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

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

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
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). 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, NYUTron, GatorTron); 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.

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 DeepSpeed35 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 and 2) low relevance between the fine-tuning data and the LLM’s application domain or task. 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.

### 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. The other study converted various data modalities into text format before feeding the text into the model. 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. 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. 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. 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. 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.

#### 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. GPT-4 was found to sometimes miss important details when converting radiological reports into a structured format. Additionally, multiple studies indicated that LLMs were not proficient in recommending appropriate treatments or examinations. One study showed that ChatGPT often provided unnecessary treatments for 55% of patients with head and neck cases, and for
67%-90% of such patients in other instances. Another study reported that unnecessary treatments were recommended by ChatGPT for 55% of patients with positive blood cultures, and ChatGPT was more likely to suggest additional treatments compared to physicians (94.3% vs. 73.5%, p<0.001). For rhinologic cases, the accuracy of GPT-4 in suggesting treatment strategies was only 16.7%

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. In the emergency department, ChatGPT provided unsafe triage in 41% of cases. 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. ChatGPT, relying on static data, lacks the ability to assess individual patient history when diagnosing conditions like bacterial tonsillitis. 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. ChatGPT also struggles with diagnosing complex diseases due to ambiguous symptoms. Two studies noted that ChatGPT might overlook compositional information and adjacent relationships of nodules when diagnosing tumor-related diseases.

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. 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. 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. 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”. ChatGPT has also been known to fabricate patient names when generating structured thyroid ultrasound reports from unstructured ultrasound reports. 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.

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.
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. Clusmann et al. emphasized that hallucinations are a significant obstacle, 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. Additionally, Wornow highlighted the lack of benchmarks and standardized evaluation techniques necessary to ensure LLM reliability in real clinical settings. 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. However, recent research has suggested that prompting strategies, such as few-shot prompting, do not always lead to performance improvements. 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. 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. Techniques like LoRA and QLoRA allow researchers to fine-tune LLMs with more affordable hardware, and the DeepSpeed algorithm can accelerate the fine-tuning process. 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. 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. 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. 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. 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, while another highlighted a gender bias, with male patients receiving more appropriate responses than female patients. 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, accurately converting radiological reports, and recommending appropriate treatments. The tendency to suggest unnecessary treatments and the high rate of unsafe triage decisions further highlight the risks associated with relying on LLMs in clinical settings. LLMs may also omit critical patient history details and provide responses that are difficult for patients to understand. Their inadequacies in handling complex diseases and ambiguous symptoms, as well as the potential for overlooking essential information, 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. 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. 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. 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. 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
No mention of the effect of prompting strategy. Using few-shot (two shots) improves ChatGPT's performances: accuracy from 0.73 to 0.87; Optimized prompts increased the models' F1 scores by 5%–15% on few-shot + chain-of-thought. Few-shot Diagnosing of benign and malignant bone tumors

Zero-shot

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.

Tips mentioned in corresponding papers for improving the LLM's performance on specific tasks:

Clinical note summarization: In prompt, do not limit the number of generated answer.

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.

Answer questions regarding glaucoma diagnosis and treatment: Instructing the model to respond as a clinician in an ophthalmology note format.

Radiology reports simplification: Request simplification at a specific grade level

Few-shot

Identifying clinical phenotypes within the intricate notes of rare genetic diseases: 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.

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.

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.

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.

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.

New-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%).

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.

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 ROUGE-2 F1 score from 35.8% to 39%-40%.

---

**Tables**

**Table 1. Studies with Prompting Strategy.**

<table>
<thead>
<tr>
<th>Prompting Strategy</th>
<th>Summary of Findings Related to Prompting</th>
</tr>
</thead>
</table>
| Zero-shot          | Decision support: The capability of clinical decision making has been tested on specialties of dermatology, emergency medicine, gastroenterology, internal medicine, neurology, obstetrics and gynecology, oncology, ophthalmology, orthopedics, otolaryngology, pathology, pediatrics, radiology, surgery, urology, and general (no specific specialty). 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. 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. Phentyping patients: Only one included paper talked about using LLM to phenotype patients. The authors found LLMs achieved 95% positive predictive value when phenotyping patients when postpartum hemorrhage. Tips mentioned in corresponding papers for improving the LLM’s performance on specific tasks: Clinical note summarization: In prompt, do not limit the number of generated answer. 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. Answer questions regarding glaucoma diagnosis and treatment: Instructing the model to respond as a clinician in an ophthalmology note format. Radiology reports simplification: Request simplification at a specific grade level |**

| Few shot           | Diagnosis 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. Identifying clinical phenotypes within the intricate notes of rare genetic diseases. 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. 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. 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. |**

| 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. 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. |**

| 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%). |**

| 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. |**

| 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 ROUGE-2 F1 score from 35.8% to 39%-40%. |
### Table 2. Studies with LLM Fine-Tuning.

<table>
<thead>
<tr>
<th>LLMs That Was Fine-Tuned</th>
<th>Fine-Tuning Algorithm</th>
<th>Fine-Tuning Hardware</th>
<th>Summary of Findings Related to Fine-Tuning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Llama</td>
<td>Full Parameter - DeepSpeed</td>
<td>2* NVIDIA A100 GPUs</td>
<td>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.</td>
</tr>
</tbody>
</table>
| Llama                    | PEFF-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.  
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 | PEFF-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.  
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 | PEFF-QLoRA | Open AI’s Cloud Resources | Phenotype recognition in clinical notes: No quantitative comparison between models before and after fine-tuning. Fine-tuned GPT-J 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). |
| BART; PEGASUS; T5; FLAN-T5; BioBART; Clinical-T5; GPT2; OPT; Llama; Alpaca | PEFF-LoRA for Llama and Alpaca; full parameter tuning for other models. | At least two NVIDIA 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. |
| 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.  
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 LLM fine-tuned using domain-specific notes outperformed much larger models trained on general-domain data. 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 | PEFF-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. |
### Table 3. Summary of Evaluation Matrices.

<table>
<thead>
<tr>
<th>Evaluation Matrices</th>
<th>Evaluation Purpose</th>
<th>Best Reported Performance</th>
<th>Clinical Task</th>
<th>Clinical Specialty</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>General Matrices</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Confusion Matrices-Based Scores</td>
<td>Correctness</td>
<td>100%</td>
<td>Diagnosing glaucoma based on specific clinical case descriptions41</td>
<td>Ophthalmology</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Generating radiology reports from concise imaging findings102</td>
<td>Radiology</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Accelerating review of historic echocardiogram reports29</td>
<td>Internal Medicine</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Interpret symptoms and management of common cardiac conditions81</td>
<td>Internal Medicine</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>The diagnosis management of bacterial tonsillitis86</td>
<td>Otolaryngology</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Classifying margin status for lung cancer79</td>
<td>Oncology</td>
</tr>
<tr>
<td></td>
<td>Correctness</td>
<td>100%</td>
<td>Summarizing clinical text76</td>
<td>No specific Specialty</td>
</tr>
<tr>
<td></td>
<td>Completeness</td>
<td>94.1%</td>
<td>Summarizing radiology reports into structured format46</td>
<td>Radiology</td>
</tr>
<tr>
<td></td>
<td>Conciseness</td>
<td>12%</td>
<td>Summarizing patient questions and progress notes28</td>
<td>No specific Specialty</td>
</tr>
<tr>
<td></td>
<td>Harmfulness*</td>
<td>2%</td>
<td>Proposing a comprehensive management plan (suspected/confirmed diagnosis, workup, antibiotic therapy, source control, follow-up) for patients with positive blood cultures50</td>
<td>Internal Medicine</td>
</tr>
<tr>
<td></td>
<td>Readability</td>
<td>80%</td>
<td>Generating radiology reports from concise imaging findings102</td>
<td>Radiology</td>
</tr>
<tr>
<td></td>
<td>Quality</td>
<td>89%</td>
<td>Impression generation for whole-body PET reports37</td>
<td>Radiology</td>
</tr>
<tr>
<td></td>
<td>Appropriateness</td>
<td>58.5%</td>
<td>Diagnosing and suggest examinations/treatments for urology patients (subgroups that had the best performance: non-oncology)44</td>
<td>Urology</td>
</tr>
<tr>
<td></td>
<td>Satisfactory</td>
<td>80%</td>
<td>Proposing a comprehensive management plan (suspected/confirmed diagnosis, workup, antibiotic therapy, source control, follow-up) for patients with positive blood cultures50</td>
<td>Internal Medicine</td>
</tr>
<tr>
<td></td>
<td>Reliability/Stability</td>
<td>70%</td>
<td>Predicting treatments for patients with aortic stenosis65</td>
<td>Internal Medicine</td>
</tr>
<tr>
<td></td>
<td>Preference over Human</td>
<td>81%</td>
<td>Summarizing clinical text76</td>
<td>No specific Specialty</td>
</tr>
<tr>
<td></td>
<td>Level of Empathy</td>
<td>61.4%</td>
<td>Generating high-quality responses to patient-submitted questions in the patient portal40</td>
<td>Dermatology</td>
</tr>
<tr>
<td></td>
<td>Hallucination Rate*</td>
<td>4%</td>
<td>Improving the readability of foot and ankle orthopedic radiology reports108</td>
<td>Radiology</td>
</tr>
<tr>
<td></td>
<td>Utility</td>
<td>81.6%</td>
<td>Impression generation for whole-body PET reports37</td>
<td>Radiology</td>
</tr>
<tr>
<td></td>
<td>Relevancy</td>
<td>40%</td>
<td>Simplifying radiological MRI findings of the knee joint103</td>
<td>Radiology</td>
</tr>
<tr>
<td></td>
<td>Other Performance</td>
<td>15.1/20.0</td>
<td>Managing cases in otolaryngology–head and neck surgery42</td>
<td>Otolaryngology</td>
</tr>
<tr>
<td>Artificial Intelligence Performance Instrument (AIPI)</td>
<td>Other Performance</td>
<td>18.4/30</td>
<td>Providing Triage for Maxillofacial Trauma Cases</td>
<td>Surgery</td>
</tr>
<tr>
<td>QAMAI Tool</td>
<td>Other Performance</td>
<td>3.88/5.00</td>
<td>Recommending differential diagnosis for laryngology and head and neck (Recommending differential diagnosis) cases49</td>
<td>Otolaryngology</td>
</tr>
<tr>
<td>Ottawa Clinic Assessment Tool</td>
<td>Other Performance</td>
<td>15.35/35</td>
<td>Diagnosing and suggest examinations/treatments for urology patients (subgroups that had the best performance: oncology, emergency, and male)54</td>
<td>Urology</td>
</tr>
<tr>
<td>DISCERN</td>
<td>Quality</td>
<td>15/35</td>
<td>Measuring the angle of correction for high tibial osteotomy97</td>
<td>Orthopedic</td>
</tr>
<tr>
<td>Root Mean Square Error</td>
<td>Error</td>
<td>2.96</td>
<td>Improving the readability of foot and ankle orthopedic radiology reports65</td>
<td>Radiology</td>
</tr>
<tr>
<td>Flesch Reading Ease</td>
<td>Readability</td>
<td>72.7%</td>
<td>Summarizing discharge summary53</td>
<td>No specific Specialty</td>
</tr>
<tr>
<td>Flesch-Kincaid Reading Grade Level*</td>
<td>Readability</td>
<td>6.2</td>
<td>Summarizing X-Ray report103</td>
<td>Radiology</td>
</tr>
<tr>
<td>Average of Gunning Fog,</td>
<td>Readability</td>
<td>7.5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Similarity Measurements for Generative NLP Models

<table>
<thead>
<tr>
<th>Evaluation Matrices</th>
<th>Correlated Evaluation (if Reported) Measured by Spearman’s Coefficient</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Reported Coefficient</td>
<td>Evaluation and Task</td>
</tr>
<tr>
<td>BLEU</td>
<td>0.412</td>
<td>Quality score of Impression generation for whole-body PET reports(^{37})</td>
</tr>
<tr>
<td></td>
<td>0.225</td>
<td>Completeness of clinical text summarization(^{38})</td>
</tr>
<tr>
<td></td>
<td>0.125</td>
<td>Correctness of clinical text summarization(^{38})</td>
</tr>
<tr>
<td></td>
<td>0.075</td>
<td>Conciseness of clinical text summarization(^{38})</td>
</tr>
<tr>
<td>BLEU-2</td>
<td></td>
<td>74.5</td>
</tr>
<tr>
<td>BLEU-3</td>
<td></td>
<td>67.8</td>
</tr>
<tr>
<td>BLEU-4</td>
<td></td>
<td>63.2</td>
</tr>
<tr>
<td>ROUGE-1</td>
<td>0.402</td>
<td>Quality score of Impression generation for whole-body PET reports(^{37})</td>
</tr>
<tr>
<td>ROUGE-2</td>
<td>0.379</td>
<td>Quality score of Impression generation for whole-body PET reports(^{37})</td>
</tr>
<tr>
<td>ROUGE-L</td>
<td>0.22</td>
<td>Completeness of clinical text summarization(^{38})</td>
</tr>
<tr>
<td></td>
<td>0.16</td>
<td>Correctness of clinical text summarization(^{38})</td>
</tr>
<tr>
<td></td>
<td>0.19</td>
<td>Conciseness of clinical text summarization(^{38})</td>
</tr>
<tr>
<td></td>
<td>0.398</td>
<td>Quality score of Impression generation for whole-body PET</td>
</tr>
<tr>
<td>Metric</td>
<td>Score</td>
<td>Description</td>
</tr>
<tr>
<td>------------------------------</td>
<td>--------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>BERTScore-Precision</td>
<td>0.18</td>
<td>Completeness of clinical text summarization</td>
</tr>
<tr>
<td>BERTScore-Recall</td>
<td>0.18</td>
<td>Correctness of clinical text summarization</td>
</tr>
<tr>
<td>BERTScore-F1</td>
<td>0.24</td>
<td>Conciseness of clinical text summarization</td>
</tr>
<tr>
<td>BARTScore+PET</td>
<td>0.0568</td>
<td>Quality score of Impression generation for whole-body PET reports</td>
</tr>
<tr>
<td>PEGASUSScore+PET</td>
<td>0.0563</td>
<td></td>
</tr>
<tr>
<td>T5Score+PET</td>
<td>0.1474</td>
<td></td>
</tr>
<tr>
<td>UniEval</td>
<td>0.0433</td>
<td></td>
</tr>
<tr>
<td>Moverscore</td>
<td>0.0420</td>
<td></td>
</tr>
<tr>
<td>ROUGE-WE-1</td>
<td>0.0397</td>
<td></td>
</tr>
<tr>
<td>ROUGE-LSUM</td>
<td>0.0396</td>
<td></td>
</tr>
<tr>
<td>ROUGE-WL-2</td>
<td>0.0395</td>
<td></td>
</tr>
<tr>
<td>METEOR</td>
<td>0.0388</td>
<td></td>
</tr>
<tr>
<td>ROUGE-WL-3</td>
<td>0.0385</td>
<td></td>
</tr>
<tr>
<td>RedGraph</td>
<td>0.0384</td>
<td></td>
</tr>
<tr>
<td>PRISM</td>
<td>0.0369</td>
<td></td>
</tr>
<tr>
<td>ROUGE-3</td>
<td>0.0345</td>
<td></td>
</tr>
<tr>
<td>S^pyr</td>
<td>0.0302</td>
<td></td>
</tr>
<tr>
<td>S^resp</td>
<td>0.0301</td>
<td></td>
</tr>
<tr>
<td>Stats-novel trigram</td>
<td>0.0292</td>
<td></td>
</tr>
<tr>
<td>Stats-density</td>
<td>0.280</td>
<td></td>
</tr>
<tr>
<td>BLANC</td>
<td>0.165</td>
<td></td>
</tr>
<tr>
<td>Stats-compression</td>
<td>0.145</td>
<td></td>
</tr>
<tr>
<td>SUPERT</td>
<td>0.082</td>
<td></td>
</tr>
<tr>
<td>Stats-coverage</td>
<td>0.078</td>
<td></td>
</tr>
<tr>
<td>SummaQA</td>
<td>0.075</td>
<td></td>
</tr>
</tbody>
</table>

* Lower value represents better performance.
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.

---

**Table 1. Eligibility Screening Criteria**

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Number (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duplicate records removed</td>
<td>1,910</td>
</tr>
<tr>
<td>Published before 2023</td>
<td>1,085</td>
</tr>
<tr>
<td>The article is not the correct type</td>
<td>2,431</td>
</tr>
<tr>
<td>Article does not include generative LLM</td>
<td>366</td>
</tr>
<tr>
<td>LLM is not used in English</td>
<td>126</td>
</tr>
<tr>
<td>LLM is not used in the area of interest</td>
<td>2,413</td>
</tr>
<tr>
<td>There is no quantitative evaluation of LLM</td>
<td>218</td>
</tr>
<tr>
<td>Data type is not EHR</td>
<td>624</td>
</tr>
<tr>
<td>EHR data is not original (e.g., synthesized data or summarization of original notes, etc.)</td>
<td>74</td>
</tr>
</tbody>
</table>

**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


64. 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.exp-prod1.hul.harvard.edu/#!content/playContent/1-s2.0-S1268773124000262?returnurl=null&referrer=null


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


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


