Generative Large Language Models in Electronic Health Records for Patient Care Since 2023: A Systematic Review

Xinsong Du[1,2,3], Zhengyang Zhou[4], Yifei Wang[4], Ya-Wen Chuang[5,6,7], Richard Yang[1,2], Wenyu Zhang[1,2,3], Xinyi Wang[1,2,3], Rui Zhang[8], Pengyu Hong[4], David W. Bates[1,2,9], Li Zhou[1,2,3]

[1] Division of General Internal Medicine and Primary Care, Brigham and Women's Hospital, Boston, Massachusetts 02115
[2] Department of Medicine, Harvard Medical School, Boston, Massachusetts 02115
[3] Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts 02115
[4] Department of Computer Science, Brandeis University, Waltham, MA 02453
[5] Division of Nephrology, Department of Internal Medicine, Taichung Veterans General Hospital, Taichung, Taiwan, 407219
[6] Department of Post-Baccalaureate Medicine, College of Medicine, National Chung Hsing University, Taichung, Taiwan, 402202
[7] School of Medicine, College of Medicine, China Medical University, Taichung, Taiwan, 404328
[8]Division of Computational Health Sciences, University of Minnesota, Minneapolis, MN 55455
[9]Department of Health Policy and Management, Harvard T.H. Chan School of Public Health, Boston, MA 02115

Corresponding author:
Xinsong Du, Ph.D.
Division of General Internal Medicine and Primary Care
Department of Medicine
Brigham and Women’s Hospital and Harvard Medical School
399 Revolution Dr, Suite 777
Somerville, MA 02145
Tel: 3522811387 | E-mail: xidu1@bwh.harvard.edu

Abstract

Background: Generative Large language models (LLMs) represent a significant advancement in natural language processing, achieving state-of-the-art performance across various tasks. However, their application in clinical settings using real electronic health records (EHRs) is still rare and presents numerous challenges.

Objective: This study aims to systematically review the use of generative LLMs in patient care-related topics involving EHRs, summarize the challenges faced, and suggest future directions.

Methods: A Boolean search for peer-reviewed articles was conducted in May 2024 using PubMed and Web of Science to include research articles published since 2023, which was one month after the release of ChatGPT. The search results were deduplicated. Multiple reviewers, including biomedical informaticians, computer scientists, and a physician, screened the publications for eligibility and extracted bibliometric and clinically relevant information. Only papers utilizing generative LLMs to analyze real EHR data were included. We summarized the use of prompt engineering, fine-tuning, multimodal EHR data, and evaluation matrices. Additionally, we identified current challenges in applying LLMs in clinical settings as reported by the included papers and proposed future directions.

Results: The initial search identified 6,328 unique studies, with 76 studies included after eligibility screening. Of these, 67 studies (88.2%) employed zero-shot prompting, five of them reported 100% accuracy on five specific clinical tasks. Nine studies used advanced prompting strategies; four tested these strategies experimentally, finding that prompt engineering improved performance, with one study noting a non-linear relationship between the number of examples in a prompt and performance improvement. Eight studies explored fine-tuning generative LLMs, all reported performance improvements on specific tasks, but three of them noted potential performance degradation after fine-tuning on certain tasks. Only two studies utilized multimodal data, which improved LLM-based decision-making and enabled accurate rare disease diagnosis and prognosis. The studies employed 55 different evaluation metrics for 22 purposes, such as correctness, completeness, and conciseness. Two studies investigated LLM bias, with one detecting no bias and the other finding that male patients received more appropriate clinical decision-making suggestions. Six studies identified hallucinations, such as fabricating patient names in structured thyroid ultrasound reports. Additional challenges included but not limited to the impersonal tone of LLM consultations, which made patients uncomfortable, and the difficulty patients had in understanding LLM responses.

Conclusion: Our review indicates that few studies have employed advanced computational techniques to enhance LLM performance. The diverse evaluation metrics used highlight the need for standardization. LLMs currently cannot replace physicians due to challenges such as bias, hallucinations, and impersonal responses.

1. Introduction

Recent advancements in transformer-based generative large language models (LLMs) have significantly transformed the landscape of natural language processing (NLP) and artificial intelligence (AI)[1–3]. These models, distinguished by their substantial size and intricate architecture, have gained widespread recognition in both academic and industrial domains due to their extraordinary capability to comprehend and generate human-like reasoning[4]. Prominent examples of generative LLMs include the GPT (Generative Pre-trained Transformer) series by OpenAI[5], the Llama series by Meta[6], and Gemini by Google. These models are pre-trained on vast corpora of text data and subsequently fine-tuned for specific downstream tasks[7]. With billions to trillions of parameters, they are exceptionally proficient in capturing complex linguistic patterns and subtleties, achieving unprecedented levels of accuracy and depth.

Given the remarkable capabilities of LLMs in processing text data, there has been a surge in research exploring their applications in healthcare. Numerous studies have reviewed and synthesized recent advancements in applying LLMs to various healthcare domains.[8–11] Some researchers have evaluated LLMs' ability to answer healthcare-related questions by analyzing their responses to queries from medical specialty associations.[12–14] Other studies have tested LLMs' performance in specific clinical tasks, often benchmarking their accuracy against that of human experts.[15–18] Additionally, comparisons have been made between LLMs and traditional AI approaches,[19] as well as search engines[20–22] to assess their relative effectiveness. Despite these advances, there are still emerging opportunities and challenges in leveraging LLMs in healthcare, particularly in the statistical analysis of where these models have been most and least effectively applied.

Electronic health records (EHRs) have revolutionized healthcare by offering a comprehensive digital repository of a patient's medical history, accessible to authorized providers across various healthcare settings. This seamless information sharing significantly enhances the quality, safety, and efficiency of patient care by integrating diverse data types, including medical history, diagnoses, medications, and test results. EHRs facilitate more accurate and timely decision-making, reduce the likelihood of medical errors, and contribute to improved patient outcomes. Additionally, they serve as an invaluable resource for healthcare research and quality improvement initiatives. However, the vast and complex datasets generated by EHRs present growing challenges for effective analysis and utilization.

Generative Large Language Models (LLMs) are emerging as powerful tools to address the challenges of analyzing complex healthcare data by uncovering intricate patterns, enhancing predictive analytics, and advancing personalized medicine. However, despite their potential, generative LLMs also introduce risks, such as biases and hallucinations, which could mislead clinical decision-making. Although LLMs have been extensively studied for healthcare data analysis, privacy concerns add another layer of complexity to their application in real-world EHR data, making such use cases relatively rare. For instance, as of April 18, 2023, an announcement explicitly prohibited the use of ChatGPT on a popular EHR data source, Medical Information Mart for Intensive Care (MIMIC) data,[23–25] underscoring the need for Health Insurance Portability and Accountability Act (HIPAA)-compliant platforms to safely utilize LLMs on EHR data.[26] While many reviews have explored the broader application of LLMs in healthcare, there is a notable gap in focused analyses of their use in EHR data to improve patient care in specific clinical tasks. To address this gap, we conducted a systematic review that provides an in-depth analysis of generative LLM applications in EHR data, highlighting the opportunities, challenges, and future directions for their implementation in real clinical settings.

2. Methods

2.1. Study Selection Process

We adhered to the PRISMA guidelines for conducting our literature search (**Figure 1**).[27] The process involved several key steps: a Boolean search, removal of duplicates, screening of papers, and data extraction. The Boolean search was conducted on May 19th, 2024, with search terms and restrictions determined through team discussions. Our search included LLM-related terms, such as "prompt engineering" and the names of various LLMs; the detailed query can be found in **Supplementary Table S1**. To focus on research articles presenting original data and quantitative results, we excluded certain article types, such as reviews. Given that the first release of ChatGPT was on November 30th, 2022, we included articles published from 2023 onward. Our search was conducted in PubMed and Web of Science, with only peer-reviewed articles included, while preprints were excluded.

2.2. Inclusion and Exclusion Criteria

Generative LLMs have been employed with various types of medical data, including but not limited to medical imaging, pharmaceutical data, public health data, genomics, biometric data, and EHR data. Our review specifically focuses on the application of LLMs to original EHR data, excluding studies using synthetic or summarized EHR data. For each included paper, we summarized the data size and data source.

The selection process involved removing duplicate articles, followed by a manual review of the deduplicated list. The exclusion criteria were as follows: 1) the article is not the correct type (e.g., preprint, review, editorial, comment); 2) the article does not involve generative LLMs (e.g., it discusses a chatbot that does not utilize an LLM, or it only includes encoder-based models like Longformer[28], NYUTron[29], GatorTron[30]); 3) the LLM in the article was not used for English-language communication; 4) the LLM application is unrelated to patient care (e.g., using LLMs for passing exams or conducting research); 5) the article lacks quantitative evaluation (e.g., some articles may only include communication records with ChatGPT); 6) the article does not involve EHR data; 7) the EHR data used in the article is not original (e.g., it uses synthetic or summarized data).

During the paper screening process, two reviewers initially screened a set of 50 identical articles. If the agreement rate was above 90%, the reviewers proceeded to independently screen the remaining articles. If not, they discussed and screened an additional 50 articles until the agreement reached 90%.

2.3. Data Extraction and Statistical Analysis

For the included papers, we extracted various categories of information, as detailed in **Table 1**. This includes data-related information, clinical information, LLM-specific details, evaluation metrics, and identified challenges. The extracted data encompasses key aspects such as the nature and source of the data, the clinical context in which the LLM was applied, the specific LLM models and techniques used, the methods of evaluation employed, and the current challenges faced in these applications.

2.3.1. Data

We extracted details on data size and data source from the included studies. Data size refers to the number of samples used in each study, while data source indicates the origin of the data, which could be from a specific hospital or a publicly available EHR dataset, such as MIMIC.[24,25]

2.3.2. Clinical Domain

We extracted information on clinical tasks and specialties from each included study. The main clinical task of each study was described in text form, while the distribution of clinical specialties was summarized using a pie chart. Specialties represented in less than 5% of the studies were consolidated into a single category within the pie chart.

2.3.3. Generative LLM-Related Information

We extracted detailed information on the generative LLMs used in each article, including the prompting methods applied, fine-tuning approaches, and the integration of multimodal EHR data with LLMs to enhance patient care. For LLMs, we calculated the frequency of their usage across studies. Regarding prompting methods, we documented the specific techniques used, the clinical tasks they were applied to, and the quantitative impact of each prompting approach on the performance of these tasks. In terms of fine-tuning, we extracted information on the base models that were fine-tuned, the specific fine-tuning methods employed, the hardware used for the fine-tuning process, and the quantitative effects on clinical task performance. For studies involving the application of LLMs to multimodal EHR data, we summarized the data modalities involved, the methods used for data integration, and the quantitative impact of multimodal integration on performance. Additionally, we provided detailed explanations of existing techniques for prompt engineering, generative LLM fine-tuning, and multimodal data integration in the **Supplementary Material**.

2.3.4. Evaluation

Researchers employ various evaluation purposes depending on the specific clinical task when assessing LLM performance. For example, in clinical decision-making, the emphasis may be on the accuracy and completeness of the LLM's output, whereas for clinical note summarization or simplification, readability and conciseness are primary evaluation criteria. Given that LLM responses may be used in clinical settings (e.g., providing clinical advice to patients), additional factors such as the potential harmfulness of the output and the level of empathy conveyed are also critical aspects of performance evaluation.

We used the original terms from the included studies to represent evaluation purposes, consolidating terms with the same meaning (e.g., reliability and stability) into a single category. A bar chart was used to illustrate the frequency of each evaluation purpose. For each evaluation metric, we summarized its purpose and the best reported value in relation to the corresponding clinical tasks. Additionally, for NLP metrics that assess the similarity between the LLM's output and the ground truth, we identified correlated metrics that require human judgment for validation.

2.3.5. Generative LLM Challenges

An introduction to the existing challenges is provided in the **Supplementary Material**. This review summarizes the current challenges of applying generative LLMs to EHR data, including bias, common errors, hallucinations, and other issues identified in the included studies.

3. Results

3.1. Study Selection Results

As illustrated in **Figure 1**, our Boolean search initially yielded 9,323 articles. After removing 1,910 duplicates and excluding 1,085 articles published before 2023, we had 6,328 articles remaining for screening. Following a thorough screening of titles, abstracts, and full papers, we ultimately included 76 eligible studies for further analysis.

3.2. Analysis Result

3.2.1. Data

The distribution of data sizes is shown in **Figure 1(A)**. Detailed information on data size and data source is provided in **Supplementary Table S2**. We found that 38 studies (50.0%) had a data size of less than 100, 21 studies (27.8%) had a data size between 100 and 1,000, and 17 studies (21.5%) had a data size greater than 1,000.

3.2.2. Clinical Domain

The distribution of clinical specialties is shown in **Figure 1(B)**. The top three clinical tasks identified are radiology (15.8%), general—no specific specialty (14.5%), and internal medicine (14.5%). Detailed information on the clinical tasks of each included study is provided in **Supplementary Table S2**.

3.2.3. Generative LLM-Related Information

3.2.3.1. Included Generative LLMs

Figure 1(C) illustrates the frequency of each generative LLM used in the included studies. ChatGPT was the most frequently used, appearing in 48 studies (63.2%), followed by GPT-4 in 32 studies (42.1%). Google Gemini was the next most common, appearing in 6 studies (7.9%).

3.2.3.2. Prompting Methods

Table 1 summarizes the findings on prompting methods used in the included studies. Nine studies employed advanced prompting techniques, while the remaining 67 studies used zero-shot prompting. Four studies specifically discussed strategies for crafting zero-shot prompts to enhance LLM performance on specific clinical tasks. Among the advanced prompting techniques, three studies used few-shot prompting, two studies employed chain-of-thought prompting, two studies utilized soft prompting, one study involved retrieval-augmented generation (RAG), and one study used another LLM to assist with prompt generation.

All studies that used advanced prompting techniques reported improvements in LLM performance due to prompt engineering, though the significance of these improvements varied depending on the clinical task. For instance, one study found that a combination of few-shot prompting, chain-of-thought, and RAG increased the LLM's F1 score by 5% to 15% on a subset of 100 reports when detecting speech recognition errors in radiology reports[31] Another study combined soft prompting with LLM-aided prompting (using an LLM to help generate prompts) for clinical note summarization and found that LLM-aided prompting improved ROUGE-1 by 1% to 3%, ROUGE-2 by 2% to 4%, and ROUGE-L by 1% to 2%, while soft prompting reduced response variability by up to 43%.[32]

3.2.3.3. Fine-Tuning Methods

Table 2 summarizes the studies that fine-tuned LLMs for specific clinical domains or tasks. Of the 76 included papers, only 8 (10.5%) involved LLM fine-tuning. Regarding the fine-tuning methods, three studies used parameter-efficient fine-tuning (PEFT) with Low-Rank Adaptation (LoRA)[33], two used PEFT-Quantized LoRA (QLoRA),[34] two utilized DeepSpeed[35] for full parameter tuning, and one paper did not specify the fine-tuning technique.

While six of the eight papers reported that fine-tuning improved performance on clinical tasks, three studies noted potential drawbacks of fine-tuning, such as 1) catastrophic forgetting[36] and 2) low relevance between the fine-tuning data and the LLM's application domain or task.[37,38] Additionally, it was observed that a smaller fine-tuned model can sometimes outperform a larger base model in specific domains and tasks. For example, in differential diagnosis for PICU patients, the fine-tuned Llama-7B achieved an average quality score of 2.88, while the Llama-65B without fine-tuning achieved an average quality score of 2.65 out of 5.00[39]

3.2.3.4.Multimodal Data Fusion for LLM

Two of the included studies utilized multimodal data. In one study, different types of data were encoded and fused within the AI model itself after encoding each input data modality.[40] The other study converted various data modalities into text format before feeding the text into the model.[41] Integrating multimodal data was shown to enhance overall performance. For instance, a Llama model trained on multimodal data achieved a higher macro F1 score (22.3%) compared to a Llama model trained solely on medical notes (macro F1 = 21.8%) for disease diagnosis.[40] Similarly, using multimodal data for pre-training and fine-tuning LLMs led to better performance in diagnosing COVID-19 (accuracy = 90.3% vs. 84.1%) and prognosticating COVID-19 (accuracy = 92.8% vs. 94.9%) when compared to using text-only data.[41] Notably, the study pre-trained the LLM on Delta COVID-19 data, fine-tuned it on 1% of Omicron data, and then evaluated it on the remaining 99% of Omicron data. This also suggests that multimodal LLMs can effectively handle scenarios where training data is scarce, such as in diagnosing or prognosticating rare diseases.

3.2.4. Evaluation

Figure 1(D) presents the statistics on the evaluation methods used in the included studies. A total of 22 evaluation purposes were identified. The three most frequently used evaluation purposes were correctness (employed in 56 studies), agreement with experts or ground truth (used in 12 studies), and completeness, reliability/stability, and readability (each used in 7 studies). For assessing accuracy, confusion matrix-based metrics were the most commonly employed.

Table 3 provides a summary of all evaluation metrics used in the included studies. A total of 55 different evaluation metrics were identified, with 35 of them being NLP metrics that measure the similarity between the generative LLM's response and the gold standard response. Four studies used Spearman's correlation to examine the relationships between evaluation metrics.[37,38,42] The findings were as follows: 1) The Artificial Intelligence Performance Instrument (AIPI) correlated with the Ottawa Clinic Assessment Tool (OCAT) when managing cases in otolaryngology–head and neck surgery ($\rho = 0.495$); 2) BERTScore correlated with the quality score derived from a Likert scale when generating impressions for whole-body PET reports ($\rho = 0.474$); 3) BERTScore correlated with conciseness when summarizing patient questions and progress notes; and 4) when generating concise and accurate layperson summaries of musculoskeletal radiology reports, BERTScore and MEDCON Score correlated with correctness ($\rho = 0.17$), and BLEU correlated with completeness ($\rho = 0.225$).

3.2.5. Challenges for Applying in Real Clinical Settings

3.2.5.1.Bias

Among the included studies, only two specifically examined the bias of LLMs. One study reported that ChatGPT did not exhibit biases related to demographic factors such as age and gender when making imaging referrals.[43] However, the other study found that male patients received more appropriate responses than female patients, indicating a potential gender bias in how ChatGPT processes information.[44]

3.2.5.2.Common Errors

Several studies highlighted common errors made by LLMs. For instance, multiple studies pointed out that the LLM made more errors when diagnosing uncommon cases.[45] GPT-4 was found to sometimes miss important details when converting radiological reports into a structured format.[46] Additionally, multiple studies indicated that LLMs were not proficient in recommending appropriate treatments or examinations[32,47]. One study showed that ChatGPT often provided unnecessary treatments for 55% of patients with head and neck cases[48], and for

67%-90% of such patients in other instances.[49] Another study reported that unnecessary treatments were recommended by ChatGPT for 55% of patients with positive blood cultures,[50] and ChatGPT was more likely to suggest additional treatments compared to physicians (94.3% vs. 73.5%, $p<0.001$).[51] For rhinologic cases, the accuracy of GPT-4 in suggesting treatment strategies was only 16.7%[52]

Several studies also found that LLMs performed poorly when triaging patients. For example, when providing triage for maxillofacial trauma cases, Gemini inadequately proposed intermaxillary fixation and missed the necessity of teeth splinting in another case.[53] In the emergency department, ChatGPT provided unsafe triage in 41% of cases.[54] Furthermore, LLMs may omit critical information in patient history. When tasked with improving the readability of clinical notes, LLMs were found to omit the history of present illness and procedures in 52.1% of cases[55] ChatGPT, relying on static data, lacks the ability to assess individual patient history when diagnosing conditions like bacterial tonsillitis.[56] Additionally, studies found that patients had difficulty understanding ChatGPT's responses, and the readability of ChatGPT-generated responses to patient-submitted questions was not as good as those produced by dermatology physicians.[57] ChatGPT also struggles with diagnosing complex diseases due to ambiguous symptoms.[56,58] Two studies noted that ChatGPT might overlook compositional information and adjacent relationships of nodules when diagnosing tumor-related diseases[59,60]

3.2.5.3.Hallucinations

LLMs can sometimes generate hallucinations, producing content that is inaccurate or fabricated. When identifying clinical phenotypes within the complex notes of rare genetic disease patients, GPT-J may invent Human Phenotype Ontology (HPO) IDs, even after fine-tuning and using few-shot prompting.[61] In another instance, when identifying confidential content in clinical notes, 87% of the 306 excerpts proposed by ChatGPT from a note containing confidential information included hallucinations.[62] Additionally, when extracting the clinical factor of neoadjuvant chemotherapy status in breast cancer patients, ChatGPT provided a yes or no answer despite the pathology report lacking any relevant information.[63] While summarizing clinical letters, ChatGPT occasionally inserted sentences that were not present in the original letter, such as "please know that we are here to support you every step of the way" and "your expertise and insights are invaluable".[64] ChatGPT has also been known to fabricate patient names when generating structured thyroid ultrasound reports from unstructured ultrasound reports.[60] Moreover, when improving the readability of radiology reports, ChatGPT incorrectly stated that a patient had a lateral ligament complex tear when the lateral ligament complex was intact or claimed there was no fracture of the lateral malleolus when a fracture was indeed present.[65]

3.2.5.4.Other Challenges

Three included studies noted that patients felt uncomfortable with ChatGPT's impersonal tone during consultations, and they often found it difficult to understand ChatGPT's responses.[57,64,66]

4. Discussion

Recent publications on LLMs in healthcare underscore their evolving role and the wide range of potential applications. Numerous reviews have been published to summarize the field's development, with a general consensus that LLMs hold significant promise in clinical settings, assisting physicians in tasks such as answering patient questions and improving the readability of medical documents. However, challenges remain in applying LLMs in clinical environments. Omiye et al. reviewed LLM applications in medicine and identified major challenges, including bias, data privacy concerns, and the unpredictability of outputs.[8] Clusmann et al. emphasized that hallucinations are a significant obstacle,[9] , while Acharya et al. attempted to address this issue by fine-tuning LLMs, only to find that this process led to the loss of previously acquired knowledge.[36] Additionally, Wornow highlighted the lack of benchmarks and standardized evaluation techniques necessary to ensure LLM reliability in real clinical settings.[10] Unlike existing reviews, our study extends previous work by summarizing the techniques, challenges, and opportunities for applying LLMs to real EHR data in actual clinical settings to improve patient care—an area where corresponding studies remain rare due to privacy concerns.

Our review found that out of the 76 included studies, 67 relied on zero-shot prompting. Among the studies that employed a specific prompting strategy, only four evaluated its effectiveness, and all four reported that using prompting strategies improved performance. For instance, one study noted that soft prompting reduced the variability of LLM outputs when summarizing clinical notes.[32] However, recent research has suggested that prompting strategies, such as few-shot prompting, do not always lead to performance improvements.[26,67] This may be due to the fact that prompting strategies can increase the length of a prompt, and a longer prompt might negatively impact the LLM's performance.[68] Furthermore, the use of prompting strategies is often limited by the maximum length constraints of an LLM. Therefore, further testing of prompting strategies in specific clinical tasks and specialties is necessary to validate their effectiveness in real clinical settings.

Unlike prompting strategies, fine-tuning an LLM enables it to fully leverage all labeled training data without concerns about maximum length limits. However, fine-tuning proprietary LLMs (e.g., ChatGPT and GPT-4) is often restricted, and fine-tuning open-source LLMs requires expensive hardware. Fortunately, one included study demonstrated that fine-tuning a smaller language model can outperform an unfine-tuned LLM.[39] Techniques like LoRA and QLoRA allow researchers to fine-tune LLMs with more affordable hardware,[33,34] and the DeepSpeed algorithm can accelerate the fine-tuning process.[35] It's important to note, however, that fine-tuning may not enhance performance if the fine-tuning dataset lacks sufficient text relevant to the specific task.[37] For instance, if the goal is to optimize LLM performance in analyzing PET reports, it would be more effective to fine-tune the model using a large corpus of PET reports rather than a mix of different clinical notes. Therefore, in clinical settings, we recommend fine-tuning a smaller, open-source language model with a domain- and task-specific corpus to achieve better results in specific domains and tasks.

Incorporating multimodal clinical data enhances the performance of clinical decision support systems and enables LLM-based support for rare diseases.[41] Notably, several studies mentioned that LLMs struggle with handling rare diseases, likely due to the limited information about rare conditions in the training data. We also observed that only two of the included studies utilized multimodal data, indicating a need for more research in the future focused on leveraging LLMs and multimodal EHR data to address challenges in rare disease diagnosis and management.

Our review indicates a pressing need for standardized evaluation metrics and solutions to reduce the labor-intensive nature of human evaluation. We found that different studies often use varying metrics to achieve the same evaluation goals, highlighting the necessity of establishing standardized metrics for each evaluation purpose to benchmark performance consistently. Although expert evaluation is considered the gold standard, it is impractical for physicians to thoroughly review all LLM outputs for performance evaluation.[37] As data sizes increase, manual review becomes increasingly labor-intensive, costly, and time-consuming. This challenge may also explain why 50% of the included studies used a small data size of less than 100 samples. Fortunately, some

studies have identified correlations between automated similarity metrics and human subjective evaluation metrics. For instance, BLEU scores showed a Spearman's correlation coefficient of 0.225 with physicians' preferences for completeness when summarizing clinical texts.[38] Therefore, developing standardized objective metrics for each evaluation purpose is crucial for ensuring fair and effective evaluations. Additionally, further investigation is needed to explore how automated evaluation metrics can replace human subjective evaluation, particularly when dealing with large datasets.

Overall, while ChatGPT and similar LLMs present innovative potential in medical diagnostics and patient interaction, significant challenges and biases persist. Although only a limited number of studies have examined biases in large language models, there is evidence of gender-related bias in ChatGPT's responses. For instance, one study found no bias in imaging referrals related to age or gender,[60] while another highlighted a gender bias, with male patients receiving more appropriate responses than female patients.[61] This finding underscores the need for ongoing evaluation and mitigation of biases in LLMs to ensure equitable and unbiased healthcare information for all users. Additionally, these models often struggle with diagnosing uncommon cases,[62] accurately converting radiological reports,[63] and recommending appropriate treatments.[51,64-48] The tendency to suggest unnecessary treatments and the high rate of unsafe triage decisions[69, 70] further highlight the risks associated with relying on LLMs in clinical settings. LLMs may also omit critical patient history details[71, 72] and provide responses that are difficult for patients to understand.[73] Their inadequacies in handling complex diseases and ambiguous symptoms,[72, 74] as well as the potential for overlooking essential information,[75, 76] suggest that LLMs currently lack the reliability needed for high-stakes medical decision-making. These findings emphasize the need for continuous improvement and careful integration of LLMs into healthcare to mitigate risks and enhance patient safety.

The findings regarding hallucinations in generative LLMs like GPT-J and ChatGPT highlight a critical issue that limits the reliability and safety of these models in clinical settings. Hallucinations, which involve the generation of fabricated or incorrect information, are particularly concerning when LLMs are used for tasks requiring high accuracy and trust, such as in healthcare. For example, GPT-J's tendency to create fictitious Human Phenotype Ontology (HPO) IDs when addressing rare genetic diseases suggests that even advanced fine-tuning and prompting techniques may not fully eliminate the risk of hallucinations.[77] This issue not only compromises the accuracy of diagnoses but also risks misleading healthcare providers who might rely on these outputs in decision-making processes.

Moreover, ChatGPT has exhibited similar issues across various medical applications. The model has been shown to insert non-existent information into clinical notes and summaries, fabricating phrases intended to convey support or even creating fictitious patient names when generating structured reports.[78-80] These errors are far from benign; they have the potential to cause real harm, especially if clinicians act on incorrect information. The implications of these hallucinations are significant. For instance, misstating the condition of the lateral ligament complex or incorrectly identifying the presence of fractures can lead to inappropriate treatment plans and delayed care.[81] Such inconsistencies and inaccuracies call into question the reliability of LLMs in clinical environments, emphasizing the need for their cautious use, particularly in high-stakes situations.

Beyond technical inaccuracies, the impersonal tone of ChatGPT's responses and the challenges patients face in understanding these responses further diminish the effectiveness of LLMs in patient interaction.[73, 80, 82] The lack of empathy and clarity in communication can erode patient trust and satisfaction, both of which are critical components of effective healthcare delivery. While LLMs hold significant promise for enhancing healthcare through automation and data processing, the risks posed by hallucinations and communication challenges must be addressed. Until these issues are resolved, the integration of LLMs into healthcare should proceed with caution, ensuring that human oversight remains central to patient care.

Our review has several strengths and weaknesses. Given the rapid development in the field, the volume of articles on LLMs in healthcare is substantial. We identified studies published since 2023 from two databases

(PubMed and Web of Science) and thoroughly screened each article based on our eligibility criteria. Every included paper was analyzed in depth, and we provided detailed summaries. However, a limitation of our review is that our Boolean search was conducted in May 2024, so papers published online after this date were not included.

5. Conclusion

We conducted a systematic literature review to summarize articles that use LLMs to analyze real EHR data for improving patient care. We found that the application of prompt engineering and fine-tuning techniques is still relatively rare. Additionally, only two studies utilized LLMs with multimodal EHR data, and they demonstrated that incorporating multimodal data can enhance decision-making performance and enable more accurate diagnoses of rare diseases. Several limitations of LLMs were identified, making them currently unsuitable for widespread use in clinical practice. These limitations include the lack of standardized evaluation methods, impersonal tone and low readability in responses to patient questions, and the presence of biases and hallucinations in generated responses.

Future research should focus on exploring more prompt engineering and fine-tuning approaches tailored to specific clinical domains and tasks to optimize their use. Additionally, important future directions include standardizing evaluation metrics, mitigating bias and hallucinations, and applying LLMs to multimodal data to further improve their performance.

Acknowledgments
This study was funded by NIH-NIA R44AG081006, NIH-NLM 1R01LM014239, and NIH-NIA R01AG080429

Tables

Table 1. Studies with Prompting Strategy.

Prompting Strategy	Summary of Findings Related to Prompting
Zero-shot	Decision support: The capability of clinical decision making has been tested on specialties of dermatology[66], emergency medicine[43,54,69–74], gastroenterology[75,76], internal medicine[50,77–85], neurology[86], obstetrics and gynecology[87], oncology[48,51,63,88–92], ophthalmology[45,93–95], orthopedic[64,96,97], otolaryngology[42,49,52,56,98–100], pathology[101], pediatrics[39], radiology[37,46,60,102–108], surgery[53,109], urology[44], and general (no specific specialty)[36,47,55,57,110–112]. Zero-shot prompting allows LLM to get promising performances on clinical tasks. For example, when classifying the emergency department patient's acuity levels, LLM achieved 89% accuracy.[74] Clinical document summarization: Zero-shot enabled LLMs to summarize clinical notes, but the performance of LLMs needs improvement. For example, when summarizing discharge summary, 52.1% inaccuracies were due to omitting key information such as history of present illness and procedures.[55] Phenotyping patients: Only one included paper talked about using LLM to phenotype patients. The authors found LLMs achieved 95% positive predictive value when phenotyping patients with postpartum hemorrhage.[87] **Tips mentioned in corresponding papers for improving the LLM's performance on specific tasks:** Clinical note summarization: In prompt, do not limit the length of generated answer.[80] Patient question summarization: In prompt, limit the length of generated answer. Without this instruction, the model might generate lengthy outputs, occasionally even longer than the input text.[38] Answer questions regarding glaucoma diagnosis and treatment: Instructing the model to respond as a clinician in an ophthalmology note format.[95] Radiology reports simplification: Request simplification at a specific grade level[105]
Few shot	Diagnosing of benign and malignant bone tumors: Few-shot (two shots) improves ChatGPT's performances: accuracy from 0.73 to 0.87; sensitivity from 0.95 to 0.99; specificity from 0.58 to 0.73; AUROC from 0.72 to 0.83[58]
	Identifying clinical phenotypes within the intricate notes of rare genetic disease patients: No mention of the effect of prompting strategy. Only mentioned that literature said few-shot learning and chain-of-thought was effective. On dataset BiolarkGSC, the best-performing LLM (GPT-J fine-tuned with training data and prompted with few-shot) achieved 83.2% F1 score. On dataset ID-68, the best performing LLM (GPT-3 fine-tuned with training data and prompted with few-shot) achieved 81.6% F1 score.[61]
	Identifying the presence of confidential content in clinical notes: No mention of the effect of prompting strategy. Using few-shot prompting, ChatGPT achieved 97% sensitivity, 18% specificity, and 34 positive predictive value.[62]
	Clinical text summarization: More examples in the prompt would lead to a better performance, but the improvement becomes less obvious when adding more and more examples. For example, on MIMIC-CXR dataset, zero-shot achieved a MEDCON score of less than 20. Using 2, 8, 32, and 128 examples led to improved MEDCON scores of 43, 50, 52, and 53 respectively.[38]
Few-shot + chain-of-thought	Classification tasks related to COVID-19 diagnosis: No mention of the effect of prompting strategy. Only mentioned that literature said few-shot learning was effective. The model achieved 96.3% accuracy.[41]
	Converting free-text clinical notes into structured data: No mention of the effect of prompting strategy. ChatGPT-3.5, GPT-4 demonstrated the ability to extract pathological classifications with an overall accuracy of 89% and 94% separately (primary tumor classification: 87% and 91%; regional lymph node involvement classification: 91% and 95%; pathology stage identification: 76% and 89%; histological diagnosis: 99% and 99%). In lung cancer dataset, outperforming the performance of two traditional NLP methods. In pediatric osteosarcoma dataset, ChatGPT-3.5 accurately classified both grades and margin status with accuracy of 98.6% and 100% respectively.[59]
Few-shot + chain-of-thought + retrieval augmented generation (RAG)	Automatic detection of speech recognition errors in radiology reports: Optimized prompts increased the models' F1 scores by 5%–15% on the subset of 100 reports assessed by three independent raters. For GPT-3.5-turbo, F1 score increased from 59.1% to 73% for clinically significant errors and 32.2% to 45% for not clinically significant errors. F1 score for GPT-4 increased from 86.9% to 91% for clinically significant errors and from 94.3% to 97% for not clinically significant errors. Further increases were achieved for text-davinci-003 (72% to 82% F1 score on clinically significant errors, 60% to 74.3% F1 score on not clinically significant errors), Llama-v2–70B-chat (58.8% to 67% F1 score, 31.2% to 41%), and Bard (34.8% to 44% F1 score, 33.2% to 39%).[31]
Soft prompting	Disease diagnosis: EHR-KnowGen-III showed decreased performance across various evaluation metrics (micro f1: from 29.7% to 28.3%; macro f1: from 23.8% to 21.1%; accuracy: from 38.1% to 34.6%), emphasizing the importance of incorporating soft prompting for multimodal learning.[40]
Soft prompting + using ChatGPT to generate prompt	Clinical notes summarization: LLM-aided prompt improved similarity score, and soft prompting improved reproducibility of LLM. Table 4 in the Chuan et al. (2024)'s paper highlighted detailed improvements quantitatively. For example, for Flan-T5 model, adding LLM-aided prompt improved ROUGE-1 F1 score from 50.7% to a range of 52%-53%, and adding soft prompting reduced the standard deviation from 0.8% to 0.5%; adding LLM-aided prompt improved ROUGH-2 F1 score from 35.8% to 39%-40%.[32]

Table 2. Studies with LLM Fine-Tuning.

LLMs That Was Fine-Tuned	Fine-Tuning Algorithm	Fine-Tuning Hardware	Summary of Findings Related to Fine-Tuning
Llama	Full Parameter - DeepSpeed	2*NVIDIA A100 GPUs	Disease diagnosis: No mention regarding the comparison of fine-tuned model and the original model. Fine-tuned Llama achieved 34.9% accuracy, 22.3% macro f1 score, and 28.7% micro f1 score.[40]
Llama	PEFT-LoRA	1*Nvidia RTX A6000 GPU	Predicting diagnosis-related group for hospitalized patients: No mention regarding the comparison of fine-tuned model and the original model. DRG-LLaMA -7B model exhibited a noteworthy macro-averaged F1 score of 0.327, a top-1 prediction accuracy of 52.0%, and a macro-averaged Area Under the Curve (AUC) of 0.986.[111] 1) Larger base-model led to better fine-tuned performance. Top-1 diagnosis accuracy of fine-tuned Llama-13B achieved 54.6%, while that of fine-tuned 7B model achieved 53.9%. 2) Longer input context from the fine-tuning data led to better fine-tuned performance. For fine-tuned Llama-13B, when the max input token size was 340, the top-1 diagnosis accuracy was 49.9%, but the accuracy was increased to 54.6% when the max token size was 1024.
Llama 2; FLAN-T5; FLAN-UL2; Vicuna Alpaca	PEFT-QLoRA	1*NVIDIA Quadro RTX 8000	Clinical text summarization: Fine-tuned FLAN-T5 improved MEDCON score from 5 to a range of 26-69 on four datasets.[38] 1) QLoRA FLAN-T5 was the best-performing fine-tuned open-source model. It achieved a MEDCON score of 59 on Open-i data, 38 on MIMIC-CXR data, 26 on MIMIC-III data, and 46 on patient questions data. 2) QLoRA typically outperformed ICL with the better models (FLAN-T5 and Llama-2); given a sufficient number of in-context examples (from 1 to 64), however, all models surpassed even the best QLoRA fine-tuned model, FLAN-T5, in at least one dataset. 3) An LLM fine-tuned with domain specific data performed worse than the original model. For example, when Alpaca achieved a BLEU value of 30, Med-Alpaca only reached 20. This highlights a distinction between domain adaptation and task adaptation.
GPT-3; GPT-J; Falcon; Llama	PEFT-QLoRA	Open AI's Cloud Resources	Phenotype recognition in clinical notes: No quantitative comparison between models before and after fine-tuning. Fine-tuned GPT-3 achieved the best performance of 81.6% F1 score on one dataset, and fine-tuned GPT-J performed the best on the other dataset (83.2% F1 score) [61]
BART; PEGASUS; T5; FLAN-T5; BioBART; Clinical-T5; GPT2; OPT; Llama; Alpaca	PEFT-LoRA for Llama and Alpaca; full parameter tuning for other models.	At least two NVDIA A100 GPUs	Generating personalized impressions for whole-body PET reports: Biomedical domain pretrained LLMs did not outperform their base models. Specifically, domain-specific fine-tuned BART model reduced the accuracy from 75.3% to 73.9%. This could be attributed to two reasons. First, our large training set diminished the benefits of medical-domain adaptation. Second, the corpora, such as MIMIC-III and PubMed, likely had limited PET related content, making pretraining less effective for our task.[37]
Llama 2	No mention	No mention	Predicting opioid use disorder (OUD), substance use disorder (SUD), and Diabetes: Fine-tuned Llama 2 achieved 92%, 93%, 74%, and 88% AUROC on four datasets for predicting SUD. Fine-tuned Llama 2 achieved 95%, 72%, 73%, and 98% AUROC on four datasets for predicting OUD. Fine-tuned Llama 2 achieved 88%, 76%, 64%, and 94% AUROC on four datasets for predicting diabetes.[36] 1) An experiment of changing instructions suggests that fine-tuning on our datasets might have induced catastrophic forgetting particularly when dealing with a large volume of data. 2) Fine-tuned Llama 2 outperformed Llama 2 without fine-tuning on diabetes prediction (AUROC increased from 50% to 88%).
Llama 2-7B; BioGPT-Large	Full Parameter - DeepSpeed	4*A40 Nvidia GPUs	Differential Diagnoses in PICU Patients: Fine-tuned model outperformed original model, but a smaller LM fine-tuned using domain-specific notes outperformed much larger models trained on general-domain data.[66] Specifically: 1) Fine-tuned Llama-7B achieved an average quality score of 2.88, while Llama-65B without fine-tuning achieved an average quality score of 2.65. 2) Fine-tuned BioGPT-Large had an average score of 2.78, while BioGPT-Large without fine-tuning had a mean score of 2.02
BART; GPT; MedLM	PEFT-LoRA	No mention	Early detection of gout flares based on nurses' chief complaint notes in the emergency department: No comparison between models before and after fine-tuning. Fine-tuned BART model (BioBART) performed the best, which achieved 0.73 and 0.67 F1 score on datasets GOUT-CC-2019-CORPUS and GROUT-CC-2020-CORPUS.[73]

Table 3. Summary of Evaluation Matrices.

General Matrices				
Evaluation Matrices	**Evaluation Purpose**	**Best Reported Performance**	**Clinical Task**	**Clinical Specialty**
Confusion Matrices-Based Scores	Correctness	100%	Diagnosing glaucoma based on specific clinical case descriptions[45]	Ophthalmology
			Generating radiology reports from concise imaging findings[102]	Radiology
			Accelerating review of historic echocardiogram reports[79]	Internal Medicine
			Interpret symptoms and management of common cardiac conditions[81]	Internal Medicine
			The diagnosis management of bacterial tonsillitis[56]	Otolaryngology
			Classifying margin status for lung cancer[59]	Oncology
Average Word Count Reduction Percentage + Recall	Balance Between Conciseness and Completeness	Average Word Count Reduction Percentage=47% when Recall=90%	Summarizing radiology reports into structured format[46]	Radiology
Self-Designed Human Evaluation (e.g., Likert-Scale)	Correctness	89.6%	Generating concise and accurate layperson summaries of musculoskeletal radiology reports[103]	Radiology
	Completeness	94.1%	Generating concise and accurate layperson summaries of musculoskeletal radiology reports[103]	Radiology
	Conciseness	12%	Summarizing patient questions and progress notes[38]	No specific specialty
	Harmfulness*	2%	Proposing a comprehensive management plan (suspected/confirmed diagnosis, workup, antibiotic therapy, source control, follow-up) for patients with positive blood cultures[50]	Internal Medicine
	Readability	80%	Generating radiology reports from concise imaging findings[102]	Radiology
	Quality	89%	Impression generation for whole-body PET reports[37]	Radiology
	Appropriateness	58.5%	Diagnosing and suggest examinations/treatments for urology patients (subgroups that had the best performance: non-oncology)[44]	Urology
	Satisfactory	80%	Proposing a comprehensive management plan (suspected/confirmed diagnosis, workup, antibiotic therapy, source control, follow-up) for patients with positive blood cultures[50]	Internal Medicine
	Reliability/Stability	70%	Predicting treatments for patients with aortic stenosis[85]	Internal Medicine
	Preference over Human	81%	Summarizing clinical text[38]	No specific Specialty
	Level of Empathy	61.4%	Generating high-quality responses to patient-submitted questions in the patient portal[66]	Dermatology
	Hallucination Rate*	4%	Improving the readability of foot and ankle orthopedic radiology reports[108]	Radiology
	Utility	81.6%	Impression generation for whole-body PET reports[37]	Radiology
	Relevancy	40%	Simplifying radiological MRI findings of the knee joint[107]	Radiology
Artificial Intelligence Performance Instrument (AIPI)	Other Performance	15.1/20.0	Managing cases in otolaryngology–head and neck surgery[42]	Otolaryngology
QAMAI Tool	Other Performance	18.4/30	Providing Triage for Maxillofacial Trauma Cases	Surgery
Ottawa Clinic Assessment Tool	Other Performance	3.88/5.00	Recommending differential diagnosis for laryngology and head and neck (Recommending differential diagnosis) cases[49]	Otolaryngology
DISCERN	Quality	15/35	Diagnosing and suggest examinations/treatments for urology patients (subgroups that had the best performance: oncology, emergency, and male)[44]	Urology
Root Mean Square Error	Error	2.96	Measuring the angle of correction for high tibial osteotomy[97]	Orthopedic
Flesch Reading Ease	Readability	72.7%	Improving the readability of foot and ankle orthopedic radiology reports[65]	Radiology
Flesch-Kincaid Reading Grade Level*	Readability	6.2	Summarizing discharge summary[55]	No specific specialty
Average of Gunning Fog,	Readability	7.5	Summarizing X-Ray report[105]	Radiology

Flesch–Kincaid Grade Level, Automated Readability, Coleman–Liau*				
Patient Education Materials Assessment Tool	Understandability	81%	Summarizing discharge summary[55]	No specific specialty
Cohen's Kappa	Reliability/Stability	κ=1.0	Head and neck oncological board decisions: deciding on neoadjuvant chemotherapy and chemoradiotherapy treatment	Oncology
	Agreement with Expert or Ground Truth	κ=0.727	Predicting the dichotomized modified Rankin Scale (mRS) score at 3 months post-thrombectomy[86]	Neurology
Cronbach's α	Agreement with Expert or Ground Truth	0.754	Managing otolaryngology cases[98]	Otolaryngology
Mann-Whitney U test	Agreement with Expert or Ground Truth	κ=0.770	Providing number of additional examinations when managing otolaryngological cases[42]	Otolaryngology
Spearman's Coefficient	Reliability/Stability	0.999	Considering the patient's symptoms and physical findings reported by practitioners when managing otolaryngology cases[98]	Otolaryngology
Percentage of Getting the Same Response to Identical Queries	Reliability/Stability	100%	Predicting hemoglobinopathies from a patient's laboratory results of CBC and ferritin values[84]	Internal Medicine
Agreement Percentage	Agreement with Expert or Ground Truth	80%	Determining disease severity for acute ulcerative colitis presentations in the setting of an emergency department[77]	Gastroenterology
Global Quality Scale	Quality	4.2	Analyzing retinal detachment cases and suggesting the best possible surgical planning[94]	Ophthalmology
Fleiss Kappa	Reliability/Stability	κ=0.786	Colonoscopy recommendations for colorectal cancer rescreening and surveillance[76]	Gastroenterology

Similarity Measurements for Generative NLP Models

Evaluation Matrices	**Correlated Evaluation (If Reported) Measured by Spearman's Coefficient**			**Performance**		
	Reported Coefficient	**Evaluation and Task**	**Clinical Specialty**	**Best Reported Value**	**Task**	**Clinical Specialty**
BLEU	0.412	Quality score of Impression generation for whole-body PET reports[37]	Radiology	24.7	Impression generation for whole-body PET reports[37]	Radiology
	0.225	Completeness of clinical text summarization[38]	No Specific Specialty			
	0.125	Correctness of clinical text summarization[38]				
	0.075	Conciseness of clinical text summarization[38]				
BLEU-2				74.5	Generating a comprehensive and coherent medical report of a given medical image from COVID-19 data[41]	Internal Medicine
BLEU-3				67.8		
BLEU-4				63.2		
ROUGE-1	0.402	Quality score of Impression generation for whole-body PET reports[37]	Radiology	57.29	Clinical notes summarization[32]	No Specific Specialty
ROUGE-2	0.379	Quality score of Impression generation for whole-body PET reports[37]	Radiology	44.32	Clinical notes summarization[32]	No Specific Specialty
ROUGE-L	0.22	Completeness of clinical text summarization[38]	No Specific Specialty	68.5	Generating a comprehensive and coherent medical report of a given medical image from COVID-19 data[41]	Internal Medicine
	0.16	Correctness of clinical text summarization[38]				
	0.19	Conciseness of clinical text summarization[38]				
	0.398	Quality score of Impression generation for whole-body PET	Radiology			

		reports[37]				
BERTScore-Precision				86.57	Clinical notes summarization[32]	No Specific Specialty
BERTScore-Recall				87.14		
BERTScore-F1	0.18	Completeness of clinical text summarization[38]	No Specific Specialty	89.4	Summarizing longitudinal aneurysm reports[104]	Radiology
	0.18	Correctness of clinical text summarization[38]				
	0.24	Conciseness of clinical text summarization[38]				
	0.407	Quality score of Impression generation for whole-body PET reports[37]	Radiology			
MEDCON	0.125	Completeness of clinical text summarization[38]		64.9	Clinical text summarization[38]	No Specific Specialty
	0.175	Correctness of clinical text summarization[38]				
	0.15	Conciseness of clinical text summarization[38]				
CIDEr	0.194	Quality score of Impression generation for whole-body PET reports[37]	Radiology	97.5	Generating a comprehensive and coherent medical report of a given medical image from COVID-19 data[41]	Internal Medicine
BARTScore+PET	0.568			-1.46	Impression generation for whole-body PET reports[37]	Radiology
PEGASUSScore+PET	0.563			-1.44		
T5Score+PET	0.542			-1.41		
UniEval	0.501			0.78		
BARTScore	0.474			-3.05		
CHRF	0.433			42.2		
Moverscore	0.420			0.607		
ROUGE-WE-1	0.403			54.8		
ROUGE-LSUM	0.397			50.8		
ROUGE-WE-2	0.396			40.7		
METEOR	0.388			0.279		
ROUGE-WE-3	0.385			42.5		
RedGraph	0.384			0.397		
PRISM	0.369			-3.24		
ROUGE-3	0.345			20.5		
S^3-pyr	0.302			0.71		
S^3-resp	0.301			0.79		
Stats-novel trigram	0.292			0.99		
Stats-density	0.280			6.51		
BLANC	0.165			0.131		
Stats-compression	0.145			8.36		
SUPERT	0.082			0.557		
Stats-coverage	0.078			8.36		
SummaQA	0.075			0.180		

* Lower value represents better performance.

Figures

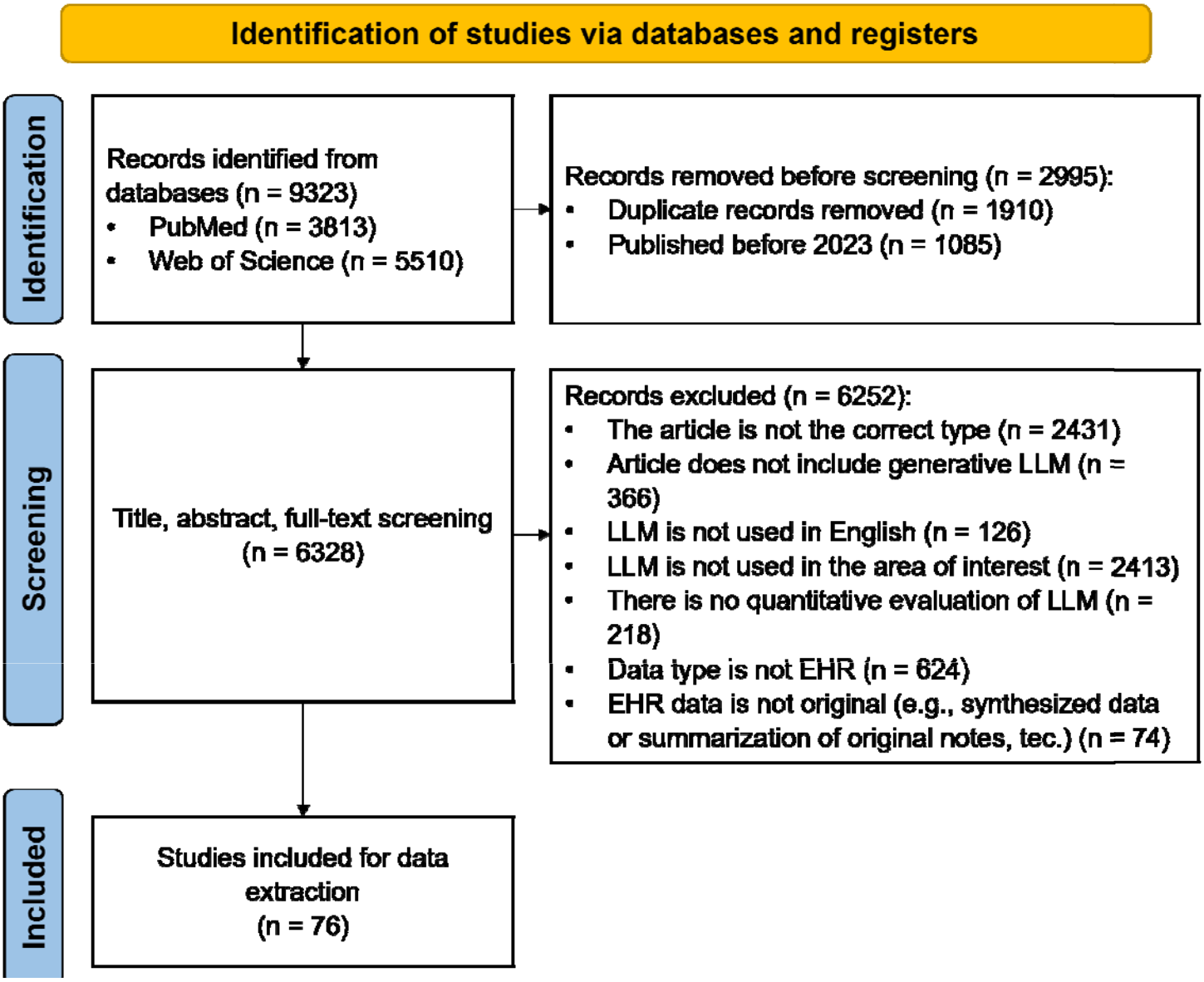

Figure 1. PRISMA Flow Chart for Eligibility Screening. We initially identified 9,323 studies from PubMed and Web of Science. After deduplication and excluding articles published before 2023, we included 6,328 studies for eligibility screening. Ultimately, 6,252 studies did not meet the inclusion criteria, leaving 76 studies for detailed analysis.

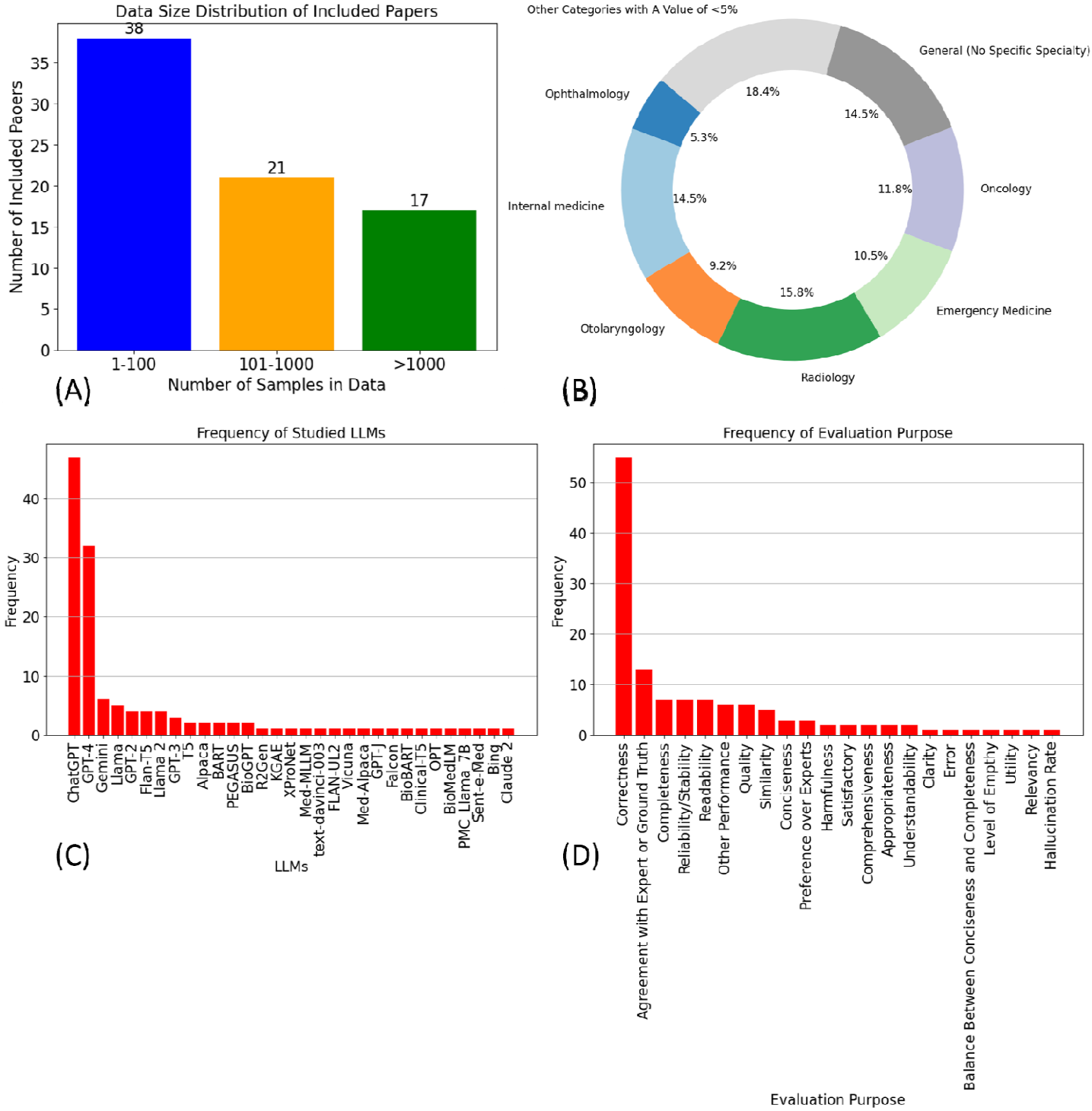

Figure 2. **Result Summarization.** (A) shows the data size distribution of the included studies, with most studies (38 out of 76, 50%) having less than 100 samples. (B) displays the distribution of clinical specialties among the included studies, with radiology being the most frequently studied specialty, accounting for 15.8% of the studies. (C) is a bar plot highlighting the frequency of studied LLMs, with ChatGPT being the most frequently used, appearing in 48 studies. (D) represents the frequency of evaluation purposes in the included studies, with correctness being the most common, evaluated in 56 studies.

References

1. Chang Y, Wang X, Wang J, et al. A Survey on Evaluation of Large Language Models. *ACM Trans Intell Syst Technol*. 2024;15(3):39:1-39:45. doi:10.1145/3641289
2. Vaswani A, Shazeer N, Parmar N, et al. Attention is All you Need. In: *Advances in Neural Information Processing Systems*. Vol 30. Curran Associates, Inc.; 2017. Accessed April 11, 2024. https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
3. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Published online May 24, 2019. doi:10.48550/arXiv.1810.04805
4. Kasneci E, Sessler K, Küchemann S, et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learn Individ Differ*. 2023;103:102274. doi:10.1016/j.lindif.2023.102274
5. OpenAI, Achiam J, Adler S, et al. GPT-4 Technical Report. Published online December 18, 2023. doi:10.48550/arXiv.2303.08774
6. Touvron H, Martin L, Stone K, et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv.org. July 18, 2023. Accessed January 17, 2024. https://arxiv.org/abs/2307.09288v2
7. Gemini Team, Anil R, Borgeaud S, et al. Gemini: A Family of Highly Capable Multimodal Models. Published online April 2, 2024. doi:10.48550/arXiv.2312.11805
8. Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. Large Language Models in Medicine: The Potentials and Pitfalls. *Ann Intern Med*. 2024;177(2):210-220. doi:10.7326/M23-2772
9. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in medicine. *Commun Med*. 2023;3(1):1-8. doi:10.1038/s43856-023-00370-1
10. Wornow M, Xu Y, Thapa R, et al. The shaky foundations of large language models and foundation models for electronic health records. *Npj Digit Med*. 2023;6(1):1-10. doi:10.1038/s41746-023-00879-8
11. Benary M, Wang XD, Schmidt M, et al. Leveraging Large Language Models for Decision Support in Personalized Oncology. *JAMA Netw Open*. 2023;6(11):e2343689. doi:10.1001/jamanetworkopen.2023.43689
12. Jeblick K, Schachtner B, Dexl J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. *Eur Radiol*. Published online October 5, 2023. doi:10.1007/s00330-023-10213-1
13. Haver HL, Ambinder EB, Bahl M, Oluyemi ET, Jeudy J, Yi PH. Appropriateness of Breast Cancer Prevention and Screening Recommendations Provided by ChatGPT. *Radiology*. 2023;307(4):e230424. doi:10.1148/radiol.230424
14. Hamilton Z, Naffakh N, Reizine NM, et al. Relevance and accuracy of ChatGPT-generated NGS reports with treatment recommendations for oncogene-driven NSCLC. *J Clin Oncol*. 2023;41(16_suppl):1555-1555. doi:10.1200/JCO.2023.41.16_suppl.1555
15. Levkovich I, Elyoseph Z. Suicide Risk Assessments Through the Eyes of ChatGPT-3.5 Versus ChatGPT-4: Vignette Study. *JMIR Ment Health*. 2023;10:e51232. doi:10.2196/51232
16. Hu X, Ran AR, Nguyen TX, et al. What can GPT-4 do for Diagnosing Rare Eye Diseases? A Pilot Study. *Ophthalmol Ther*. 2023;12(6):3395-3402. doi:10.1007/s40123-023-00789-8
17. Pillai J, Pillai K. Accuracy of generative artificial intelligence models in differential diagnoses of familial Mediterranean fever and deficiency of Interleukin-1 receptor antagonist. *J Transl Autoimmun*. 2023;7:100213. doi:10.1016/j.jtauto.2023.100213
18. Hirosawa T, Mizuta K, Harada Y, Shimizu T. Comparative Evaluation of Diagnostic Accuracy Between Google Bard and Physicians. *Am J Med*. 2023;136(11):1119-1123.e18. doi:10.1016/j.amjmed.2023.08.003
19. Caruccio L, Cirillo S, Polese G, Solimando G, Sundaramurthy S, Tortora G. Can ChatGPT provide intelligent diagnoses? A comparative study between predictive models and ChatGPT to define a new medical diagnostic bot. *Expert Syst Appl*. 2024;235:121186. doi:10.1016/j.eswa.2023.121186
20. Nanji K, Yu CW, Wong TY, et al. Evaluation of postoperative ophthalmology patient instructions from ChatGPT and Google Search. *Can J Ophthalmol*. 2024;59(1):e69-e71. doi:10.1016/j.jcjo.2023.10.001
21. Liu HY, Alessandri Bonetti M, De Lorenzi F, Gimbel ML, Nguyen VT, Egro FM. Consulting the Digital Doctor: Google Versus ChatGPT as Sources of Information on Breast Implant-Associated Anaplastic Large Cell Lymphoma and Breast Implant Illness. *Aesthetic Plast Surg*. 2024;48(4):590-607. doi:10.1007/s00266-023-03713-4
22. Hristidis V, Ruggiano N, Brown EL, Ganta SRR, Stewart S. ChatGPT vs Google for Queries Related to Dementia and Other Cognitive Decline: Comparison of Results. *J Med Internet Res*. 2023;25:e48966. doi:10.2196/48966
23. Responsible use of MIMIC data with online services like GPT. Accessed August 11, 2024. https://physionet.org/news/post/gpt-responsible-use
24. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. 2016;3(1):160035. doi:10.1038/sdata.2016.35

25. Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data*. 2023;10(1):1. doi:10.1038/s41597-022-01899-x
26. Du X, Novoa-Laurentiev J, Plasaek JM, et al. Enhancing Early Detection of Cognitive Decline in the Elderly: A Comparative Study Utilizing Large Language Models in Clinical Notes. *medRxiv*. Published online May 6, 2024:2024.04.03.24305298. doi:10.1101/2024.04.03.24305298
27. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71. doi:10.1136/bmj.n71
28. Beltagy I, Peters ME, Cohan A. Longformer: The Long-Document Transformer. Published online December 2, 2020. doi:10.48550/arXiv.2004.05150
29. Jiang LY, Liu XC, Nejatian NP, et al. Health system-scale language models are all-purpose prediction engines. *Nature*. 2023;619(7969):357-362. doi:10.1038/s41586-023-06160-y
30. Yang X, Chen A, PourNejatian N, et al. A large language model for electronic health records. *Npj Digit Med*. 2022;5(1):1-9. doi:10.1038/s41746-022-00742-2
31. Schmidt RA, Seah JCY, Cao K, Lim L, Lim W, Yeung J. Generative Large Language Models for Detection of Speech Recognition Errors in Radiology Reports. *Radiol Artif Intell*. 2024;6(2):e230205. doi:10.1148/ryai.230205
32. Chuang YN, Tang R, Jiang X, Hu X. SPeC: A Soft Prompt-Based Calibration on Performance Variability of Large Language Model in Clinical Notes Summarization. *J Biomed Inform*. 2024;151:104606. doi:10.1016/j.jbi.2024.104606
33. Hu EJ, Shen Y, Wallis P, et al. LoRA: Low-Rank Adaptation of Large Language Models. Published online October 16, 2021. doi:10.48550/arXiv.2106.09685
34. Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L. QLoRA: Efficient Finetuning of Quantized LLMs. *Adv Neural Inf Process Syst*. 2023;36:10088-10115.
35. Rasley J, Rajbhandari S, Ruwase O, He Y. DeepSpeed: System Optimizations Enable Training Deep Learning Models with Over 100 Billion Parameters. In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD ’20. Association for Computing Machinery; 2020:3505-3506. doi:10.1145/3394486.3406703
36. Acharya A, Shrestha S, Chen A, et al. Clinical risk prediction using language models: benefits and considerations. *J Am Med Inform Assoc*. Published online February 27, 2024:ocae030. doi:10.1093/jamia/ocae030
37. Tie X, Shin M, Pirasteh A, et al. Personalized Impression Generation for PET Reports Using Large Language Models. *J Imaging Inform Med*. 2024;37(2):471-488. doi:10.1007/s10278-024-00985-3
38. Van Veen D, Van Uden C, Blankemeier L, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nat Med*. 2024;30(4):1134-1142. doi:10.1038/s41591-024-02855-5
39. Akhondi-Asl A, Yang Y, Luchette M, Burns JP, Mehta NM, Geva A. Comparing the Quality of Domain-Specific Versus General Language Models for Artificial Intelligence-Generated Differential Diagnoses in PICU Patients. *Pediatr Crit Care Med J Soc Crit Care Med World Fed Pediatr Intensive Crit Care Soc*. 2024;25(6):e273-e282. doi:10.1097/PCC.0000000000003468
40. Niu S, Ma J, Bai L, Wang Z, Guo L, Yang X. EHR-KnowGen: Knowledge-enhanced multimodal learning for disease diagnosis generation. *Inf Fusion*. 2024;102:102069. doi:10.1016/j.inffus.2023.102069
41. Liu F, Zhu T, Wu X, et al. A medical multimodal large language model for future pandemics. *Npj Digit Med*. 2023;6(1):226. doi:10.1038/s41746-023-00952-2
42. Lechien JR, Naunheim MR, Maniaci A, et al. Performance and Consistency of ChatGPT-4 Versus Otolaryngologists: A Clinical Case Series. *Otolaryngol Neck Surg*. 2024;170(6):1519-1526. doi:10.1002/ohn.759
43. Rosen S, Saban M. Evaluating the reliability of ChatGPT as a tool for imaging test referral: a comparative study with a clinical decision support system. *Eur Radiol*. Published online October 13, 2023. doi:10.1007/s00330-023-10230-0
44. Cocci A, Pezzoli M, Lo Re M, et al. Quality of information and appropriateness of ChatGPT outputs for urology patients. *Prostate Cancer Prostatic Dis*. 2024;27(1):103-108. doi:10.1038/s41391-023-00705-y
45. Delsoz M, Raja H, Madadi Y, et al. The Use of ChatGPT to Assist in Diagnosing Glaucoma Based on Clinical Case Reports. *Ophthalmol Ther*. 2023;12(6):3121-3132. doi:10.1007/s40123-023-00805-x
46. Mallio CA, Bernetti C, Sertorio AC, Zobel BB. ChatGPT in radiology structured reporting: analysis of ChatGPT-3.5 Turbo and GPT-4 in reducing word count and recalling findings. *Quant Imaging Med Surg*. 2024;14(2):2096102-2092102. doi:10.21037/qims-23-1300
47. Bužančić I, Belec D, Držaić M, et al. Clinical decision making in benzodiazepine deprescribing by healthcare providers *vs* . AI assisted approach. *Br J Clin Pharmacol*. 2024;90(3):662-674. doi:10.1111/bcp.15963
48. Lechien JR, Chiesa-Estomba CM, Baudouin R, Hans S. Accuracy of ChatGPT in head and neck oncological board decisions: preliminary findings. *Eur Arch Otorhinolaryngol*. 2024;281(4):2105-2114. doi:10.1007/s00405-023-08326-w

49. Lechien JR, Georgescu BM, Hans S, Chiesa-Estomba CM. ChatGPT performance in laryngology and head and neck surgery: a clinical case-series. *Eur Arch Otorhinolaryngol*. 2024;281(1):319-333. doi:10.1007/s00405-023-08282-5
50. Maillard A, Micheli G, Lefevre L, et al. Can Chatbot Artificial Intelligence Replace Infectious Diseases Physicians in the Management of Bloodstream Infections? A Prospective Cohort Study. *Clin Infect Dis*. Published online October 12, 2023:ciad632. doi:10.1093/cid/ciad632
51. Gebrael G, Sahu KK, Chigarira B, et al. Enhancing Triage Efficiency and Accuracy in Emergency Rooms for Patients with Metastatic Prostate Cancer: A Retrospective Analysis of Artificial Intelligence-Assisted Triage Using ChatGPT 4.0. *Cancers*. 2023;15(14):3717. doi:10.3390/cancers15143717
52. Radulesco T, Saibene AM, Michel J, Vaira LA, Lechien JR. ChatGPT-4 performance in rhinology: A clinical case series. *Int Forum Allergy Rhinol*. 2024;14(6):1123-1130. doi:10.1002/alr.23323
53. Frosolini A, Catarzi L, Benedetti S, et al. The Role of Large Language Models (LLMs) in Providing Triage for Maxillofacial Trauma Cases: A Preliminary Study. *Diagnostics*. 2024;14(8):839. doi:10.3390/diagnostics14080839
54. Fraser H, Crossland D, Bacher I, Ranney M, Madsen T, Hilliard R. Comparison of Diagnostic and Triage Accuracy of Ada Health and WebMD Symptom Checkers, ChatGPT, and Physicians for Patients in an Emergency Department: Clinical Data Analysis Study. *JMIR MHealth UHealth*. 2023;11:e49995. doi:10.2196/49995
55. Zaretsky J, Kim JM, Baskharoun S, et al. Generative Artificial Intelligence to Transform Inpatient Discharge Summaries to Patient-Friendly Language and Format. *JAMA Netw Open*. 2024;7(3):e240357. doi:10.1001/jamanetworkopen.2024.0357
56. Mayo-Yáñez M, González-Torres L, Saibene AM, et al. Application of ChatGPT as a support tool in the diagnosis and management of acute bacterial tonsillitis. *Health Technol*. 2024;14(4):773-779. doi:10.1007/s12553-024-00858-3
57. Samala AD, Rawas S. Generative AI as Virtual Healthcare Assistant for Enhancing Patient Care Quality. *Int J Online Biomed Eng IJOE*. 2024;20(05):174-187. doi:10.3991/ijoe.v20i05.45937
58. Yang F, Yan D, Wang Z. Large-Scale assessment of ChatGPT's performance in benign and malignant bone tumors imaging report diagnosis and its potential for clinical applications. *J Bone Oncol*. 2024;44:100525. doi:10.1016/j.jbo.2024.100525
59. Huang J, Yang DM, Rong R, et al. A critical assessment of using ChatGPT for extracting structured data from clinical notes. *Npj Digit Med*. 2024;7(1):1-13. doi:10.1038/s41746-024-01079-8
60. Jiang H, Xia S, Yang Y, et al. Transforming free-text radiology reports into structured reports using ChatGPT: A study on thyroid ultrasonography. *Eur J Radiol*. 2024;175:111458. doi:10.1016/j.ejrad.2024.111458
61. Yang J, Liu C, Deng W, et al. Enhancing phenotype recognition in clinical notes using large language models: PhenoBCBERT and PhenoGPT. *Patterns*. 2024;5(1):100887. doi:10.1016/j.patter.2023.100887
62. Rabbani N, Brown C, Bedgood M, et al. Evaluation of a Large Language Model to Identify Confidential Content in Adolescent Encounter Notes. *JAMA Pediatr*. 2024;178(3):308-310. doi:10.1001/jamapediatrics.2023.6032
63. Choi HS, Song JY, Shin KH, Chang JH, Jang BS. Developing prompts from large language model for extracting clinical information from pathology and ultrasound reports in breast cancer. *Radiat Oncol J*. 2023;41(3):209-216. doi:10.3857/roj.2023.00633
64. Stoneham AC, Walker LC, Newman MJ, Nicholls A, Avis D. Can artificial intelligence make elective hand clinic letters easier for patients to understand? *J Hand Surg Eur Vol*. Published online April 20, 2024:17531934241246479. doi:10.1177/17531934241246479
65. From jargon to clarity: Improving the readability of foot and ankle radiology reports with an artificial intelligence large language model - ClinicalKey. Accessed July 14, 2024. https://www-clinicalkey-com.ezp-prod1.hul.harvard.edu/#!/content/playContent/1-s2.0-S1268773124000262?returnurl=null&referrer=null
66. Reynolds K, Nadelman D, Durgin J, et al. Comparing the quality of ChatGPT- and physician-generated responses to patients' dermatology questions in the electronic medical record. *Clin Exp Dermatol*. 2024;49(7):715-718. doi:10.1093/ced/llad456
67. Chen Q, Du J, Hu Y, et al. Large language models in biomedical natural language processing: benchmarks, baselines, and recommendations. Published online January 20, 2024. doi:10.48550/arXiv.2305.16326
68. Levy M, Jacoby A, Goldberg Y. Same Task, More Tokens: the Impact of Input Length on the Reasoning Performance of Large Language Models. Published online July 10, 2024. doi:10.48550/arXiv.2402.14848
69. Berg HT, Van Bakel B, Van De Wouw L, et al. ChatGPT and Generating a Differential Diagnosis Early in an Emergency Department Presentation. *Ann Emerg Med*. 2024;83(1):83-86. doi:10.1016/j.annemergmed.2023.08.003
70. Paslı S, Şahin AS, Beşer MF, Topçuoğlu H, Yadigaroğlu M, İmamoğlu M. Assessing the precision of artificial intelligence in ED triage decisions: Insights from a study with ChatGPT. *Am J Emerg Med*. 2024;78:170-175. doi:10.1016/j.ajem.2024.01.037

71. Huang T, Socrates V, Gilson A, et al. Identifying incarceration status in the electronic health record using large language models in emergency department settings. *J Clin Transl Sci*. 2024;8(1):e53. doi:10.1017/cts.2024.496
72. Haim GB, Braun A, Eden H, et al. AI in the ED: Assessing the efficacy of GPT models vs. physicians in medical score calculation. *Am J Emerg Med*. 2024;79:161-166. doi:10.1016/j.ajem.2024.02.016
73. Oliveira LL, Jiang X, Babu AN, Karajagi P, Daneshkhah A. Effective Natural Language Processing Algorithms for Early Alerts of Gout Flares from Chief Complaints. *Forecasting*. 2024;6(1):224-238. doi:10.3390/forecast6010013
74. Williams CYK, Zack T, Miao BY, et al. Use of a Large Language Model to Assess Clinical Acuity of Adults in the Emergency Department. *JAMA Netw Open*. 2024;7(5):e248895. doi:10.1001/jamanetworkopen.2024.8895
75. Y W, Y H, Ir N, et al. Validation of GPT-4 for clinical event classification: A comparative analysis with ICD codes and human reviewers. *J Gastroenterol Hepatol*. Published online April 16, 2024. doi:10.1111/jgh.16561
76. Chang PW, Amini MM, Davis RO, et al. ChatGPT4 Outperforms Endoscopists for Determination of Postcolonoscopy Rescreening and Surveillance Recommendations. *Clin Gastroenterol Hepatol Off Clin Pract J Am Gastroenterol Assoc*. Published online May 9, 2024:S1542-3565(24)00429-4. doi:10.1016/j.cgh.2024.04.022
77. Levartovsky A, Ben-Horin S, Kopylov U, Klang E, Barash Y. Towards AI-Augmented Clinical Decision-Making: An Examination of ChatGPT's Utility in Acute Ulcerative Colitis Presentations. *Am J Gastroenterol*. 2023;118(12):2283-2289. doi:10.14309/ajg.0000000000002483
78. Al Tibi G, Alexander M, Miller S, Chronos N. A Retrospective Comparison of Medication Recommendations Between a Cardiologist and ChatGPT-4 for Hypertension Patients in a Rural Clinic. *Cureus*. 2024;16(3):e55789. doi:10.7759/cureus.55789
79. Vaid A, Duong SQ, Lampert J, et al. Local large language models for privacy-preserving accelerated review of historic echocardiogram reports. *J Am Med Inform Assoc*. Published online April 30, 2024:ocae085. doi:10.1093/jamia/ocae085
80. Li Q, Ma H, Song D, Bai Y, Zhao L, Xie K. Early prediction of sepsis using chatGPT-generated summaries and structured data. *Multimed Tools Appl*. Published online February 9, 2024. doi:10.1007/s11042-024-18378-7
81. Harskamp RE, De Clercq L. Performance of ChatGPT as an AI-assisted decision support tool in medicine: a proof-of-concept study for interpreting symptoms and management of common cardiac conditions (AMSTELHEART-2). *Acta Cardiol*. 2024;79(3):358-366. doi:10.1080/00015385.2024.2303528
82. van Nuland M, Snoep JD, Egberts T, Erdogan A, Wassink R, van der Linden PD. Poor performance of ChatGPT in clinical rule-guided dose interventions in hospitalized patients with renal dysfunction. *Eur J Clin Pharmacol*. 2024;80(8):1133-1140. doi:10.1007/s00228-024-03687-5
83. Kopitar L, Fister I, Stiglic G. Using Generative AI to Improve the Performance and Interpretability of Rule-Based Diagnosis of Type 2 Diabetes Mellitus. *Information*. 2024;15(3):162. doi:10.3390/info15030162
84. Kurstjens S, Schipper A, Krabbe J, Kusters R. Predicting hemoglobinopathies using ChatGPT. *Clin Chem Lab Med CCLM*. 2024;62(3):e59-e61. doi:10.1515/cclm-2023-0885
85. Baladrón C, Sevilla T, Carrasco-Moraleja M, Gómez-Salvador I, Peral-Oliveira J, San Román JA. Assessing the accuracy of ChatGPT as a decision support tool in cardiology. *Rev Espanola Cardiol Engl Ed*. 2024;77(5):433-435. doi:10.1016/j.rec.2023.11.011
86. Pedro T, Sousa JM, Fonseca L, et al. Exploring the use of ChatGPT in predicting anterior circulation stroke functional outcomes after mechanical thrombectomy: a pilot study. *J Neurointerventional Surg*. Published online March 7, 2024:jnis-2024-021556. doi:10.1136/jnis-2024-021556
87. Alsentzer E, Rasmussen MJ, Fontoura R, et al. Zero-shot Interpretable Phenotyping of Postpartum Hemorrhage Using Large Language Models. Published online June 1, 2023. doi:10.1101/2023.05.31.23290753
88. Choo JM, Ryu HS, Kim JS, et al. Conversational artificial intelligence (chatGPT^{TM}) in the management of complex colorectal cancer patients: early experience. *ANZ J Surg*. 2024;94(3):356-361. doi:10.1111/ans.18749
89. Lukac S, Dayan D, Fink V, et al. Evaluating ChatGPT as an adjunct for the multidisciplinary tumor board decision-making in primary breast cancer cases. *Arch Gynecol Obstet*. 2023;308(6):1831-1844. doi:10.1007/s00404-023-07130-5
90. Vela Ulloa J, King Valenzuela S, Riquoir Altamirano C, Urrejola Schmied G. Artificial intelligence-based decision-making: can ChatGPT replace a multidisciplinary tumour board? *Br J Surg*. 2023;110(11):1543-1544. doi:10.1093/bjs/znad264
91. Haemmerli J, Sveikata L, Nouri A, et al. ChatGPT in glioma adjuvant therapy decision making: ready to assume the role of a doctor in the tumour board? *BMJ Health Care Inform Online*. 2023;30(1):e100775. doi:10.1136/bmjhci-2023-100775
92. Sushil M, Zack T, Mandair D, et al. A comparative study of large language model-based zero-shot inference and task-specific supervised classification of breast cancer pathology reports. *J Am Med Inform Assoc*. Published online June 20, 2024:ocae146. doi:10.1093/jamia/ocae146

93. Carlà MM, Gambini G, Baldascino A, et al. Large language models as assistance for glaucoma surgical cases: a ChatGPT vs. Google Gemini comparison. *Graefes Arch Clin Exp Ophthalmol Albrecht Von Graefes Arch Klin Exp Ophthalmol*. Published online April 4, 2024. doi:10.1007/s00417-024-06470-5
94. Carlà MM, Gambini G, Baldascino A, et al. Exploring AI-chatbots' capability to suggest surgical planning in ophthalmology: ChatGPT versus Google Gemini analysis of retinal detachment cases. *Br J Ophthalmol*. Published online March 6, 2024. doi:10.1136/bjo-2023-325143
95. Huang AS, Hirabayashi K, Barna L, Parikh D, Pasquale LR. Assessment of a Large Language Model's Responses to Questions and Cases About Glaucoma and Retina Management. *JAMA Ophthalmol*. 2024;142(4):371-375. doi:10.1001/jamaophthalmol.2023.6917
96. Daher M, Koa J, Boufadel P, Singh J, Fares MY, Abboud JA. Breaking barriers: can ChatGPT compete with a shoulder and elbow specialist in diagnosis and management? *JSES Int*. 2023;7(6):2534-2541. doi:10.1016/j.jseint.2023.07.018
97. Dhivakaran G, Saggi SS, Rahmatullah BARH. Pre-operative Planning of High Tibial Osteotomy With ChatGPT: Are We There Yet? *Cureus*. 2024;16(2). doi:10.7759/cureus.54858
98. Lechien JR, Maniaci A, Gengler I, Hans S, Chiesa-Estomba CM, Vaira LA. Validity and reliability of an instrument evaluating the performance of intelligent chatbot: the Artificial Intelligence Performance Instrument (AIPI). *Eur Arch Otorhinolaryngol*. 2024;281(4):2063-2079. doi:10.1007/s00405-023-08219-y
99. Makhoul M, Melkane AE, Khoury PE, Hadi CE, Matar N. A cross-sectional comparative study: ChatGPT 3.5 versus diverse levels of medical experts in the diagnosis of ENT diseases. *Eur Arch Oto-Rhino-Laryngol Off J Eur Fed Oto-Rhino-Laryngol Soc EUFOS Affil Ger Soc Oto-Rhino-Laryngol - Head Neck Surg*. 2024;281(5):2717-2721. doi:10.1007/s00405-024-08509-z
100. Sievert M, Conrad O, Mueller SK, et al. Risk stratification of thyroid nodules: Assessing the suitability of ChatGPT for text-based analysis. *Am J Otolaryngol*. 2024;45(2):104144. doi:10.1016/j.amjoto.2023.104144
101. Oon ML, Syn NL, Tan CL, Tan K, Ng S. Bridging bytes and biopsies: A comparative analysis of ChatGPT and histopathologists in pathology diagnosis and collaborative potential. *Histopathology*. 2024;84(4):601-613. doi:10.1111/his.15100
102. Nakaura T, Yoshida N, Kobayashi N, et al. Preliminary assessment of automated radiology report generation with generative pre-trained transformers: comparing results to radiologist-generated reports. *Jpn J Radiol*. 2024;42(2):190-200. doi:10.1007/s11604-023-01487-y
103. Kuckelman IJ, Wetley K, Yi PH, Ross AB. Translating musculoskeletal radiology reports into patient-friendly summaries using ChatGPT-4. *Skeletal Radiol*. 2024;53(8):1621-1624. doi:10.1007/s00256-024-04599-2
104. Chien A, Tang H, Jagessar B, et al. AI-Assisted Summarization of Radiologic Reports: Evaluating GPT3davinci, BARTcnn, LongT5booksum, LEDbooksum, LEDlegal, and LEDclinical. *Am J Neuroradiol*. 2024;45(2):244-248. doi:10.3174/ajnr.A8102
105. Doshi R, Amin KS, Khosla P, Bajaj SS, Chheang S, Forman HP. Quantitative Evaluation of Large Language Models to Streamline Radiology Report Impressions: A Multimodal Retrospective Analysis. *Radiology*. 2024;310(3):e231593. doi:10.1148/radiol.231593
106. Wang WH, Wang SY, Huang JY, et al. An investigation study on the interpretation of ultrasonic medical reports using OpenAI 's GPT □3.5□turbo model. *J Clin Ultrasound*. 2024;52(2):105-111. doi:10.1002/jcu.23590
107. Schmidt S, Zimmerer A, Cucos T, Feucht M, Navas L. Simplifying radiologic reports with natural language processing: a novel approach using ChatGPT in enhancing patient understanding of MRI results. *Arch Orthop Trauma Surg*. 2023;144(2):611-618. doi:10.1007/s00402-023-05113-4
108. Butler JJ, Harrington MC, Tong Y, et al. From jargon to clarity: Improving the readability of foot and ankle radiology reports with an artificial intelligence large language model. *Foot Ankle Surg Off J Eur Soc Foot Ankle Surg*. 2024;30(4):331-337. doi:10.1016/j.fas.2024.01.008
109. Gakuba C, Le Barbey C, Sar A, et al. Evaluation of ChatGPT in Predicting 6-Month Outcomes After Traumatic Brain Injury. *Crit Care Med*. 2024;52(6):942-950. doi:10.1097/CCM.0000000000006236
110. Shea YF, Lee CMY, Ip WCT, Luk DWA, Wong SSW. Use of GPT-4 to Analyze Medical Records of Patients With Extensive Investigations and Delayed Diagnosis. *JAMA Netw Open*. 2023;6(8):e2325000. doi:10.1001/jamanetworkopen.2023.25000
111. Wang H, Gao C, Dantona C, Hull B, Sun J. DRG-LLaMA□: tuning LLaMA model to predict diagnosis-related group for hospitalized patients. *Npj Digit Med*. 2024;7(1):1-9. doi:10.1038/s41746-023-00989-3
112. Chiu WHK, Ko WSK, Cho WCS, Hui SYJ, Chan WCL, Kuo MD. Evaluating the Diagnostic Performance of Large Language Models on Complex Multimodal Medical Cases. *J Med Internet Res*. 2024;26:e53724. doi:10.2196/53724