. Inform.*, vol. 118, p. 103799, Jun. 2021, doi: 10.1016/j.jbi.2021.103799.

[37] A. Roy and S. Pan, “Incorporating medical knowledge in BERT for clinical relation extraction,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W.-T. Yih, Eds., Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 5357–5366. doi: 10.18653/v1/2021.emnlp-main.435.
[38] J. Lee *et al.*, “BioBERT: a pre-trained biomedical language representation model for biomedical text mining,” *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020, doi: 10.1093/bioinformatics/btz682.
[39] E. Alsentzer *et al.*, “Publicly Available Clinical BERT Embeddings,” *arXiv [cs.CL]*, Apr. 06, 2019. [Online]. Available: http://arxiv.org/abs/1904.03323
[40] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, “Large language models in medicine,” *Nat. Med.*, vol. 29, no. 8, pp. 1930–1940, Aug. 2023, doi: 10.1038/s41591-023-02448-8.
[41] T. Kuroiwa *et al.*, “The Potential of ChatGPT as a Self-Diagnostic Tool in Common Orthopedic Diseases: Exploratory Study,” *J. Med. Internet Res.*, vol. 25, p. e47621, Sep. 2023, doi: 10.2196/47621.
[42] L. Caruccio, S. Cirillo, G. Polese, G. Solimando, S. Sundaramurthy, and G. Tortora, “Can ChatGPT provide intelligent diagnoses? A comparative study between predictive models and ChatGPT to define a new medical diagnostic bot,” *Expert Syst. Appl.*, vol. 235, p. 121186, Jan. 2024, doi: 10.1016/j.eswa.2023.121186.
[43] S. Koga, N. B. Martin, and D. W. Dickson, “Evaluating the performance of large language models: ChatGPT and Google Bard in generating differential diagnoses in clinicopathological conferences of neurodegenerative disorders,” *Brain Pathol.*, p. e13207, Aug. 2023, doi: 10.1111/bpa.13207.
[44] M. Jin *et al.*, “Health-LLM: Personalized Retrieval-Augmented Disease Prediction System,” *arXiv [cs.CL]*, Feb. 01, 2024. [Online]. Available: http://arxiv.org/abs/2402.00746
[45] X. Yang *et al.*, “A large language model for electronic health records,” *NPJ Digit Med*, vol. 5, no. 1, p. 194, Dec. 2022, doi: 10.1038/s41746-022-00742-2.
[46] F. Y. Al-Ashwal, M. Zawiah, L. Gharaibeh, R. Abu-Farha, and A. N. Bitar, “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.*, vol. 15, pp. 137–147, Sep. 2023, doi: 10.2147/DHPS.S425858.
[47] Z. Gao, L. Li, S. Ma, Q. Wang, L. Hemphill, and R. Xu, “Examining the Potential of ChatGPT on Biomedical Information Retrieval: Fact-Checking Drug-Disease Associations,” *Ann. Biomed. Eng.*, Oct. 2023, doi: 10.1007/s10439-023-03385-w.
[48] G. Eysenbach, “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*, vol. 9, p. e46885, Mar. 2023, doi: 10.2196/46885.
[49] H. Lee, “The rise of ChatGPT: Exploring its potential in medical education,” *Anat. Sci. Educ.*, Mar. 2023, doi: 10.1002/ase.2270.
[50] M. Sallam, “ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns,” *Healthcare (Basel)*, vol. 11, no. 6, Mar. 2023, doi: 10.3390/healthcare11060887.
[51] L. Li, Z. Ma, L. Fan, S. Lee, H. Yu, and L. Hemphill, “ChatGPT in education: a discourse analysis of worries and concerns on social media,” *Education and Information Technologies*, Oct. 2023, doi: 10.1007/s10639-023-12256-9.
[52] T. H. Kung *et al.*, “Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models,” *PLOS Digit Health*, vol. 2, no. 2, p. e0000198, Feb. 2023, doi: 10.1371/journal.pdig.0000198.
[53] J. Li, A. Dada, B. Puladi, J. Kleesiek, and J. Egger, “ChatGPT in healthcare: A taxonomy and systematic review,” *Comput. Methods Programs Biomed.*, vol. 245, p. 108013, Mar.

2024, doi: 10.1016/j.cmpb.2024.108013.
[54] S. Thapa and S. Adhikari, “ChatGPT, Bard, and Large Language Models for Biomedical Research: Opportunities and Pitfalls,” *Ann. Biomed. Eng.*, vol. 51, no. 12, pp. 2647–2651, Dec. 2023, doi: 10.1007/s10439-023-03284-0.
[55] Y. Chang *et al.*, “A Survey on Evaluation of Large Language Models,” *ACM Trans. Intell. Syst. Technol.*, Jan. 2024, doi: 10.1145/3641289.
[56] S. Tian *et al.*, “Opportunities and challenges for ChatGPT and large language models in biomedicine and health,” *Brief. Bioinform.*, vol. 25, no. 1, Nov. 2023, doi: 10.1093/bib/bbad493.
[57] L. Fan, L. Li, Z. Ma, S. Lee, H. Yu, and L. Hemphill, “A Bibliometric Review of Large Language Models Research from 2017 to 2023,” *arXiv [cs.DL]*, Apr. 03, 2023. [Online]. Available: http://arxiv.org/abs/2304.02020
[58] Z. Guo, L. Zhu, and L. Han, “Research on Short Text Classification Based on RoBERTa-TextRCNN,” in *2021 International Conference on Computer Information Science and Artificial Intelligence (CISAI)*, IEEE, Sep. 2021, pp. 845–849. doi: 10.1109/CISAI54367.2021.00171.
[59] Z. Xu, “RoBERTa-wwm-ext Fine-Tuning for Chinese Text Classification,” *arXiv [cs.CL]*, Feb. 24, 2021. [Online]. Available: http://arxiv.org/abs/2103.00492
[60] W.-C. Chang, H.-F. Yu, K. Zhong, Y. Yang, and I. S. Dhillon, “Taming Pretrained Transformers for Extreme Multi-label Text Classification,” in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, in KDD ’20. New York, NY, USA: Association for Computing Machinery, Aug. 2020, pp. 3163–3171. doi: 10.1145/3394486.3403368.
[61] W. Yin, J. Hay, and D. Roth, “Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach,” *arXiv [cs.CL]*, Aug. 31, 2019. [Online]. Available: http://arxiv.org/abs/1909.00161
[62] VOSviewer, “VOSviewer - Visualizing scientific landscapes,” VOSviewer. Accessed: Mar. 12, 2024. [Online]. Available: https://www.vosviewer.com/
[63] A. Gilson *et al.*, “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*, vol. 9, p. e45312, Feb. 2023, doi: 10.2196/45312.
[64] L. C. Adams *et al.*, “Leveraging GPT-4 for Post Hoc Transformation of Free-text Radiology Reports into Structured Reporting: A Multilingual Feasibility Study,” *Radiology*, vol. 307, no. 4, p. e230725, May 2023, doi: 10.1148/radiol.230725.
[65] H. L. Haver, E. B. Ambinder, M. Bahl, E. T. Oluyemi, J. Jeudy, and P. H. Yi, “Appropriateness of Breast Cancer Prevention and Screening Recommendations Provided by ChatGPT,” *Radiology*, vol. 307, no. 4, p. e230424, May 2023, doi: 10.1148/radiol.230424.
[66] Z. Sun *et al.*, “Evaluating GPT4 on Impressions Generation in Radiology Reports,” *Radiology*, vol. 307, no. 5, p. e231259, Jun. 2023, doi: 10.1148/radiol.231259.
[67] R. Bhayana, S. Krishna, and R. R. Bleakney, “Performance of ChatGPT on a Radiology Board-style Examination: Insights into Current Strengths and Limitations,” *Radiology*, vol. 307, no. 5, p. e230582, Jun. 2023, doi: 10.1148/radiol.230582.
[68] A. Thurzo, M. Strunga, R. Urban, J. Surovková, and K. I. Afrashtehfar, “Impact of Artificial Intelligence on Dental Education: A Review and Guide for Curriculum Update,” *Education Sciences*, vol. 13, no. 2, p. 150, Jan. 2023, doi: 10.3390/educsci13020150.
[69] H. Huang *et al.*, “ChatGPT for shaping the future of dentistry: the potential of multi-modal large language model,” *Int. J. Oral Sci.*, vol. 15, no. 1, p. 29, Jul. 2023, doi: 10.1038/s41368-023-00239-y.
[70] J. Surovková, S. Haluzová, M. Strunga, R. Urban, M. Lifková, and A. Thurzo, “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.*, vol. 13, no. 8, p. 5212, Apr. 2023, doi: 10.3390/app13085212.
[71] M. R. Cervera *et al.*, "Assessment of Artificial Intelligence Language Models and Information Retrieval Strategies for QA in Hematology," *Blood*, vol. 142, no. Supplement 1, pp. 7175–7175, Nov. 2023, doi: 10.1182/blood-2023-178528.
[72] M. Agrawal, S. Hegselmann, H. Lang, Y. Kim, and D. Sontag, "Large language models are few-shot clinical information extractors," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds., Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 1998–2022. doi: 10.18653/v1/2022.emnlp-main.130.
[73] Z. Gero *et al.*, "Self-Verification Improves Few-Shot Clinical Information Extraction," *arXiv [cs.CL]*, May 30, 2023. [Online]. Available: http://arxiv.org/abs/2306.00024
[74] A. Goel *et al.*, "LLMs Accelerate Annotation for Medical Information Extraction," *arXiv [cs.CL]*, Dec. 04, 2023. [Online]. Available: http://arxiv.org/abs/2312.02296
[75] D. Hu, B. Liu, X. Zhu, X. Lu, and N. Wu, "Zero-shot information extraction from radiological reports using ChatGPT," *Int. J. Med. Inform.*, vol. 183, p. 105321, Mar. 2024, doi: 10.1016/j.ijmedinf.2023.105321.
[76] J. Chen, P. Chen, and X. Wu, "Generating Chinese Event Extraction Method Based on ChatGPT and Prompt Learning," *NATO Adv. Sci. Inst. Ser. E Appl. Sci.*, vol. 13, no. 17, p. 9500, Aug. 2023, doi: 10.3390/app13179500.
[77] L. Wang, Y. Ma, W. Bi, H. Lv, and Y. Li, "An Entity Extraction pipeline for Medical Text Records Utilizing Large Language Models: An Analytical Study," JMIR Preprints. Accessed: Mar. 12, 2024. [Online]. Available: https://preprints.jmir.org/preprint/54580
[78] H. Sousa, N. Guimarães, A. Jorge, and R. Campos, "GPT Struct Me: Probing GPT Models on Narrative Entity Extraction," *arXiv [cs.CL]*, Nov. 24, 2023. [Online]. Available: http://arxiv.org/abs/2311.14583
[79] S. Mohammed, J. Fiaidhi, and H. Shaik, "Empowering Transformers for Evidence-Based Medicine," *medRxiv*, p. 2023.12.25.23300520, Dec. 28, 2023. doi: 10.1101/2023.12.25.23300520.
[80] I. Goenaga, A. Atutxa, K. Gojenola, M. Oronoz, and R. Agerri, "Explanatory Argument Extraction of Correct Answers in Resident Medical Exams," *arXiv [cs.CL]*, Dec. 01, 2023. [Online]. Available: http://arxiv.org/abs/2312.00567
[81] N. Jethani *et al.*, "Evaluating ChatGPT in Information Extraction: A Case Study of Extracting Cognitive Exam Dates and Scores," *medRxiv*, p. 2023.07.10.23292373, Jul. 12, 2023. doi: 10.1101/2023.07.10.23292373.
[82] D. S. Bitterman *et al.*, "An End-to-End Natural Language Processing System for Automatically Extracting Radiation Therapy Events From Clinical Texts," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 117, no. 1, pp. 262–273, Sep. 2023, doi: 10.1016/j.ijrobp.2023.03.055.
[83] S. Chen *et al.*, "Natural Language Processing to Automatically Extract the Presence and Severity of Esophagitis in Notes of Patients Undergoing Radiotherapy," *JCO Clin Cancer Inform*, vol. 7, p. e2300048, Jul. 2023, doi: 10.1200/CCI.23.00048.
[84] D. Mahajan, J. J. Liang, C.-H. Tsou, and Ö. Uzuner, "Overview of the 2022 n2c2 shared task on contextualized medication event extraction in clinical notes," *J. Biomed. Inform.*, vol. 144, p. 104432, Aug. 2023, doi: 10.1016/j.jbi.2023.104432.
[85] A. Chen, Z. Yu, X. Yang, Y. Guo, J. Bian, and Y. Wu, "Contextualized Medication Information Extraction Using Transformer-based Deep Learning Architectures," *arXiv [cs.CL]*, Mar. 14, 2023. [Online]. Available: http://arxiv.org/abs/2303.08259
[86] W. Trevena *et al.*, "Utilizing Open-Source Language Models and ChatGPT for Zero-Shot Identification of Drug Discontinuation Events in Online Forums: Development and Validation

Study," JMIR Preprints. Accessed: Mar. 12, 2024. [Online]. Available: https://preprints.jmir.org/preprint/54601
[87] H. Tu, L. Han, and G. Nenadic, "Extraction of Medication and Temporal Relation from Clinical Text using Neural Language Models," *arXiv [cs.CL]*, Oct. 03, 2023. [Online]. Available: http://arxiv.org/abs/2310.02229
[88] W. Abu-Ashour, S. Emil, and D. Poenaru, "Using Artificial Intelligence To Label Free-Text Operative And Ultrasound Reports For Grading Pediatric Appendicitis," *medRxiv*, p. 2023.08.30.23294850, Sep. 01, 2023. doi: 10.1101/2023.08.30.23294850.
[89] G. K. Ramachandran *et al.*, "Prompt-based Extraction of Social Determinants of Health Using Few-shot Learning," *arXiv [cs.CL]*, Jun. 12, 2023. [Online]. Available: http://arxiv.org/abs/2306.07170
[90] N. Bhate, A. Mittal, Z. He, and X. Luo, "Zero-shot Learning with Minimum Instruction to Extract Social Determinants and Family History from Clinical Notes using GPT Model," *arXiv [cs.CL]*, Sep. 11, 2023. [Online]. Available: http://arxiv.org/abs/2309.05475
[91] C. Chakraborty, M. Bhattacharya, and S.-S. Lee, "Need an AI-Enabled, Next-Generation, Advanced ChatGPT or Large Language Models (LLMs) for Error-Free and Accurate Medical Information," *Ann. Biomed. Eng.*, vol. 52, no. 2, pp. 134–135, Feb. 2024, doi: 10.1007/s10439-023-03297-9.
[92] M. Guevara *et al.*, "Large language models to identify social determinants of health in electronic health records," *NPJ Digit Med*, vol. 7, no. 1, p. 6, Jan. 2024, doi: 10.1038/s41746-023-00970-0.
[93] A. Derton *et al.*, "Natural Language Processing Methods to Empirically Explore Social Contexts and Needs in Cancer Patient Notes," *JCO Clin Cancer Inform*, vol. 7, p. e2200196, May 2023, doi: 10.1200/CCI.22.00196.
[94] M. Alfertshofer *et al.*, "Sailing the Seven Seas: A Multinational Comparison of ChatGPT's Performance on Medical Licensing Examinations," *Ann. Biomed. Eng.*, Aug. 2023, doi: 10.1007/s10439-023-03338-3.
[95] H. Zong, J. Li, E. Wu, R. Wu, J. Lu, and B. Shen, "Performance of ChatGPT on Chinese national medical licensing examinations: A five-year examination evaluation study for physicians, pharmacists and nurses," *bioRxiv*, Jul. 09, 2023. doi: 10.1101/2023.07.09.23292415.
[96] Y. Jin, M. Chandra, G. Verma, Y. Hu, M. De Choudhury, and S. Kumar, "Better to Ask in English: Cross-Lingual Evaluation of Large Language Models for Healthcare Queries," *arXiv [cs.CL]*, Oct. 19, 2023. [Online]. Available: http://arxiv.org/abs/2310.13132
[97] H. Khorshidi *et al.*, "Application of ChatGPT in multilingual medical education: How does ChatGPT fare in 2023's Iranian residency entrance examination," *Informatics in Medicine Unlocked*, vol. 41, p. 101314, Jan. 2023, doi: 10.1016/j.imu.2023.101314.
[98] Y. H. Yeo *et al.*, "GPT-4 outperforms ChatGPT in answering non-English questions related to cirrhosis," *bioRxiv*, May 05, 2023. doi: 10.1101/2023.05.04.23289482.
[99] C. Fang *et al.*, "How does ChatGPT4 preform on Non-English National Medical Licensing Examination? An Evaluation in Chinese Language," *bioRxiv*, May 05, 2023. doi: 10.1101/2023.05.03.23289443.
[100] R. Ackerman and R. Balyan, "Automatic Multilingual Question Generation for Health Data Using LLMs," in *AI-generated Content*, Springer Nature Singapore, 2024, pp. 1–11. doi: 10.1007/978-981-99-7587-7_1.
[101] M. Li and X. Zheng, "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*, vol. 11, pp. 131453–131464, 2023, doi: 10.1109/ACCESS.2023.3330212.
[102] Y.-Q. Lee *et al.*, "Unlocking the Secrets Behind Advanced Artificial Intelligence Language Models in Deidentifying Chinese-English Mixed Clinical Text: Development and Validation

Study," *J. Med. Internet Res.*, vol. 26, p. e48443, Jan. 2024, doi: 10.2196/48443.
[103] J. Frei and F. Kramer, "Annotated dataset creation through large language models for non-english medical NLP," *J. Biomed. Inform.*, vol. 145, p. 104478, Sep. 2023, doi: 10.1016/j.jbi.2023.104478.
[104] X. Fontaine, F. Gaschi, P. Rastin, and Y. Toussaint, "Multilingual Clinical NER: Translation or Cross-lingual Transfer?," *arXiv [cs.CL]*, Jun. 07, 2023. [Online]. Available: http://arxiv.org/abs/2306.04384
[105] M. Buholayka, R. Zouabi, and A. Tadinada, "The Readiness of ChatGPT to Write Scientific Case Reports Independently: A Comparative Evaluation Between Human and Artificial Intelligence," *Cureus*, vol. 15, no. 5, p. e39386, May 2023, doi: 10.7759/cureus.39386.
[106] H. Liu, M. Azam, S. Bin Naeem, and A. Faiola, "An overview of the capabilities of ChatGPT for medical writing and its implications for academic integrity," *Health Info. Libr. J.*, vol. 40, no. 4, pp. 440–446, Dec. 2023, doi: 10.1111/hir.12509.
[107] A. S. Doyal, D. Sender, M. Nanda, and R. A. Serrano, "ChatGPT and Artificial Intelligence in Medical Writing: Concerns and Ethical Considerations," *Cureus*, vol. 15, no. 8, p. e43292, Aug. 2023, doi: 10.7759/cureus.43292.
[108] D. Najafali, C. Hinson, J. M. Camacho, L. G. Galbraith, R. Gupta, and C. M. Reid, "Can Chatbots Assist With Grant Writing in Plastic Surgery? Utilizing ChatGPT to Start an R01 Grant," *Aesthet. Surg. J.*, vol. 43, no. 8, pp. NP663–NP665, Jul. 2023, doi: 10.1093/asj/sjad116.
[109] C. A. Gao *et al.*, "Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers," *NPJ Digit Med*, vol. 6, no. 1, p. 75, Apr. 2023, doi: 10.1038/s41746-023-00819-6.
[110] I. A. Huespe *et al.*, "Clinical Research With Large Language Models Generated Writing-Clinical Research with AI-assisted Writing (CRAW) Study," *Crit Care Explor*, vol. 5, no. 10, p. e0975, Oct. 2023, doi: 10.1097/CCE.0000000000000975.
[111] F. Ufuk, H. Peker, E. Sagtas, and A. B. Yagci, "Distinguishing GPT-4-generated Radiology Abstracts from Original Abstracts: Performance of Blinded Human Observers and AI Content Detector," *medRxiv*, p. 2023.04.28.23289283, May 03, 2023. doi: 10.1101/2023.04.28.23289283.
[112] W. Liao *et al.*, "Differentiate ChatGPT-generated and Human-written Medical Texts," *arXiv [cs.CL]*, Apr. 23, 2023. [Online]. Available: http://arxiv.org/abs/2304.11567
[113] J. T. H. Wang, "Is the laboratory report dead? AI and ChatGPT," *Microbiol. Aust.*, pp. 144–148, Apr. 2023, doi: 10.1071/MA23042.
[114] O. Abuyaman, "Strengths and Weaknesses of ChatGPT Models for Scientific Writing About Medical Vitamin B12: Mixed Methods Study," *JMIR Form Res*, vol. 7, p. e49459, Nov. 2023, doi: 10.2196/49459.
[115] T. R. Grigio, H. Timmerman, and A. P. Wolff, "ChatGPT in anaesthesia research: risk of fabrication in literature searches," *Br. J. Anaesth.*, vol. 131, no. 1, pp. e29–e30, Jul. 2023, doi: 10.1016/j.bja.2023.04.009.
[116] M. Májovský, M. Černý, M. Kasal, M. Komarc, and D. Netuka, "Artificial Intelligence Can Generate Fraudulent but Authentic-Looking Scientific Medical Articles: Pandora's Box Has Been Opened," *J. Med. Internet Res.*, vol. 25, p. e46924, May 2023, doi: 10.2196/46924.
[117] A. A. Hamed and X. Wu, "Detection of ChatGPT Fake Science with the xFakeBibs Learning Algorithm," *arXiv [cs.CL]*, Aug. 15, 2023. [Online]. Available: http://arxiv.org/abs/2308.11767
[118] I. Katib, F. Y. Assiri, H. A. Abdushkour, D. Hamed, and M. Ragab, "Differentiating Chat Generative Pretrained Transformer from Humans: Detecting ChatGPT-Generated Text and Human Text Using Machine Learning," *Sci. China Ser. A Math.*, vol. 11, no. 15, p. 3400, Aug. 2023, doi: 10.3390/math11153400.

[119] T. I. Leung, T. de Azevedo Cardoso, A. Mavragani, and G. Eysenbach, “Best Practices for Using AI Tools as an Author, Peer Reviewer, or Editor,” *J. Med. Internet Res.*, vol. 25, p. e51584, Aug. 2023, doi: 10.2196/51584.
[120] E. Waisberg, J. Ong, M. Masalkhi, N. Zaman, and A. Tavakkoli, “Chat Generative Pretrained Transformer to optimize accessibility for cataract surgery postoperative management,” *The Pan-American Journal of Ophthalmology*, vol. 5, no. 1, Nov. 2023, doi: 10.4103/pajo.pajo_51_23.
[121] S. Lim and R. Schmälzle, “Artificial intelligence for health message generation: an empirical study using a large language model (LLM) and prompt engineering,” *Frontiers in Communication*, vol. 8, 2023, doi: 10.3389/fcomm.2023.1129082.
[122] Y. Xie, I. Seth, D. J. Hunter-Smith, W. M. Rozen, R. Ross, and M. Lee, “Aesthetic Surgery Advice and Counseling from Artificial Intelligence: A Rhinoplasty Consultation with ChatGPT,” *Aesthetic Plast. Surg.*, vol. 47, no. 5, pp. 1985–1993, Oct. 2023, doi: 10.1007/s00266-023-03338-7.
[123] E. Karinshak, S. X. Liu, J. S. Park, and J. T. Hancock, “Working With AI to Persuade: Examining a Large Language Model’s Ability to Generate Pro-Vaccination Messages,” *Proc. ACM Hum.-Comput. Interact.*, vol. 7, no. CSCW1, pp. 1–29, Apr. 2023, doi: 10.1145/3579592.
[124] J. Liu, T. Hu, Y. Zhang, X. Gai, Y. Feng, and Z. Liu, “A ChatGPT Aided Explainable Framework for Zero-Shot Medical Image Diagnosis,” *arXiv [eess.IV]*, Jul. 05, 2023. [Online]. Available: http://arxiv.org/abs/2307.01981
[125] H. Ying, Z. Zhao, Y. Zhao, S. Zeng, and S. Yu, “CoRTEx: Contrastive Learning for Representing Terms via Explanations with Applications on Constructing Biomedical Knowledge Graphs,” *arXiv [cs.CL]*, Dec. 13, 2023. [Online]. Available: http://arxiv.org/abs/2312.08036
[126] B. Meskó, “The Impact of Multimodal Large Language Models on Health Care’s Future,” *J. Med. Internet Res.*, vol. 25, p. e52865, Nov. 2023, doi: 10.2196/52865.
[127] R. Temsah, I. Altamimi, K. Alhasan, M.-H. Temsah, and A. Jamal, “Healthcare’s New Horizon With ChatGPT's Voice and Vision Capabilities: A Leap Beyond Text,” *Cureus*, vol. 15, no. 10, p. e47469, Oct. 2023, doi: 10.7759/cureus.47469.
[128] E. Waisberg *et al.*, “GPT-4 and medical image analysis: strengths, weaknesses and future directions,” *J. Med. Artif. Intell.*, vol. 6, pp. 29–29, Dec. 2023, doi: 10.21037/jmai-23-94.
[129] X. Li *et al.*, “Artificial General Intelligence for Medical Imaging,” *arXiv [cs.AI]*, Jun. 08, 2023. [Online]. Available: http://arxiv.org/abs/2306.05480
[130] M. Hu, S. Pan, Y. Li, and X. Yang, “Advancing Medical Imaging with Language Models: A Journey from N-grams to ChatGPT,” *arXiv [cs.CV]*, Apr. 11, 2023. [Online]. Available: http://arxiv.org/abs/2304.04920
[131] Z. Liu *et al.*, “Holistic Evaluation of GPT-4V for Biomedical Imaging,” Nov. 2023, Accessed: Mar. 13, 2024. [Online]. Available: https://paperswithcode.com/paper/holistic-evaluation-of-gpt-4v-for-biomedical
[132] J. Z. T. Sim, K. N. Bhanu Prakash, W. M. Huang, and C. H. Tan, “Harnessing artificial intelligence in radiology to augment population health,” *Front Med Technol*, vol. 5, p. 1281500, Nov. 2023, doi: 10.3389/fmedt.2023.1281500.
[133] H. Daungsupawong and V. Wiwanitkit, “Transforming Radiology With AI Visual Chatbot,” *J. Am. Coll. Radiol.*, vol. 21, no. 1, p. 3, Jan. 2024, doi: 10.1016/j.jacr.2023.10.022.
[134] N. M. Davies, “Adapting artificial intelligence into the evolution of pharmaceutical sciences and publishing: Technological darwinism,” *J. Pharm. Pharm. Sci.*, vol. 26, p. 11349, Mar. 2023, doi: 10.3389/jpps.2023.11349.
[135] A. Awan, A. Gonzalez, and M. Sharma, “A Neoteric Approach toward Social Media in Public Health Informatics: A Narrative Review of Current Trends and Future Directions,”

Dec. 26, 2023. doi: 10.20944/preprints202312.2102.v1.
[136] H. Demirhan and W. Zadrozny, “Survey of Multimodal Medical Question Answering,” *BioMedInformatics*, vol. 4, no. 1, pp. 50–74, Dec. 2023, doi: 10.3390/biomedinformatics4010004.
[137] Q. Chen, X. Hu, Z. Wang, and Y. Hong, “MedBLIP: Bootstrapping Language-Image Pre-training from 3D Medical Images and Texts,” *arXiv [cs.CV]*, May 18, 2023. [Online]. Available: http://arxiv.org/abs/2305.10799
[138] J. Liu, Z. Wang, Q. Ye, D. Chong, P. Zhou, and Y. Hua, “Qilin-Med-VL: Towards Chinese Large Vision-Language Model for General Healthcare,” *arXiv [cs.CV]*, Oct. 2023, [Online]. Available: https://arxiv.org/abs/2310.17956
[139] X. Yang, L. Xu, H. Li, and S. Zhang, “ViLaM: A Vision-Language Model with Enhanced Visual Grounding and Generalization Capability,” *arXiv [cs.CV]*, Nov. 21, 2023. [Online]. Available: http://arxiv.org/abs/2311.12327
[140] Q. Li *et al.*, “From Beginner to Expert: Modeling Medical Knowledge into General LLMs,” *arXiv [cs.CL]*, Dec. 02, 2023. [Online]. Available: http://arxiv.org/abs/2312.01040
[141] W. Gao *et al.*, “OphGLM: Training an Ophthalmology Large Language-and-Vision Assistant based on Instructions and Dialogue,” *arXiv [cs.CV]*, Jun. 21, 2023. [Online]. Available: http://arxiv.org/abs/2306.12174
[142] H. Zhu, R. Togo, T. Ogawa, and M. Haseyama, “A Medical Domain Visual Question Generation Model via Large Language Model,” in *2023 International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan)*, IEEE, Jul. 2023, pp. 163–164. doi: 10.1109/ICCE-Taiwan58799.2023.10227045.
[143] M. Muñoz-Echeverría *et al.*, “The hydrostatic-to-lensing mass bias from resolved X-ray and optical-IR data,” *arXiv [astro-ph.CO]*, Dec. 02, 2023. [Online]. Available: http://arxiv.org/abs/2312.01154
[144] X. Zhang *et al.*, “PMC-VQA: Visual Instruction Tuning for Medical Visual Question Answering,” *arXiv [cs.CV]*, May 17, 2023. [Online]. Available: http://arxiv.org/abs/2305.10415
[145] F. Liu *et al.*, “A medical multimodal large language model for future pandemics,” *NPJ Digit Med*, vol. 6, no. 1, p. 226, Dec. 2023, doi: 10.1038/s41746-023-00952-2.
[146] T. Zhu *et al.*, “A Large Language Modelling Deep Learning Framework for the Next Pandemic,” May 2023, doi: 10.21203/rs.3.rs-2777372/v1.
[147] F. Mehboob, K. M. Malik, A. K. J. Saudagar, A. Rauf, and A. AlTameem, “Medical Report Generation and Chatbot for COVID_19 Diagnosis Using Open-AI,” Feb. 08, 2023. doi: 10.21203/rs.3.rs-2563448/v1.
[148] S. Wu, B. Yang, Z. Ye, H. Wang, H. Zheng, and T. Zhang, “Improving Medical Report Generation with Adapter Tuning and Knowledge Enhancement in Vision-Language Foundation Models,” *arXiv [cs.CV]*, Dec. 07, 2023. [Online]. Available: http://arxiv.org/abs/2312.03970
[149] L. Yang, Z. Wang, and L. Zhou, “MedXChat: Bridging CXR Modalities with a Unified Multimodal Large Model,” *arXiv [cs.CV]*, Dec. 04, 2023. [Online]. Available: http://arxiv.org/abs/2312.02233
[150] A. Nicolson, J. Dowling, and B. Koopman, “Improving Chest X-Ray Report Generation by Leveraging Warm Starting,” *arXiv [cs.CV]*, Jan. 24, 2022. [Online]. Available: http://arxiv.org/abs/2201.09405
[151] O. Thawakar *et al.*, “XrayGPT: Chest radiographs summarization using medical vision-language models,” *ArXiv*, vol. abs/2306.07971, Jun. 2023, doi: 10.48550/arXiv.2306.07971.
[152] Y. Lu, S. Hong, Y. Shah, and P. Xu, “Effectively Fine-tune to Improve Large Multimodal Models for Radiology Report Generation,” *arXiv [cs.CV]*, Dec. 03, 2023. [Online]. Available: http://arxiv.org/abs/2312.01504

[153] S. V. C. Sai, E. T. Nikhil, R. Ponraj, and K. K., “Comprehensive Strategy for Analyzing Dementia Brain Images and Generating Textual Reports through ViT,” in *2023 First International Conference on Advances in Electrical, Electronics and Computational Intelligence (ICAEECI)*, unknown, Oct. 2023, pp. 1–10. doi: 10.1109/ICAEECI58247.2023.10370864.
[154] C. Niu and G. Wang, “CT Multi-Task Learning with a Large Image-Text (LIT) Model,” *arXiv [eess.IV]*, Apr. 03, 2023. [Online]. Available: http://arxiv.org/abs/2304.02649
[155] Y. Gu *et al.*, “BiomedJourney: Counterfactual Biomedical Image Generation by Instruction-Learning from Multimodal Patient Journeys,” *arXiv [cs.CV]*, Oct. 16, 2023. [Online]. Available: http://arxiv.org/abs/2310.10765
[156] A. Selivanov, O. Y. Rogov, D. Chesakov, A. Shelmanov, I. Fedulova, and D. V. Dylov, “Medical Image Captioning via Generative Pretrained Transformers,” *arXiv [cs.CV]*, Sep. 28, 2022. [Online]. Available: http://arxiv.org/abs/2209.13983
[157] G.-Y. Kim, B.-D. Oh, C. Kim, and Y.-S. Kim, “Convolutional Neural Network and Language Model-Based Sequential CT Image Captioning for Intracerebral Hemorrhage,” *NATO Adv. Sci. Inst. Ser. E Appl. Sci.*, vol. 13, no. 17, p. 9665, Aug. 2023, doi: 10.3390/app13179665.
[158] Z. Zhang *et al.*, “Sam-Guided Enhanced Fine-Grained Encoding with Mixed Semantic Learning for Medical Image Captioning,” *arXiv [cs.CV]*, Nov. 02, 2023. [Online]. Available: http://arxiv.org/abs/2311.01004
[159] Y. Zhang and D. Z. Chen, “GPT4MIA: Utilizing Generative Pre-trained Transformer (GPT-3) as A Plug-and-Play Transductive Model for Medical Image Analysis,” *arXiv [cs.CV]*, Feb. 17, 2023. [Online]. Available: http://arxiv.org/abs/2302.08722
[160] N. Lei *et al.*, “A Two-Stage Chinese Medical Video Retrieval Framework with LLM,” in *Natural Language Processing and Chinese Computing*, Springer Nature Switzerland, 2023, pp. 211–220. doi: 10.1007/978-3-031-44699-3_19.
[161] J. Kim, S. Yoon, T. Choi, and S. Sull, “Unsupervised Video Anomaly Detection Based on Similarity with Predefined Text Descriptions,” *Sensors* , vol. 23, no. 14, Jul. 2023, doi: 10.3390/s23146256.
[162] R. Wang *et al.*, “ECAMP: Entity-centered Context-aware Medical Vision Language Pre-training,” *arXiv [cs.CV]*, Dec. 20, 2023. [Online]. Available: http://arxiv.org/abs/2312.13316
[163] J. Zhou, X. Chen, and X. Gao, “Path to Medical AGI: Unify Domain-specific Medical LLMs with the Lowest Cost,” *arXiv [cs.AI]*, Jun. 19, 2023. [Online]. Available: http://arxiv.org/abs/2306.10765
[164] V. Sorin, B. S. Glicksberg, Y. Barash, E. Konen, G. Nadkarni, and E. Klang, “Diagnostic accuracy of GPT multimodal analysis on USMLE questions including text and visuals,” *bioRxiv*, Oct. 31, 2023. doi: 10.1101/2023.10.29.23297733.
[165] Z. Yang *et al.*, “Performance of Multimodal GPT-4V on USMLE with Image: Potential for Imaging Diagnostic Support with Explanations,” *medRxiv*, p. 2023.10.26.23297629, Nov. 03, 2023. doi: 10.1101/2023.10.26.23297629.
[166] T. Nakao *et al.*, “Capability of GPT-4V(ision) in Japanese National Medical Licensing Examination,” *medRxiv*, p. 2023.11.07.23298133, Nov. 08, 2023. doi: 10.1101/2023.11.07.23298133.
[167] H. Sun *et al.*, “An AI Dietitian for Type 2 Diabetes Mellitus Management Based on Large Language and Image Recognition Models: Preclinical Concept Validation Study,” *J. Med. Internet Res.*, vol. 25, p. e51300, Nov. 2023, doi: 10.2196/51300.
[168] R. Noda, Y. Izaki, F. Kitano, J. Komatsu, D. Ichikawa, and Y. Shibagaki, “Performance of ChatGPT and Bard in Self-Assessment Questions for Nephrology Board Renewal,” *bioRxiv*, Jun. 12, 2023. doi: 10.1101/2023.06.06.23291070.
[169] M. Angel *et al.*, “AI and Veterinary Medicine: Performance of Large Language Models on

the North American Licensing Examination," in *2023 Tenth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, IEEE, Nov. 2023, pp. 1–4. doi: 10.1109/SNAMS60348.2023.10375414.
[170] Y. Li *et al.*, "A Comprehensive Study of GPT-4V's Multimodal Capabilities in Medical Imaging." Accessed: Mar. 13, 2024. [Online]. Available: https://www.medrxiv.org/content/10.1101/2023.11.03.23298067v1.abstract
[171] R. Chen *et al.*, "GPT-4 Vision on Medical Image Classification -- A Case Study on COVID-19 Dataset," *arXiv [eess.IV]*, Oct. 27, 2023. [Online]. Available: http://arxiv.org/abs/2310.18498
[172] Y. Cao, X. Xu, C. Sun, X. Huang, and W. Shen, *Towards Generic Anomaly Detection and Understanding: Large-scale Visual-linguistic Model (GPT-4V) Takes the Lead*. Github, 2023. Accessed: Mar. 13, 2024. [Online]. Available: https://github.com/caoyunkang
[173] T. Han, L. C. Adams, S. Nebelung, J. N. Kather, K. K. Bressem, and D. Truhn, "Multimodal large language models are generalist medical image interpreters," *bioRxiv*, Dec. 22, 2023. doi: 10.1101/2023.12.21.23300146.
[174] E. Waisberg, J. Ong, M. Masalkhi, S. A. Kamran, and A. Tavakkoli, "Automated ophthalmic imaging analysis in the era of Generative Pre-Trained Transformer-4," *The Pan-American Journal of Ophthalmology*, vol. 5, no. 1, Nov. 2023, doi: 10.4103/pajo.pajo_62_23.
[175] C. Wu *et al.*, "Can GPT-4V(ision) Serve Medical Applications? Case Studies on GPT-4V for Multimodal Medical Diagnosis," *arXiv [cs.CV]*, Oct. 15, 2023. [Online]. Available: http://arxiv.org/abs/2310.09909
[176] M. C. Schubert, M. Lasotta, F. Sahm, W. Wick, and V. Venkataramani, "Evaluating the Multimodal Capabilities of Generative AI in Complex Clinical Diagnostics," *medRxiv*, p. 2023.11.01.23297938, Nov. 02, 2023. doi: 10.1101/2023.11.01.23297938.
[177] H. Yang, R. Wang, C. Wang, H. Gao, and H. Cai, "GPT-4 and Neurologists in Screening for Mild Cognitive Impairment in the Elderly: A Comparative Analysis Study," *medRxiv*, Dec. 2023, doi: 10.1101/2023.12.02.23299327.
[178] S. M. Senthujan *et al.*, "GPT-4V(ision) Unsuitable for Clinical Care and Education: A Clinician-Evaluated Assessment," *medRxiv*, p. 2023.11.15.23298575, Nov. 16, 2023. doi: 10.1101/2023.11.15.23298575.
[179] L. A. Cox Jr, "Pushing Back on AI: A Dialogue with ChatGPT on Causal Inference in Epidemiology," in *AI-ML for Decision and Risk Analysis: Challenges and Opportunities for Normative Decision Theory*, L. A. Cox Jr, Ed., Cham: Springer International Publishing, 2023, pp. 407–423. doi: 10.1007/978-3-031-32013-2_13.
[180] K. R. Kanakarajan and M. Sankarasubbu, "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 *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, and E. Sartori, Eds., Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 995–1003. doi: 10.18653/v1/2023.semeval-1.137.
[181] J. A. Lossio-Ventura *et al.*, "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*, vol. 11, p. e50150, Jan. 2024, doi: 10.2196/50150.
[182] S. De and S. Vats, "Decoding Concerns: Multi-label Classification of Vaccine Sentiments in Social Media," *arXiv [cs.CL]*, Dec. 17, 2023. [Online]. Available: http://arxiv.org/abs/2312.10626
[183] K. E. Abramski, S. Citraro, L. Lombardi, G. Rossetti, and M. Stella, "Cognitive network science reveals bias in GPT-3, ChatGPT, and GPT-4 mirroring math anxiety in high-school students," May 2023. doi: 10.31234/osf.io/27u6z.

[184] P. Clarke *et al.*, “From a Large Language Model to Three-Dimensional Sentiment,” Aug. 2023. doi: 10.31234/osf.io/kaeqy.
[185] S. Mittal and M. De Choudhury, “Moral Framing of Mental Health Discourse and Its Relationship to Stigma: A Comparison of Social Media and News,” in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, in CHI ’23, no. 484. New York, NY, USA: Association for Computing Machinery, Apr. 2023, pp. 1–19. doi: 10.1145/3544548.3580834.
[186] Z. Szántó, B. Bánáti, and T. Zombori, “Enhancing Medication Event Classification with Syntax Parsing and Adversarial Learning,” in *Artificial Intelligence Applications and Innovations*, Springer Nature Switzerland, 2023, pp. 114–124. doi: 10.1007/978-3-031-34111-3_11.
[187] X. Zhang and A. A. Ansah, “A mobile app for tracking psychological mood changes and providing E-therapy using natural language processing and GPT-3,” in *Artificial Intelligence & Applications*, Academy & Industry Research Collaboration Center, Oct. 2023. doi: 10.5121/csit.2023.131925.
[188] L. Gómez-Zaragozá, M. E. Minissi, J. Llanes-Jurado, A. Altozano, M. Alcañiz Raya, and J. Marín-Morales, “Linguistic Indicators of Depressive Symptoms in Conversations with Virtual Humans,” in *Collaborative Networks in Digitalization and Society 5.0*, Springer Nature Switzerland, 2023, pp. 521–534. doi: 10.1007/978-3-031-42622-3_37.
[189] H. Qi *et al.*, “Supervised Learning and Large Language Model Benchmarks on Mental Health Datasets: Cognitive Distortions and Suicidal Risks in Chinese Social Media,” Nov. 2023, doi: 10.21203/rs.3.rs-3523508/v1.
[190] E. Theophilou *et al.*, “Learning to Prompt in the Classroom to Understand AI Limits: A pilot study,” *arXiv [cs.HC]*, Jul. 04, 2023. [Online]. Available: http://arxiv.org/abs/2307.01540
[191] N. Forman, J. Udvaros, and M. S. Avornicului, “ChatGPT: A new study tool shaping the future for high school students,” *IJANSER*, vol. 7, no. 4, pp. 95–102, May 2023, doi: 10.59287/ijanser.562.
[192] N. Abouammoh *et al.*, “Exploring perceptions and experiences of ChatGPT in medical education: A qualitative study among medical college faculty and students in Saudi Arabia,” *bioRxiv*, Jul. 16, 2023. doi: 10.1101/2023.07.13.23292624.
[193] B. C. Gin, O. ten Cate, P. S. O’Sullivan, and C. K. Boscardin, “Trainee versus supervisor viewpoints of entrustment: using artificial intelligence language models to detect thematic differences and potential biases,” Aug. 2023, doi: 10.21203/rs.3.rs-3223749/v1.
[194] Perlis Roy H. and Jones David S., “High-Impact Medical Journals Reflect Negative Sentiment Toward Psychiatry,” *NEJM AI*, vol. 1, no. 1, p. AIcs2300066, Dec. 2023, doi: 10.1056/AIcs2300066.
[195] T. Susnjak, “Applying BERT and ChatGPT for Sentiment Analysis of Lyme Disease in Scientific Literature,” *arXiv [cs.CL]*, Feb. 07, 2023. [Online]. Available: http://arxiv.org/abs/2302.06474
[196] M. Li and R. Zhang, “How far is Language Model from 100% Few-shot Named Entity Recognition in Medical Domain,” *arXiv [cs.CL]*, Jul. 01, 2023. [Online]. Available: http://arxiv.org/abs/2307.00186
[197] M. Iscoe *et al.*, “Identifying Signs and Symptoms of Urinary Tract Infection from Emergency Department Clinical Notes Using Large Language Models,” *medRxiv*, p. 2023.10.20.23297156, Oct. 26, 2023. doi: 10.1101/2023.10.20.23297156.
[198] E. Jo, D. A. Epstein, H. Jung, and Y.-H. Kim, “Understanding the Benefits and Challenges of Deploying Conversational AI Leveraging Large Language Models for Public Health Intervention,” in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, in CHI ’23, no. 18. New York, NY, USA: Association for Computing Machinery, Apr. 2023, pp. 1–16. doi: 10.1145/3544548.3581503.
[199] S. Akilesh, A. A. Sheik, R. Abinaya, S. Dhanushkodi, and R. Sekar, “A Novel AI-based

chatbot Application for Personalized Medical Diagnosis and review using Large Language Models," in *2023 International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE)*, IEEE, Nov. 2023, pp. 1–5. doi: 10.1109/RMKMATE59243.2023.10368616.

[200] S. Chen *et al.*, "Use of Artificial Intelligence Chatbots for Cancer Treatment Information," *JAMA oncology*, vol. 9, no. 10. pp. 1459–1462, Oct. 01, 2023. doi: 10.1001/jamaoncol.2023.2954.

[201] S. Chen *et al.*, "The impact of responding to patient messages with large language model assistance," *arXiv e-prints*, p. arXiv:2310.17703, Oct. 2023, doi: 10.48550/arXiv.2310.17703.

[202] B. Laker and E. Currell, "ChatGPT: a novel AI assistant for healthcare messaging-a commentary on its potential in addressing patient queries and reducing clinician burnout," *BMJ Lead*, Sep. 2023, doi: 10.1136/leader-2023-000844.

[203] T. F. Heston, "Safety of Large Language Models in Addressing Depression," *Cureus*, vol. 15, no. 12, p. e50729, Dec. 2023, doi: 10.7759/cureus.50729.

[204] L. Campillos-Llanos, C. Thomas, É. Bilinski, A. Neuraz, S. Rosset, and P. Zweigenbaum, "Lessons Learned from the Usability Evaluation of a Simulated Patient Dialogue System," *J. Med. Syst.*, vol. 45, no. 7, p. 69, May 2021, doi: 10.1007/s10916-021-01737-4.

[205] A. Osmanovic-Thunström and S. Steingrimsson, "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*, vol. 3, no. 1, p. 12, Apr. 2023, doi: 10.1007/s44163-023-00055-7.

[206] B. N. Hryciw, A. J. E. Seely, and K. Kyeremanteng, "Guiding principles and proposed classification system for the responsible adoption of artificial intelligence in scientific writing in medicine," *Front Artif Intell*, vol. 6, p. 1283353, Nov. 2023, doi: 10.3389/frai.2023.1283353.

[207] M. Abu-Jeyyab, S. Alrosan, and I. Alkhawaldeh, "Harnessing Large Language Models in Medical Research and Scientific Writing: A Closer Look to The Future: LLMs in Medical Research and Scientific Writing," *HYMR*, vol. 1, no. 2, Dec. 2023, doi: 10.59707/hymrFBYA5348.

[208] M. C. Schubert, W. Wick, and V. Venkataramani, "Performance of Large Language Models on a Neurology Board-Style Examination," *JAMA Netw Open*, vol. 6, no. 12, p. e2346721, Dec. 2023, doi: 10.1001/jamanetworkopen.2023.46721.

[209] A. Abd-Alrazaq *et al.*, "Large Language Models in Medical Education: Opportunities, Challenges, and Future Directions," *JMIR Med Educ*, vol. 9, p. e48291, Jun. 2023, doi: 10.2196/48291.

[210] T. F. Heston, "Evaluating risk progression in mental health chatbots using escalating prompts," *bioRxiv*, Sep. 12, 2023. doi: 10.1101/2023.09.10.23295321.

[211] J. M. Liu, D. Li, H. Cao, T. Ren, Z. Liao, and J. Wu, "ChatCounselor: A Large Language Models for Mental Health Support," *arXiv [cs.CL]*, Sep. 27, 2023. [Online]. Available: http://arxiv.org/abs/2309.15461

[212] H. Song *et al.*, "Evaluating the Performance of Different Large Language Models on Health Consultation and Patient Education in Urolithiasis," *J. Med. Syst.*, vol. 47, no. 1, p. 125, Nov. 2023, doi: 10.1007/s10916-023-02021-3.

[213] H. Di and Y. Wen, "Evaluating the Effectiveness of Artificial Intelligence-powered Large Language Models Application in Disseminating Appropriate and Readable Health Information in Urology. Letter," *J. Urol.*, vol. 210, no. 5, pp. 735–736, Nov. 2023, doi: 10.1097/JU.0000000000003655.

[214] Z. Liu *et al.*, "Evaluating Large Language Models for Radiology Natural Language Processing," *arXiv [cs.CL]*, Jul. 25, 2023. [Online]. Available: http://arxiv.org/abs/2307.13693

[215] S. Reddy, “Evaluating large language models for use in healthcare: A framework for translational value assessment,” *Informatics in Medicine Unlocked*, vol. 41, p. 101304, Jan. 2023, doi: 10.1016/j.imu.2023.101304.
[216] Z. Liu *et al.*, “RadLLM: A Comprehensive Healthcare Benchmark of Large Language Models for Radiology,” *arXiv [cs.CL]*, Jul. 25, 2023. [Online]. Available: http://arxiv.org/abs/2307.13693
[217] H. Jin, S. Chen, M. Wu, and K. Q. Zhu, “PsyEval: A Comprehensive Large Language Model Evaluation Benchmark for Mental Health,” *arXiv [cs.CL]*, Nov. 15, 2023. [Online]. Available: http://arxiv.org/abs/2311.09189
[218] Z. He *et al.*, “MedEval: A Multi-Level, Multi-Task, and Multi-Domain Medical Benchmark for Language Model Evaluation,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 8725–8744. doi: 10.18653/v1/2023.emnlp-main.540.
[219] Y. Liao, Y. Meng, H. Liu, Y. Wang, and Y. Wang, “An Automatic Evaluation Framework for Multi-turn Medical Consultations Capabilities of Large Language Models,” *arXiv [cs.CL]*, Sep. 05, 2023. [Online]. Available: http://arxiv.org/abs/2309.02077
[220] L. Tang *et al.*, “Evaluating large language models on medical evidence summarization,” *NPJ Digit Med*, vol. 6, no. 1, p. 158, Aug. 2023, doi: 10.1038/s41746-023-00896-7.
[221] X. Yao, M. Mikhelson, S. Craig Watkins, E. Choi, E. Thomaz, and K. de Barbaro, “Development and Evaluation of Three Chatbots for Postpartum Mood and Anxiety Disorders,” *arXiv [cs.CL]*, Aug. 14, 2023. doi: 10.1145/nnnnnnn.nnnnnnn.
[222] D. Duong and B. D. Solomon, “Analysis of large-language model versus human performance for genetics questions,” *medRxiv*, Jan. 2023, doi: 10.1101/2023.01.27.23285115.
[223] E. Fournier-Tombs and J. McHardy, “A Medical Ethics Framework for Conversational Artificial Intelligence,” *J. Med. Internet Res.*, vol. 25, p. e43068, Jul. 2023, doi: 10.2196/43068.
[224] S. Perni, L. S. Lehmann, and D. S. Bitterman, “Patients should be informed when AI systems are used in clinical trials,” *Nat. Med.*, vol. 29, no. 8, pp. 1890–1891, Aug. 2023, doi: 10.1038/s41591-023-02367-8.
[225] L. G. Valiña and I. Mastroleo, “The ethical and scientific challenges of ChatGPT in health: utopianism, technophobia and pragmatism,” Aug. 2023. doi: 10.31219/osf.io/kvj45.
[226] I. G. Cohen, “What Should ChatGPT Mean for Bioethics?,” *Am. J. Bioeth.*, vol. 23, no. 10, pp. 8–16, Oct. 2023, doi: 10.1080/15265161.2023.2233357.
[227] H. Li, J. T. Moon, S. Purkayastha, L. A. Celi, H. Trivedi, and J. W. Gichoya, “Ethics of large language models in medicine and medical research,” *Lancet Digit Health*, vol. 5, no. 6, pp. e333–e335, Jun. 2023, doi: 10.1016/S2589-7500(23)00083-3.
[228] A. Piñeiro-Martín, C. Garcia-Mateo, L. Docío-Fernández, and M. del C. López Pérez, “Ethical challenges in the development of virtual assistants powered by Large Language Models,” *Preprints*, Jun. 02, 2023. doi: 10.20944/preprints202306.0196.v1.
[229] R. D’Souza and A. Sousa, “Ethics in managing big data: Ensuring privacy and data security while using ChatGPT in healthcare,” *Glob Bioeth Enq J*, Jan. 2023, doi: 10.38020/gbe.11.1.2023.1-4.
[230] H. Mazumdar, C. Chakraborty, M. Sathvik, S. Mukhopadhyay, and P. K. Panigrahi, “GPTFX: A Novel GPT-3 Based Framework for Mental Health Detection and Explanations,” *IEEE J Biomed Health Inform*, vol. PP, Oct. 2023, doi: 10.1109/JBHI.2023.3328350.
[231] G. Fu *et al.*, “Enhancing Psychological Counseling with Large Language Model: A Multifaceted Decision-Support System for Non-Professionals,” *arXiv [cs.AI]*, Aug. 29, 2023. [Online]. Available: http://arxiv.org/abs/2308.15192
[232] H. Li, R. Zhang, Y.-C. Lee, R. E. Kraut, and D. C. Mohr, “Systematic review and

meta-analysis of AI-based conversational agents for promoting mental health and well-being," *NPJ Digit Med*, vol. 6, no. 1, p. 236, Dec. 2023, doi: 10.1038/s41746-023-00979-5.
[233] F. Agbavor and H. Liang, "Predicting dementia from spontaneous speech using large language models," *PLOS Digit Health*, vol. 1, no. 12, p. e0000168, Dec. 2022, doi: 10.1371/journal.pdig.0000168.
[234] H. Cai *et al.*, "Multimodal Approaches for Alzheimer's Detection Using Patients' Speech and Transcript," in *Brain Informatics*, Springer Nature Switzerland, 2023, pp. 395–406. doi: 10.1007/978-3-031-43075-6_34.
[235] R. H. Perlis, "Research Letter: Application of GPT-4 to select next-step antidepressant treatment in major depression," *medRxiv*, Apr. 2023, doi: 10.1101/2023.04.14.23288595.
[236] K. Yang, S. Ji, T. Zhang, Q. Xie, Z. Kuang, and S. Ananiadou, "Towards Interpretable Mental Health Analysis with Large Language Models," *arXiv [cs.CL]*, Apr. 06, 2023. [Online]. Available: http://arxiv.org/abs/2304.03347
[237] B. Lamichhane, "Evaluation of ChatGPT for NLP-based Mental Health Applications," *arXiv [cs.CL]*, Mar. 28, 2023. [Online]. Available: http://arxiv.org/abs/2303.15727
[238] S. Tripathy, R. Singh, and M. Ray, "Natural Language Processing for Covid-19 Consulting System," *Procedia Comput. Sci.*, vol. 218, pp. 1335–1341, Jan. 2023, doi: 10.1016/j.procs.2023.01.112.
[239] L. Zhang, S. Tashiro, M. Mukaino, and S. Yamada, "Use of artificial intelligence large language models as a clinical tool in rehabilitation medicine: a comparative test case," *J. Rehabil. Med.*, vol. 55, p. jrm13373, Sep. 2023, doi: 10.2340/jrm.v55.13373.
[240] M. A. Ahmad, I. Yaramis, and T. D. Roy, "Creating Trustworthy LLMs: Dealing with Hallucinations in Healthcare AI," *arXiv [cs.CL]*, Sep. 26, 2023. [Online]. Available: http://arxiv.org/abs/2311.01463
[241] N. C. Chung, G. Dyer, and L. Brocki, "Challenges of Large Language Models for Mental Health Counseling," *arXiv [cs.CL]*, Nov. 23, 2023. [Online]. Available: http://arxiv.org/abs/2311.13857
[242] M. De Choudhury, S. R. Pendse, and N. Kumar, "Benefits and Harms of Large Language Models in Digital Mental Health," *arXiv [cs.CL]*, Nov. 07, 2023. [Online]. Available: http://arxiv.org/abs/2311.14693

Appendices

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

Table 2 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 the co-occurrence, the higher the link strength. Occurrence is the number of times a given keyword appears across the corpus.

Table 2. Top 50 Concepts in Keyword Co-occurrence Network

Keyword	Occurrences	Total Link Strength
computer science	1394	16099
artificial intelligence	932	11246
medicine	1033	10710
psychology	719	8286
political science	483	5850
law	461	5602
natural language processing	371	4845
mathematics	344	4749
data science	363	4606
biology	344	4585
programming language	334	4471
economics	303	4262
health care	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
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

Appendix II. Representative Papers for each LLM Task

Table 3 presents the representative papers for each LLM task and their respective DOI.

Table 3. 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
Dialogue 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	https://doi.org/10.1136/leader-2023-000844

	reducing clinician burnout	
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
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
Image, Vision, Video and Multimodality	ChatCAD: Interactive Computer-Aided Diagnosis on Medical Image 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-review 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	https://doi.org/10.18653/v1/2023.semeval-1.137

	Multi-evidence Natural Language Inference in Clinical Trial Data	
	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
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

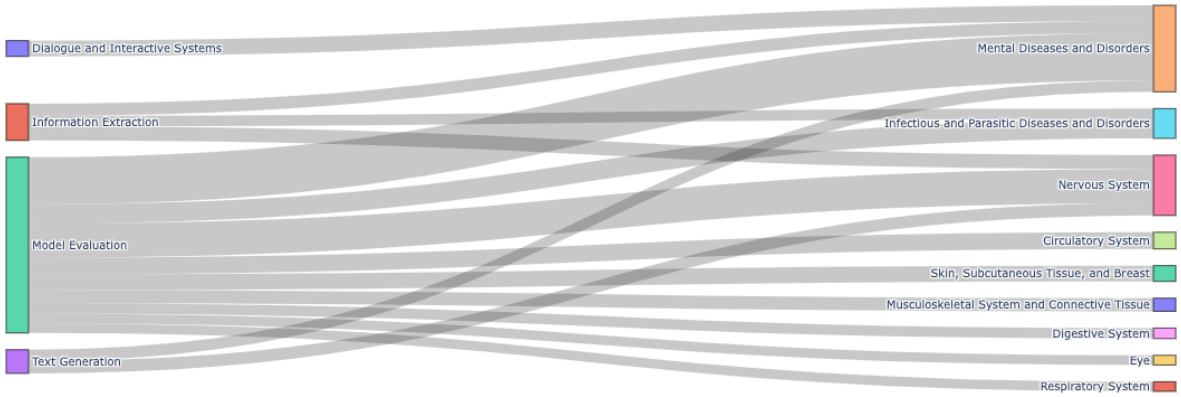
Appendix III: Analysis of LLM Task: Meta-Analysis and Literature Review

Systematic reviews and meta-analyses in this domain critically assess LLMs, focusing on their capacity to revolutionize various aspects of medical practice (Ghim & Ahn, 2023; M. Hu et al., 2023; Lawson McLean, 2023; Miao et al., 2023; Sanii et al., 2023; T. Zhu et al., 2023) and provide guidelines on their applications (Abi-Rafeh et al., 2024; P. Yu et al., 2023). 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 (Dossantos et al., 2023; Nasarian et al., 2023; Rammohan et al., 2023; Sohail, 2023). Detailed investigations into the methodologies reveal how advanced techniques like generative pre-trained transformers (Tanaka et al., 2023) and fine-tuning (Z. Liu et al., 2024) on medical datasets are applied to create innovative applications, from automated medical reporting to virtual patient engagement tools (Lun et al., 2023). The other literature suggests future developments, such as emphasizing the need for richer training data (Cazzato et al., 2023) (Schukow et al., 2024), enhancing interdisciplinary research collaborations (Suppadungsuk et al., 2023), and setting up stringent ethical standards to ensure that LLMs can be safely integrated into patient care (Dossantos et al., 2023; Gödde et al., 2023). But 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.

Appendix IV. 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 (**Figure 7** and **Table 4**). The classification tasks are usually the focus of model evaluation.

Figure 7: Sankey Diagram of LLM Tasks and Disease Categories (with Paper Count more than 10)



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 [235], explanation generation of analysis results [236], detection of mental diseases and disorders [237], 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 [137], [237], [238]. For instance, [237] evaluates LLM-based ChatGPT on mental health classification tasks with three publically 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 publically available, researchers come up with classification tasks on their own in specific scenarios [235], [239]. For example, [235]
created brief vignettes about the decision of selecting 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 [230], [235], [236]. In addition to popular automatic evaluation metrics like perplexity, BLEU-n, and ROUGE-1 [230], [238], studies also use human annotation for evaluation and for benchmarking automatic evaluation metrics [236], [240]. Additionally, approaches based on prompt engineering are also taken to evaluate the interaction between LLMs and agents by analyzing their mental health

referral patterns [210]. Apart from technical literature, other research has also examined and identified the benefits and harms of using LLMs for mental health counseling [241], [242] and the issues of hallucination [240].

Table 4: 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	https://doi.org/10.1007/978-3-031-48044-7_26

		Vaccination Tweets	
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
Digestive System	12	Advanced prompting as a catalyst:	https://doi.org/10.59717/j.xinn-med.2023.100

		Empowering large language models in the management of gastrointestinal cancers	019
		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