Adapting Generative Large Language Models for Electronic Health Record Applications: A Systematic Review of Methodologies, Evaluation, and Hallucinations

Xinsong Du[1,2,3]*, Zhengyang Zhou[4]*, Yifei Wang[4]*, Ya-Wen Chuang[5,6,7], Yiming Li[1,2], Richard Yang[1,2], Wenyu Zhang[1,2,3], Xinyi Wang[1,2,3], Xinyu Chen[1,2,3], John Lian[1], Pengyu Hong[4], David W. Bates[1,2,8], 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] Department of Health Policy and Management, Harvard T.H. Chan School of Public Health, Boston, MA 02115

*Equal contribution

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
E-mail: xidu1@bwh.harvard.edu

Abstract

Objectives: The use of generative large language models (LLMs) with electronic health record (EHR) data is rapidly expanding to support clinical and research tasks. However, challenges persist, including effective adaptation to downstream tasks and the absence of standardized evaluation metrics. This systematic review synthesizes current strategies, challenges, and future directions for adapting and evaluating generative LLMs in EHR analyses and applications.

Methods: We followed the PRISMA guidelines to conduct a systematic review of articles from PubMed and Web of Science published between January 1, 2023 and November 9, 2024. Studies were included if they used generative LLMs to analyze real-world EHR data and reported quantitative performance evaluations. The review identified key clinical applications, summarized adaptation trends and strategies, and examined evaluation metrics. By synthesizing included studies, we proposed a framework for conducting future LLM research using EHR data and facilitating LLM deployment in healthcare settings.

Results: Of the 18,735 articles retrieved, 196 met our criteria. Our review revealed a surge in publications in this area, alongside a shift from ChatGPT dominance to wider adoption of newer, open-source models (e.g., GPT-4, Llama-2) and domain-adapted variants. Early research focused predominately on clinical decision support (84.6% in 2023), but later studies showed a shift toward broader applications, particularly information extraction (25.6%) and documentation (20.9%) by 2024 H2. Most studies (88.8%) did not quantitatively evaluate the LLM performance improvement strategies, with the rest twenty-two studies (11.2%) quantitatively evaluated the effectiveness of in context learning (8 studies), fine-tuning (14 studies), multimodal integration (7), and ensemble learning (2). A wide range of evaluation metrics were used across tasks, including confusion-matrix based scores, natural language processing (NLP)-based measures, human evaluation, rating scales, and task-specific indicators. in context learning and fine-tuning consistently improved LLM performance across clinical tasks, with fine-tuning showing particular benefit for domain-specific applications; however, three studies reported performance degradation in certain scenarios. Multimodal and hybrid approaches, combining text with other data types or traditional models, improved diagnostic and reasoning performance. Additionally, two studies reported bias related to demographic factors. Reported hallucinations include but were not limited to 1) creating non-existent medical conditions when doing documentation tasks, 2) lacking numeric accuracy when doing information extraction tasks. Our proposed framework includes three major stages: 1) define clinical task, 2) select LLM performance improvement strategy based on clinical tasks, and 3) evaluate performance.

Conclusion: This review highlights a clear trend toward broader, domain-adapted models and more advanced adaptation strategies. Evaluation remains inconsistent due to a lack of standardized metrics. Our proposed framework links clinical tasks with adaptation strategies and evaluation metrics, which promotes rigor, reproducibility and safe deployment of generative LLMs in healthcare.

1. Introduction

Transformer architecture, introduced by Vaswani et al. in 2017, revolutionized artificial intelligence (AI) and natural language processing (NLP) by enabling efficient parallel processing of input sequences through self-attention mechanisms. This design allows models to capture long-range dependencies and contextual relationships more effectively than prior approaches.[1] Large language models (LLMs) built on the Transformer architecture vary in design depending on their intended applications. Models like BERT,[2] Longformer,[3] NYUTron,[4] GatorTron[5] use only the encoder component of the Transformer architecture to model bidirectional context and generate dense representations that capture semantic and linguistic nuances. They perform well for language understanding tasks such as classification and named entity recognition. In contrast, generative models are designed to produce coherent text and are typically categorized as decode-only and encoder-decoder models. Decoder-only models, such as the GPT and Llama series,[2] rely on the Transformer's decoder architecture and are trained using autoregressive (next-token prediction) objectives. They are well-suited for tasks such as content generation, dialogue systems, and reasoning. Encoder-decoder models, such as T5 and BART, combine both encode and decoder components to transform input sequences into target outputs, supporting tasks like machine translation and summarization.

With the release of ChatGPT[6] in November 2022, transformer-based generative LLMs have significantly accelerated progress in NLP and AI.[1,2,7] These models, characterized by large parameter counts and autoregressive architectures, exhibit remarkable capabilities in language generation, reasoning, and contextual understanding, leading to widespread adoption in academia and industry.[8] Despite increasing interest in healthcare, real-world deployment of generative LLMs for electronic health record (EHR) analyses and applications remain limited.[9] Yet, these models offer unique potential to synthesize, summarize, and interpret unstructured clinical text, with applications in clinical decision support, documentation automation, and patient-provider communication.

However, translating these models into clinical practice requires rigorous adaptation and evaluation. Techniques such as prompt engineering, fine-tuning, and multimodal integration are essential for improving performance, aligning output with clinical objectives, and mitigating risks like hallucinations or biased predictions. Despite promising developments, a systematic understanding of how generative LLMs is adapted and assessed for EHR use remains lacking. To address this gap, we conducted a systematic review to synthesize existing methodologies, identify evidence gaps, and inform future research and implementation of generative LLMs in healthcare. We focused on studies that quantitatively evaluated performance enhancement strategies for applying generative LLMs to real-world EHR data. Based on this synthesis, we propose a structured framework that categorizes clinical task types, maps effective adaptation techniques, and recommends evaluation strategies aligned with specific clinical objectives to support future research and deployment in healthcare settings.

2. Methods

2.1. Data Sources and Searches

We followed the Preferred Reporting Items for Systematic Review and Meta-Analyses (PRISMA) guidelines and conducted systematic searches in PubMed and Web of Science (**Figure 1)**.[10] The process involved several key steps: a Boolean search, removal of duplicates, screening of studies, and data extraction. The Boolean search was conducted on November 09, 2024, using search terms and filters developed through team discussions. The search included LLM-related terms (e.g., "large language model", "prompt engineering") and the names of specific LLMs. The complete search query is provided in **Supplementary Table S1**.

2.2. Inclusion and Exclusion Criteria

To focus on research articles presenting original data and quantitative results, we excluded certain article types, such as reviews. Given that the initial release of ChatGPT was on November 30th, 2022, we included articles published from 2023 onward, with only peer-reviewed articles included, while preprints were excluded.

The selection process began with the removal of duplicate articles, followed by a manual review of the deduplicated list. We excluded articles that met one or more of the following criteria: (1) Articles that were not of the appropriate type (e.g., preprints, reviews, editorials, comments). (2) Articles that did not involve generative LLMs; for example, those discussing chatbots but did not utilize LLMs were not considered. Although some encoder-based models like Longformer[3], NYUTron[4], GatorTron[5] are powerful and widely used, studies involving only these encoder-based models were excluded. (3) Articles where the LLM was not used for English-language communication. (4) Articles where the LLM application was unrelated to healthcare (e.g., LLMs used for passing exams or conducting research). (5) Articles that lacked quantitative evaluation (e.g., those that only presented communication records with ChatGPT). (6) Articles that did not involve EHR data. (7) Articles where the EHR data used were not original (e.g., synthetic or summarized data).

During the eligibility 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

We provided an overview of studies that applied generative LLMs to EHR data, grouped publications into three time period: 2023, the first half of 2024 (2024H1), and the second half of 2024 (2024H2). All papers published in 2023 were grouped since we found only two included articles published during the first half of 2023. We generated word clouds from study abstracts. We produced pie-charts for clinical tasks and LLMs used in included studies. Notably, clinical tasks were annotated by our physician using plain text and then classified into categories. In LLM pie-charts, LLMs appearing in two or fewer articles were categorized as "Others." For the final included studies, we categorized and summarized methods and effectiveness for the applied performance improvement strategies. Finally, a structured framework was proposed to support future research and generative LLMs deployment in healthcare settings.

3. Results

3.1.Trend of Applying Generative LLMs to EHR

Out of 18,735 unique records, 196 studies applied generative LLMs to EHR snice 2023 (**Figure 1**). As illustrated in **Figure 2**, The number of publications rose sharply from 26 in 2023 to 99 in the first half of 2024 (2024H1). Based on the 71 articles published during the first four months of 2024H2 included in our dataset, we estimated approximately 106 articles (71/4*6 ≈106) for the full half-year period 2024H2. ChatGPT was the most commonly used model in 2023 (47.5%) but showed decreasing use over time (32.7% in 2024H1, 21.6% in 2024H2), while GPT-4 usage increased (25.0%, 29.8%, and 34.2%, respectively). The share of less frequently used models—grouped as "Others"—also grew over time, reaching 36.0% in 2024H2. Llama-2 and Gemini gained traction in 2024, each surpassing 5% usage. Word cloud analysis revealed a gradual shift in thematic focus: while "ChatGPT," "patient," and "clinical" dominated early literature, terms such as "LLM", "accuracy," and "evaluation" became more prominent in 2024.

LLMs are being applied to a variety of clinical tasks, falling into the following major categories (Figure 2): 1) Clinical Decision Support (e.g., generating suggestions for diagnosis or treatment based on a patient's history; 2) Information Extraction (e.g., extracting clinical entities and relationships from free-text clinical notes); 3) Documentation (e.g., generating clinical reports, notes or referral letters) 4) Summarization (e.g., summarizing longitudinal patient records); 5) Patient Communications (e.g., developing a conversational agent that answers clinical queries from patients). Our analysis revealed that across all time periods, clinical decision support remained the most common application of generative LLMs, although its relative share decreased over time—from 84.6% in 2023 to 52.3% in 2024H1, and further to 48.8% in 2024H2. Meanwhile, other task categories became more prominent. Documentation use cases increased steadily, rising from 7.7% in 2023 to 19.6% in 2024H1 and 20.9% in 2024H2. Notably, Information Extraction tasks grew from 7.7% in 2023 to 15.9% in 2024H1, and reached 25.6% by 2024H2—surpassing documentation in terms of relative share. Summarization tasks first appeared in 2024H1 (6.5%) and remained present in 2024H2 (less than 5.0%), while conversational agents were consistently underrepresented (5.6% in 2024H1 and less than 5.0% in 2024H2).

3.2.Bias

Among the included studies, only five specifically investigated potential biases in LLMs. One study found that ChatGPT did not exhibit biases related to demographic factors such as age and gender when generating imaging referrals.[11] Another study reported that when performing disease classification, GPT-4 provided contextually appropriate and scientifically accurate explanations without showing evident bias.[12] In contrast, one study observed that male patients received more appropriate responses than female patients, suggesting a potential gender bias in ChatGPT-3.5's outputs.[13] Additionally, another study found that ChatGPT-3.5 exhibited potential racial bias in identifying emergency department visits for symptomatic kidney stones.[14] Furthermore, it was noted that bias may be introduced through the training data or the examples provided in prompts, leading to better performance when the input EHR data during inference time closely resembles the model's training data or in-context examples.[15]

3.3.Performance, Errors, and Hallucinations

As illustrated in **Table 1**, we summarized performances, errors, and hallucinations. Included studies highlighted a diverse range of performances and challenges associated with LLMs in analyzing EHR for clinical tasks.

3.3.1. Clinical Decision Support

122 included studies used LLMs to perform clinical decision support related tasks. LLMs demonstrated varying performance based on task complexity and specialization. Diagnostic accuracy ranged widely, with some tasks such as melanoma diagnosis achieving as low as 36% accuracy for the top-ranked diagnosis,[16] while others, like emergency medicine, reached an accuracy of 89%.[17,18] Notably, domain-specific fine-tuning significantly improved performance, with clinical trial matching algorithms reaching about 63% accuracy,[19] and NLP pipelines for gastrointestinal bleeding and radiology achieving exceptional sensitivity and specificity over

90%.[20] Common errors and hallucinations included the generation of incorrect diagnoses, inappropriate clinical recommendations, and temporal reasoning errors.

3.3.2. Information Extraction

39 Studies used LLMs to perform information extraction related tasks. LLMs like GPT-4 and ChatGPT-3.5 showed high accuracy, ranging from 64% to 98%, in extracting structured clinical data from unstructured sources such as radiology reports and pathology data.[21–24] However, errors in complex tasks, such as the misclassification of pathological terms or misinterpretation of tumor staging, were observed. For instance, GPT-3.5 struggled with detailed data extraction in lung cancer reports and made critical errors in privacy-sensitive contexts, generating false confidential health information.[22–25] These findings underscore the importance of model refinement in sensitive and specialized medical data extraction.

3.3.3. Summarization

We identified 10 studies that included a summarization task. LLMs demonstrated strong performance in summarization tasks, with accuracy metrics (precision, recall, and F1 scores) ranging from 0.77 to 0.84. Physician evaluations revealed that LLM-generated summaries were rated equal or superior to human-generated summaries in approximately 81% of cases.[26,27] Additionally, LLMs significantly improved readability, reducing the complexity of medical texts. However, the completeness of summaries remained a concern, with a completion rate around 56%, and high-level accuracy was reported at approximately 54%.[28] Errors, including misinterpretation of laterality and omission of critical medical details, highlight the need for continued oversight to ensure patient safety.[26,28]

3.3.4. Documentation

40 included studies used generative LLMs for clinical documentation tasks. GPT-4 exhibited high accuracy in extracting and structuring clinical information, with performance ranging from 88.6% to 98.9%.[22] Moreover, LLMs improved the readability of clinical documentation, lowering the Flesch-Kincaid grade levels from postgraduate levels to around grades 7-8.[29,30] Despite these improvements, hallucinations such as incorrect diagnoses, non-existent medical conditions, and misinterpretation of clinical stages remained a concern. The rate of such hallucinations varied from 0.26% to 2.82%,[29] but their clinical implications necessitate careful validation by healthcare professionals before implementation.

3.3.5. Patient Communication

We found 11 studies that used generative LLMs to help with patient communication tasks. Overall, LLMs contributed to enhancing the readability and patient-friendliness of medical texts, such as radiology reports and discharge summaries. In some cases, like smoking cessation therapy, GPT-4 provided acceptable responses 88% of the time.[31] However, safety concerns arose when LLMs generated discharge summaries, where inaccuracies or omissions of critical information could potentially harm patients if used without expert review.[28,29,32] These findings indicate that while LLMs can improve patient communication, they require careful oversight to ensure accuracy, particularly in clinical environments where patient safety is paramount.

3.4. Adaptation Strategies

Out of 196 studies applied generative LLMs to EHR, we identified 22 of them that developed a performance improvement strategy and compared with baseline methods quantitatively. **Table 2-5** summarize the findings. We categorized performance improvement strategies into four categories: prompt engineering (8 articles), fine-tuning (14 articles), multimodal integration (7 articles), and ensemble with traditional AI models (2 articles).

3.4.1. In Context Learning

Among the eight studies included in our review, seven reported that prompt engineering led to improved performance of generative large language models (LLMs) on clinical tasks, while one study found that few-shot prompting did not outperform zero-shot prompting. Clinical applications that benefited from prompt engineering included disease diagnosis, clinical text summarization, emergency department admission prediction, detection of early cognitive decline, radiology report error detection, MRI image analysis, and tumor

classification. For example, targeted prompting raised ChatGPT's diagnostic accuracy for bone tumors from 0.73 to 0.87, and boosted GPT-4's accuracy in detecting mild cognitive impairment from 86% (zero-shot) to 93%. In another case, explicit prompts improved MRI brain tumor detection accuracy from 66.5% to 87.0% and modality recognition accuracy from as low as 3–46% up to 74%. Similarly, GPT-4's AUC in predicting patient admissions increased from 0.79 to 0.87 with retrieval-augmented prompts. The only exception was in identifying ED visits due to symptomatic kidney stones, where few-shot prompting slightly reduced GPT-4's macro-F1 score from 0.8556 to 0.8507, despite prompt augmentation with demographic and history information still showing a slight improvement.[14] Overall, the findings indicate that thoughtfully designed prompts—especially those incorporating clinical context, examples, or auxiliary information—can substantially enhance LLM performance across a wide range of EHR-based tasks.

3.4.2. Fine-Tuning

Across the 14 studies reviewed, 13 demonstrated that fine-tuning significantly improved the performance of generative LLMs on diverse EHR-based clinical tasks, while one study reported that a non-fine-tuned general-purpose model outperformed a medically fine-tuned counterpart. Clinical applications benefiting from fine-tuning included clinical text summarization, phenotype recognition, DRG coding, disease classification, de-identification, and social information extraction. For example, in summarization tasks, fine-tuned FLAN-T5 using QLoRA improved BERTScore by ~5 points and ROUGE-L by 3–4 points compared to ICL. In the MeQSum task, QLoRA-fine-tuned models consistently outperformed prompt-based versions under low-context conditions. Similarly, in phenotype recognition, fine-tuned GPT-J achieved an F1 score of 0.832, outperforming traditional and BERT-based models. In the de-identification task, fine-tuning boosted GPT-3.5's accuracy from 94.38% (zero-shot) to 99.20%. The extraction of sleep-related information also saw large performance gains, with fine-tuned Llama2 achieving an F1 score of 0.96 for "night wakings," compared to just 0.39 using a prompting-based strategy. Four studies directly compared fine-tuning and prompting methods, revealing that while GPT-4 with few-shot prompting often outperformed fine-tuned open-source models in human evaluation (e.g., correctness and completeness of summaries), fine-tuning remained more effective for domain-specific tasks when large proprietary models were not available. The only exception was a scoliosis severity classification task, where general-purpose models like GPT-4 and Copilot achieved 100% accuracy, outperforming the medically fine-tuned PMC-LLaMA 13B model (73%). Overall, the findings support fine-tuning as a powerful approach for enhancing LLM performance in clinical tasks, especially when tailored to domain-specific data and when proprietary models are not accessible.

3.4.3. Multimodal Data Fusion

Across the seven studies that involved multimodal data, all demonstrated that multimodal data fusion—integrating textual clinical data with other modalities such as medical imaging, laboratory results, or structured patient information—consistently improved the performance of generative large language models (LLMs) across diverse clinical tasks. These tasks included disease diagnosis, radiology report generation, prognosis prediction, treatment planning, and clinical error detection. For example, the EHR-KnowGen model, which fused clinical notes, events, and lab results, improved Micro F1 by up to 14.6% and accuracy by up to 13.5% compared to text-only baselines. Similarly, in COVID-19-related tasks, the multimodal Med-MLLM achieved a BLEU-4 of 42.2 and an AUC of 98.7%, outperforming unimodal baselines by large margins even when trained on just 1% of labeled data. A YOLOv7-based multimodal system for pressure injury classification reached an average F1 score of 0.9238—substantially higher than earlier unimodal models. Three studies specifically employed GPT-4V, showcasing significant performance gains when combining textual patient histories with radiological images. For instance, GPT-4V achieved 77.01% accuracy in USMLE-style diagnosis tasks with image and text input, compared to 19.54% with images alone—a 294% relative improvement. Another GPT-4V study showed 17% accuracy gains in answering complex radiological questions, and superior performance across 15 of 16 radiologic subspecialties. Overall, these findings provide strong quantitative evidence that multimodal data fusion can significantly enhance LLM-based clinical reasoning, outperforming single-modality approaches across a wide range of healthcare applications.

3.4.4. Ensemble of The LLM and Traditional AI

One examined the ensemble of LLMs and traditional AI methods, the combined approach demonstrated superior performance over individual models for detecting early signs of cognitive decline from clinical notes. While GPT-4, optimized via error-analysis-based prompting, achieved a precision of 71.7%, recall of 91.4%, and F1 score of 80.3%, it was outperformed by traditional models such as an attention-based deep neural network (DNN) and XGBoost, which reached F1 scores of 84.5% and 85.4%, respectively. However, the ensemble model that integrated GPT-4, DNN, and XGBoost using majority voting achieved the highest performance, with a precision of 90.2%, recall of 94.2%, and F1 score of 92.1%. These improvements were statistically significant and underscore the effectiveness of ensemble learning in leveraging complementary model strengths to enhance overall diagnostic accuracy in clinical NLP tasks.

3.5.The Framework for Applying LLMs to EHR for Clinical Tasks

Based on our comprehensive review, we proposed a structured framework for conducting research using LLMs to analyze EHR for clinical applications (**Figure 3**). This framework comprises three sequential steps: defining the clinical task, selecting performance improvement strategies, and conducting evaluations. Clinical tasks commonly addressed using LLMs include clinical decision support, information extraction, summarization, documentation, and patient communication. To enhance LLM performance for these tasks, researchers may apply various strategies such as prompt engineering (e.g., few-shot learning, chain-of-thought prompting, retrieval-augmented generation), fine-tuning (e.g., LoRA, QLoRA, DeepSpeed), multimodal data fusion (e.g., early, intermediate, or late fusion of EHR modalities), and integrating traditional AI through ensemble methods (e.g., majority voting or hybrid prompting). The effectiveness of these approaches has been demonstrated in prior studies targeting specific clinical tasks (**Tables 2–5**). Finally, model evaluation should be tailored to the task's goals. For instance, summarization tasks may prioritize correctness, as well as the balance between completeness and conciseness, while patient communication tasks may emphasize readability and patient-friendliness.

4. Discussion

Generative LLMs have shown exceptional potential across NLP tasks, yet their deployment on real-world EHR remains at an early stage. In this review, we systematically examined LLM's applications on real-world EHR data across diverse clinical tasks. Our contributions are threefold: 1) categorized clinical tasks into five domains: clinical decision support, information extraction, summarization, documentation, and patient communication, and summarized performances, common errors and hallucinations; 2) we categorized the improvement strategies into four types—prompt engineering, fine-tuning, multimodal data fusion, and ensemble with traditional AI—and synthesized evidence of their quantitative effectiveness; 3) we proposed a framework for conducting research using generative LLMs to investigate EHR data, which clearly summarized commonly used performance improvement strategies and evaluation methods. By mapping out the evolving landscape of LLM enhancements, our work aims to guide the responsible development and deployment of AI tools that are better aligned with the complexities of clinical data and workflows.

Our results show that when applying performance improvement strategies to generative LLMs in the healthcare domain, several practical considerations must be taken into account. First, the choice of strategy often entails trade-offs between performance gains and resource demands. For example, fine-tuning typically yields substantial accuracy improvements but may require extensive labeled clinical data, computational resources, and careful model versioning, particularly when applied to large-scale models.[33,34] Similarly, multimodal data fusion can significantly enhance model contextual understanding, yet it depends on the availability and harmonization of structured data, imaging, or other modalities alongside unstructured text.[35] Prompt engineering, in contrast, offers a more lightweight and accessible method, especially when computational budgets are constrained, but its effectiveness may be sensitive to task complexity and model limitations.[36] Moreover, ensemble strategies, though potentially powerful in integrating complementary strengths of different model types, remain underexplored and require more research to understand their implementation feasibility in production environments. Developers and clinical informatics teams must therefore carefully evaluate the cost-benefit profile of each method, consider data availability and institutional constraints, and plan for ongoing evaluation and updating of strategies as models evolve.[37,38] Additionally, we found that several studies lacked sufficient methodological detail, such as descriptions of fine-tuning approaches and hardware specifications. To support responsible performance optimization, it is essential to ensure transparency in reporting the rationale behind strategies, thereby promoting reproducibility and enabling others to build upon validated methods.

Another key insight from our review is that the effectiveness of performance improvement strategies is often task specific. For instance, prompt engineering and few-shot learning were particularly effective for diagnostic classification and decision support, where clear examples could guide model reasoning. In contrast, fine-tuning consistently outperformed prompting in complex tasks such as clinical summarization and phenotype normalization, where domain-specific language and nuanced interpretations were essential. Multimodal data integration demonstrated particular value in tasks involving medical imaging, pathology, or structured data, where textual context alone was insufficient. These findings suggest that no single strategy is universally optimal; instead, researchers should consider the nature of their clinical task and available resources when selecting or designing enhancement methods. Future benchmarking efforts may benefit from systematically comparing strategies across task types, to establish more formal guidance on method selection based on clinical objectives.

Several broader themes emerged from the reviewed studies that merit further consideration. First, the performance gap between proprietary LLMs (e.g., GPT-4) and open-source models remains evident in some applications, particularly when prompt engineering is leveraged, yet open-source models offer greater customizability, transparency, and data privacy—making them preferable for institution-level deployments when properly fine-tuned. Second, some studies demonstrated performance plateaus, where adding more in-context examples or extending fine-tuning epochs yielded diminishing returns,[26] pointing to the need for efficient optimization strategies rather than brute-force scaling. Third, reproducibility remains a concern; studies varied in their documentation of method details (e.g., fine-tuning methods and used hardware are missing in

some studies), underscoring the importance of developing and adopting minimum reporting standards for LLM-based clinical studies. Additionally, given the evolving nature of clinical data, performance improvements achieved at one time point may degrade over time due to temporal drift, making continual learning and prompt revalidation essential components of any long-term deployment strategy.[39] Addressing these challenges will require ongoing interdisciplinary collaboration between clinicians, AI researchers, and informatics professionals to ensure that LLMs in healthcare are not only powerful but also trustworthy and sustainable.

Moreover, the limited attention to bias in existing studies highlights a critical gap in the application of generative LLMs to EHR for clinical tasks. Although two studies reported minimal or no evidence of bias,[11,12] some others revealed disparities in model performance based on patient demographics,[13,14] underscoring the risk of perpetuating or amplifying health inequities through LLM-generated outputs. These inconsistencies suggest that LLMs may be sensitive to both the composition of their training data and the structure of input prompts, particularly when prompts contain examples that resemble certain subpopulations more closely than others. Given the potential for bias to impact clinical decision-making and patient safety, future work should prioritize comprehensive, systematic evaluations of LLM fairness across diverse patient groups and clinical contexts. Furthermore, transparency in model development and the use of bias mitigation strategies—such as balanced training datasets, adversarial testing, and fairness-aware prompting—are essential to ensure equitable and trustworthy LLM deployment in healthcare settings. As LLMs become increasingly integrated into clinical workflows, addressing these ethical and methodological concerns will be critical for realizing their potential to support, rather than undermine, equitable patient care.

Finally, our proposed framework offers a systematic and extensible approach for applying generative LLMs in clinical research using EHR data. By explicitly delineating the clinical task, performance optimization strategies, and evaluation criteria, this workflow promotes methodological transparency, reproducibility, and adaptability to diverse use cases. For the research community, it serves as a scaffold to design, compare, and improve LLM-based interventions across a wide range of clinical applications, from supporting diagnostic decisions to enhancing documentation efficiency. Importantly, the integration of multimodal data fusion, fine-tuning, and ensemble methods encourages the blending of LLM capabilities with traditional domain-specific models, which may lead to more robust and context-aware outputs. For the general population, particularly patients, this framework facilitates the development of more accurate, readable, and patient-friendly tools that can improve communication, foster shared decision-making, and ultimately enhance the delivery of personalized care. As the use of generative AI continues to expand in healthcare, adopting such structured frameworks is critical for aligning technological advancements with real-world clinical impact.

This review has several strengths, including its timely focus on the post-ChatGPT era, its exclusive inclusion of real-world EHR-based studies, and its comprehensive coverage of four major performance improvement strategies. By analyzing peer-reviewed articles across diverse clinical tasks and data modalities, our work offers a current and practical synthesis for researchers and implementers seeking to improve the application of generative LLMs in healthcare. However, several limitations should be acknowledged. The studies included are drawn from a rapidly evolving and relatively new field, meaning that some findings are preliminary and potentially subject to publication bias, favoring positive results. The heterogeneity of clinical tasks, models, and evaluation metrics also limited our ability to conduct quantitative meta-analysis, and the focus on peer-reviewed literature from 2023 onward may have excluded valuable insights from preprints or earlier work. Looking ahead, future research should prioritize external validation of performance strategies, explore ensemble and multimodal data fusion more extensively, improving research reproducibility by adopting more transparent reporting and setting up minimum reporting guidelines, and advance the safety and reliability of LLM outputs through hallucination detection, bias mitigation, and fail-safe mechanisms. Moreover, effective integration into clinical workflows—ensuring usability, interoperability, and alignment with healthcare regulations—will be essential for translating technical advances into real-world impact. Collectively, these directions will be vital to support the responsible and effective adoption of generative LLMs in clinical settings.

5. Conclusion

Our systematic review reveals a rapid evolution in the application of generative LLMs to real-world EHR data, with a clear shift from general-purpose models like ChatGPT to more diverse and domain-adapted alternatives. While early efforts focused largely on clinical decision support, recent studies increasingly address a wider array of tasks such as information extraction and documentation. Among the reviewed adaptation strategies, in-context learning and fine-tuning demonstrated consistent performance gains, with fine-tuning proving particularly effective for specialized clinical tasks. Multimodal and ensemble approaches further improved diagnostic reasoning by leveraging complementary data sources and model types. However, only a small portion of studies quantitatively evaluated the effectiveness of these strategies, and evaluation practices remain heterogeneous due to the absence of standardized metrics. Our proposed framework connects clinical tasks with appropriate adaptation strategies and evaluation approaches, offering a structured guide to improve reproducibility, optimize model performance, and ensure responsible deployment of LLMs in healthcare. Future research should prioritize standardized evaluation protocols, external validation, and bias mitigation to advance the safe and equitable use of generative LLMs in clinical practice.

Acknowledgments

This study was funded by NIH-NLM 1R01LM014239, NIH-NIA R01AG080429, and OpenAI Researcher Access Program.

Figures

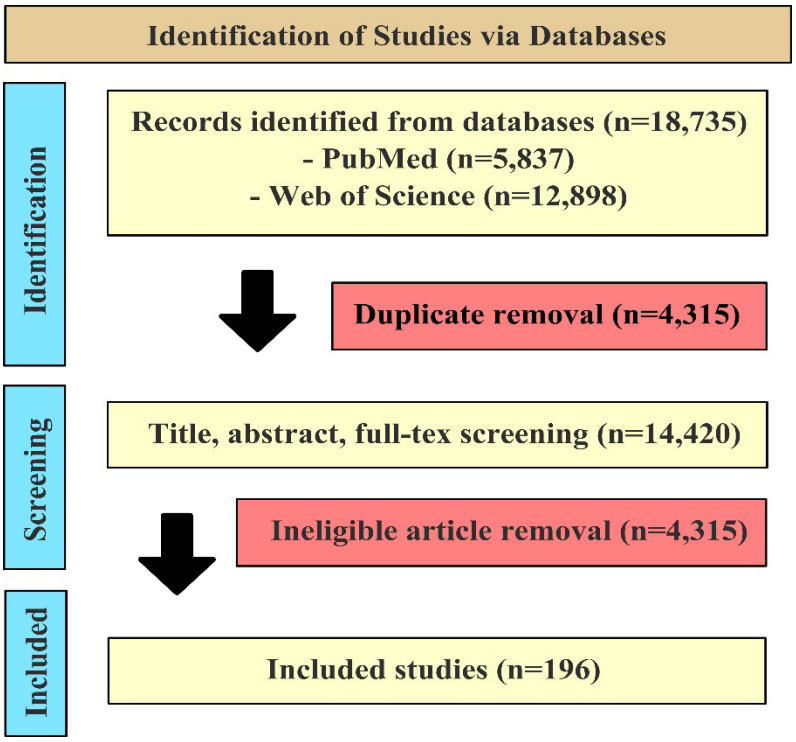

Figure 1. PRISMA Flow Chart for Eligibility Screening. A total of 18,735 records were identified from two databases: PubMed (n=5,837) and Web of Science (n=12,898). After removing 4,315 duplicates, 14,420 records underwent title, abstract, and full-text screening. Among them, 10,104 studies were excluded for not applying generative LLMs to real-world EHR data. Finally, 196 studies were included.

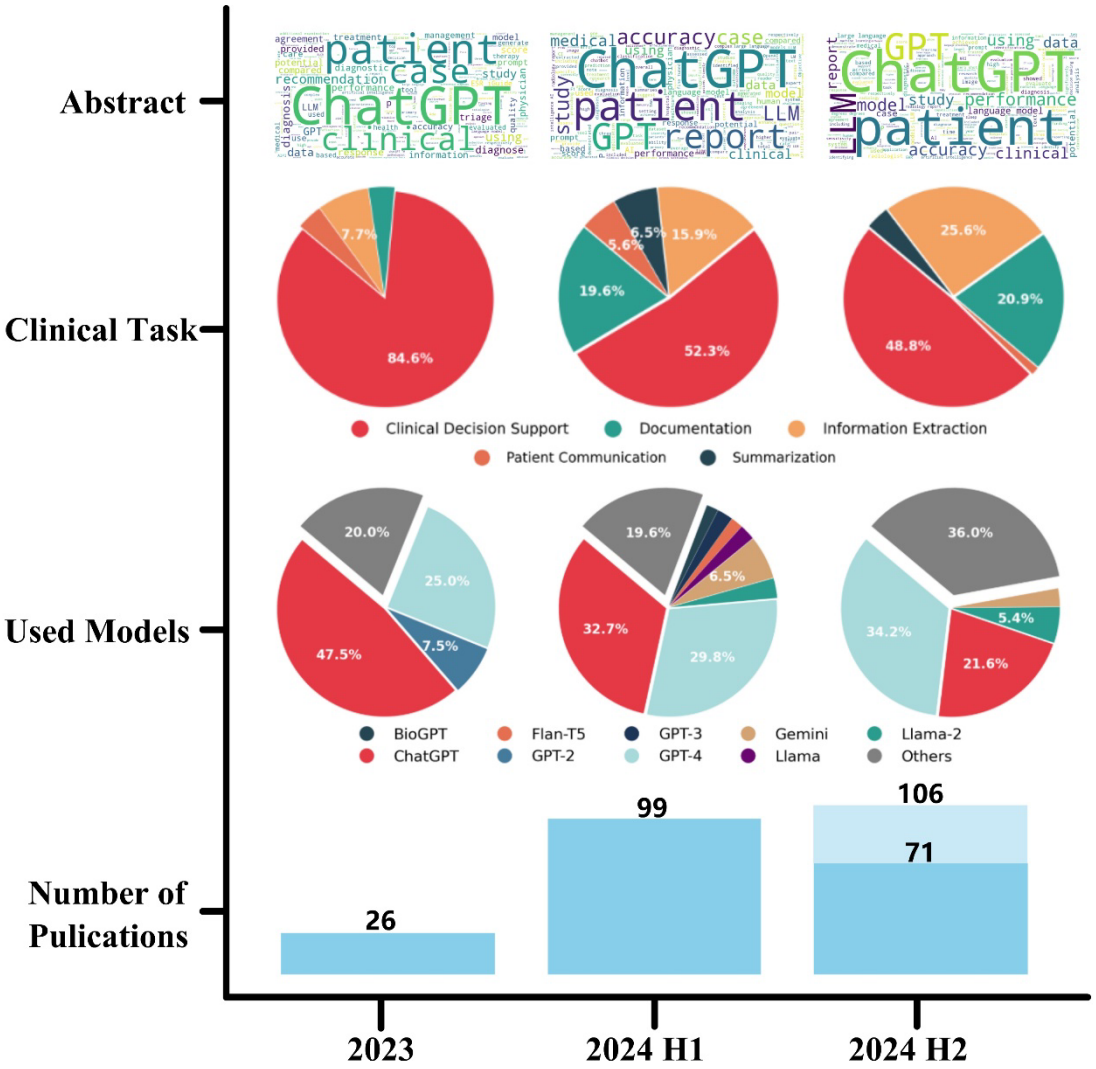

Figure 2. Summarization of Applying Generative LLM to EHR. The number of publications rose from 26 in 2023 to 99 in 2024H1, with an estimated 106 in 2024H2 based on data from the first four months. ChatGPT remained the most used model but declined over time (47.5% in 2023 to 21.6% in 2024H2), while GPT-4 and other models—including Llama-2 and Gemini—gained traction. Word clouds show a shift from terms like "ChatGPT" and "clinical" to "LLM," "accuracy," and "evaluation." Clinical Decision Support (CDS) was the dominant task throughout, though its share dropped (84.6% to 48.8%) as use cases diversified. Documentation and Information Extraction saw steady growth, with the latter surpassing documentation by 2024H2. Summarization emerged in 2024, and Patient Communication remained minimal.

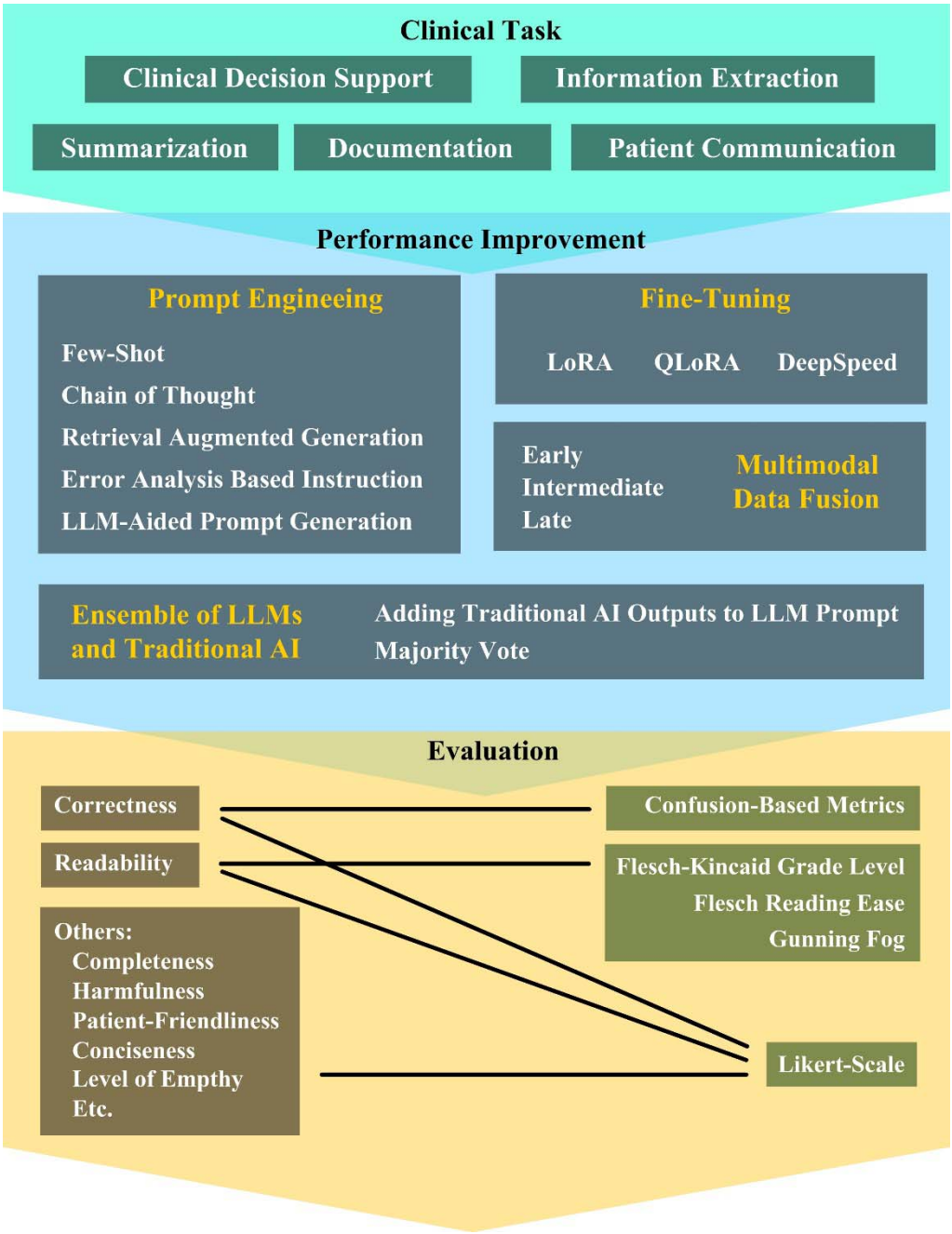

Figure 3. Framework for Using LLM to Analyze EHR data. First, define the clinical task. This framework outlines the steps for applying LLMs to clinical tasks. First, the specific clinical task is defined, which may fall into one or more categories: clinical decision support, information extraction, summarization, documentation, and patient communication. Researchers can then choose one or more performance improvement strategies based on the task, such as prompt engineering (e.g., few-shot learning, chain-of-thought prompting), fine-tuning methods (e.g., LoRA, QLoRA, DeepSpeed), and multimodal data fusion. Additionally, an ensemble of LLMs and traditional AI models can be applied, incorporating traditional AI outputs to LLM prompts through majority voting. Existing evaluations about their effectiveness are summarized in **Table 2-5**. Finally, the framework includes the selection of an evaluation strategy, with commonly used metrics like confusion-based metrics, readability assessments (e.g., Flesch-Kincaid, Flesch Reading Ease, and Gunning Fog), and Likert scale ratings. The summary of performances from included studies for each type of clinical task are illustrated in **Table 1**. The glossary of all evaluation methods used in included studies are summarized in **Supplementary Table S2-S3**.

Tables

Table 1. Summary of Evaluation, Errors, and Hallucinations

Clinical Task Category (Number of Studies)	Common Evaluate Methods	Performance Summary	Summary of Errors and Hallucinations
Clinical Decision Support (122)	• Correctness: Confusion-Based Metrics; Agreement with Human Expert; Likert-Scale • Completeness: Likert-Scale	Performance variability: LLM performance varied widely based on clinical task, complexity, and specialty. Diagnostic Accuracy in Dermatology: Accuracy ranged from 36% (top-ranked melanoma diagnosis) to 55% (when considering multiple differential diagnoses).[16] Strong Performance in Ophthalmology: GPT-4 matched or surpassed fellowship-trained specialists in diagnostic accuracy and completeness.[40] Emergency Medicine**:** LLMs achieved up to 89% accuracy for clinical acuity assessment, comparable to human physicians; performance declined for nuanced tasks (e.g., hospital admissions, imaging, antibiotic prescriptions).[17,18] Impact of Fine-Tuning: Domain-specific fine-tuning improved performance—clinical trial matching accuracy reached ~63%, comparable to experts.[19] High-Precision NLP Pipelines: Pipelines for gastrointestinal bleeding and radiology follow-up showed exceptional sensitivity and specificity (exceeding 90%), outperforming traditional NLP methods.[20]	Clinically Plausible Errors: LLMs may generate incorrect but clinically plausible details (e.g., non-existent diagnoses, inaccurate treatments) due to inadequate domain-specific training or validation.[18] Misprioritization of Differential Diagnoses: LLMs may prioritize less relevant conditions over clinically significant ones, with accuracy improving when considering multiple differential diagnoses (e.g., dermatology applications).[16] Misinterpretation of Clinical Context: LLMs may misidentify historical or negated symptoms as current, necessitating hybrid NLP methods combining LLMs and rule-based approaches (e.g., gastrointestinal bleeding detection).[20] Inappropriate Clinical Recommendations: LLMs have generated overly conservative or aggressive recommendations (e.g., misidentifying cognitive decline, recommending unnecessary admissions or imaging).[18,21] Temporal Reasoning Errors: LLMs may inaccurately capture disease progression or resolution, complicating decision-making in acute care settings (e.g., emergency acuity assessment).[17] Demographic Biases: LLMs may inherit biases from training data, leading to systematic skewing of outputs and underscoring the need for ongoing model refinement and human oversight.[19]
Information Extraction (39)	• Correctness: Confusion-Based Metrics; Agreement with Human Expert; Likert-Scale • Completeness: Likert-Scale	Overall Correctness: LLMs (particularly GPT-4 and ChatGPT-3.5) showed high accuracy, precision, recall, and F1 scores when extracting structured clinical data from unstructured texts, with accuracy ranging from ~64% to >98%, depending on task complexity and prompt specificity.[21–24] Completeness (Recall): Models demonstrated strong but variable recall, with structured data extraction (e.g., tumor status, imaging findings) achieving recall rates of ~88.6% to 94.4%, indicating substantial completeness.[22,24]	TNM Staging Errors: ChatGPT-3.5 misinterpreted tumor-node-metastasis (TNM) staging rules and misclassified specialized pathological terms in lung cancer reports.[23] Hallucinations in Sensitive Data: GPT-3.5 generated inaccurate excerpts with non-existent confidential health information when extracting adolescent encounter data, highlighting privacy risks.[25] Procedural Data Errors: GPT-3.5 incorrectly recorded critical neuroradiology procedural details, such as thrombectomy maneuvers and anatomical references (e.g., mislabeling "M2" as "M1").[24] Numeric Accuracy Challenges: GPT-4 showed reduced accuracy (88.6%) when summing tumor diameters, compared to simpler classification tasks.[22]
Summarization (10)	• Correctness: Confusion-Based Metrics; Likert-Scale • Completeness: Likert-Scale • Readability: Flesch-Kincaid Grade Level; Flesch Reading Ease; Gunning Fog	Overall Correctness: LLMs achieved precision, recall, and F1 scores typically ranging from ~0.77 to 0.84 across diverse clinical applications.[26,27] Physician Evaluations: About 81% of LLM-generated summaries were rated equivalent or superior to those by human experts.[26,27] Patient-Focused Readability: Reading difficulty improved from grade level 11 to ~grade 6, and understandability scores rose from 13% to 81%.[28] Remaining Concerns: Completeness (~56%) and high-level accuracy (~54%) remain challenges, raising safety and reliability concerns.[28]	Typical Mistakes: LLMs commonly confused laterality (right vs. left), omitted critical historical details, included nonexistent medical conditions, and made numeric inaccuracies (e.g., patient measurements, clinical scores).[26,28] Error Rates: Errors occurred in ~6–13% of LLM outputs, compared to ~9–25% in human expert summaries.[26] Safety Implications: Despite lower error rates, mistakes present significant safety concerns, emphasizing the need for ongoing human oversight and further model refinement.[26]
Documentation (40)	• Correctness: Confusion-Based Metrics; Agreement with Human Expert; Likert-Scale • Readability: Flesch-Kincaid Grade Level; Flesch Reading Ease; Gunning Fog; Likert-Scale • Conciseness: Word	Data Extraction: GPT-4 achieved 88.6%-98.9% accuracy (F1: 0.922-0.948)[22] for hepatocellular carcinoma imaging and F1 score of 0.997 for pancreatic cancer report structuring.[41] Improved Readability: LLM-simplified reports reduced Flesch-Kincaid grade levels from ~13.2 (postgraduate) to ~7–8.[29,30] Laterality Error Detection: GPT-4 identified laterality errors in radiology reports with up to 98.3% accuracy.[42] Error and Hallucination Rates: Incorrect interpretations ranged from 0.44%-3.44%, and hallucinations from 0.26%-2.82%.[29] Clinical Utility and Oversight: GPT-4 substantially improved documentation	Types of Hallucinations: LLMs may misinterpret medical details, incorrectly extract data, or invent clinical findings not supported by source texts. Examples of Hallucinations: GPT-4 occasionally falsely states diagnoses, creates non-existent conditions, or misidentifies disease stages or severity (hallucination rates: 0.26%-2.82%).[29] Typical Errors: Common mistakes include laterality mismatches (left vs. right confusion), numerical inaccuracies (e.g., tumor sizes, lesion counts), omission of critical data, and misclassification of medical conditions or treatments.[42,43] Error Rates: Overall incorrect interpretations range from 0.44% to 3.44%. Clinical Significance: Despite relatively low error rates, mistakes are clinically

	Count • Completeness: Recall; Likert-Scale • Patient-Friendliness: Likert-Scale	quality and patient comprehension but still requires clinician oversight.	significant and require careful professional verification before use in clinical care.
Patient Communication (11)	• Correctness: Confusion-Based Metrics; Likert-Scale • Patient-Friendliness: Likert-Scale • Preference Over Experts: Likert-Scale • Level of Empathy: Likert-Scale • Readability: Likert-Scale • Harmfulness: Likert-Scale	Improved Readability: LLMs reduced the Flesch-Kincaid Grade Level of medical texts (e.g., radiology reports) from ~13.2 to as low as 7.45, enhancing patient comprehension.[29] Application-Specific Performance: GPT-4 generated acceptable responses 88% of the time for smoking cessation therapy.[31] Performance Limitations: LLMs were less successful in generating hand clinic letters[32] and raised safety concerns in ~18% of discharge summaries due to inaccuracies.[28] Need for Oversight: While LLMs improve patient-friendliness, careful professional review is essential to ensure accuracy and safety.	Risk of Harmful Advice: In discharge summaries, LLMs sometimes included incorrect statements or medical advice that could potentially harm patients if not reviewed by experts.[28] Omission of Critical Information: Important medical details were sometimes omitted in LLM-simplified discharge summaries, posing safety risks.[28] Oversimplification Risks: LLMs may oversimplify or misinterpret technical details, leading to loss of essential information. Hand Clinic Letters: ChatGPT failed to effectively simplify or improve the readability of hand clinic letters.[32] Pathology Reports: Simplified pathology reports occasionally distorted or misrepresented critical medical context, risking patient misunderstanding.[29]

Table 2. The Effectiveness of In Context Learning towards Performance Improvement of Applying Generative LLMs to EHR in Clinical Tasks

Clinical Task Category	Detailed Clinical Task	In Context Learning Method	Quantitative Performance Summary	Notable Findings
Clinical Decision Support	Diagnosing of benign and malignant bone tumors	Few Shot	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[44]	Few-shot prompting improved performance
	Automatic detection of speech recognition errors in radiology reports	RAG, Chain-of-Thought	The study applied RAG and Chain-of-Thought. 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. [45]	RAG and Chain-of-Thought improved performance
	Disease diagnosis (Chronic Obstructive Pulmonary Disease, Chronic Kidney Disease, and Primary Biliary Cirrhosis)	Few-shot	Few shot prompting improved performance of all LLMs (ChatGPT, Llama3, and GPT-4). Prompt engineering using few-shot prompts significantly enhanced precision and specificity for ChatGPT, GPT-4, and LLaMA3 in identifying patients with Chronic Obstructive Pulmonary Disease, Chronic Kidney Disease, and Primary Biliary Cirrhosis from electronic health records, with precision improvements up to 42.54 percentage points (e.g., ChatGPT's precision for PBC improved from 41.67% to 84.21%). However, this improvement came at the cost of decreased sensitivity, which dropped by up to 44.44 percentage points (e.g., LLaMA3's sensitivity for COPD decreased from 80.95% to 36.51%). Among all models, GPT-4 exhibited the most balanced performance, maintaining high sensitivity while significantly improving precision and specificity.[12]	Few-shot prompting improved precision and specificity. However, this improvement came at the cost of decreased sensitivity, which dropped by up to 44.44 percentage points (e.g., LLaMA3's sensitivity for COPD decreased from 80.95% to 36.51%)
	Predicting patient admissions from the emergency department	RAG	The baseline zero-shot GPT-4 achieved an accuracy of 77.5%, AUC of 0.79, and AUPRC of 0.48, while incorporating retrieval-augmented generation (RAG) few-shot prompts boosted accuracy to 81.3%, AUC to 0.82, and AUPRC to 0.56.[46]	RAG Few-shot prompting improved performance
	Detecting early signs of cognitive decline (such as mild cognitive impairment and dementia) from clinical notes	Few-shot, RAG, error analysis-based instruction	Prompt engineering, especially the use of error-analysis-based instructions, significantly improved GPT-4's accuracy from 86% (zero-shot) to 93%. Compared to the best-performing few-shot prompting method (85% accuracy with Retrieval Augmented Generation (RAG)-dynamic five-shot), the error-analysis-based instructions further increased performance by 8 percentage points. The optimized GPT-4 achieved a precision of 71.7%, recall of 91.4%, and F1-score of 80.3%.[21]	Error analysis-based instruction prompting outperformed few-shot and RAG. Few-shot did not outperform zero-shot.
	Identifying ED visits due to symptomatic kidney stones	Few-shot	Few-shot prompting (providing model examples) did not yield performance gains over zero-shot prompting, as GPT-4's macro-F1 decreased slightly from 0.8556 (zero-shot) to 0.8507 (two-shot).[14]	Few-shot prompting did not improve performance
Summarization	Clinical text summarization	Few-shot	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.[26]	More examples in the prompt would lead to better performance, but the improvement becomes less obvious when adding more and more examples.
	Clinical text summarization	LLM-aided prompt generation	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%.[47]	LLM-aided prompt improved similarity score, and soft prompting improved reproducibility of LLM.

Table 3. The Effectiveness of Fine-Tuning towards Performance Improvement of Applying Generative LLMs to EHR in Clinical Tasks

Clinical Task Category	Detailed Clinical Task	Fine-Tuning Method	Fine-Tuned LLM	Fine-Tuning Hardware	Quantitative Performance Summary	Notable Findings
Clinical Decision Support	The assignment of Diagnosis-Related Group (DRG) codes for hospitalized patients	LoRA	LLaMA-7B	One Nvidia RTX A6000 GPU with 48GB of graphics memory	Fine-tuning LLaMA improved macro F1 by approximately 40.3% over ClinicalBERT and by 35.7% over CAML.[48]	Fine-Tuning improved performance.
	Generating differential diagnoses from pediatric intensive care unit (PICU) admission notes	Deepspeed	BioGPT-Large, LLaMa-7B	Four A40 Nvidia GPU	The fine-tuned LLaMa-7B outperformed the larger LLaMa-65B by 0.23 Likert units and BioGPT-Large by 0.86 units. Fine-tuned BioGPT-Large also improved by 0.76 units over its base version. In ranking clinical differentials, fine-tuned LLaMa-7B and BioGPT-Large placed first in 29% and 24% of cases, respectively, compared to only 10% for the non-fine-tuned BioGPT-Large, highlighting the value of domain-specific fine-tuning.[49]	Fine-tuning a small model can outperform a large model.
	Identification of pathologic complete response (pCR) in breast cancer patients following neoadjuvant chemotherapy	LoRA	GPT-2	One NVIDIA Tesla V100 16GB GPU	The non-fine-tuned models tested (e.g., BERT, BART, T5, and GPT-2 without fine-tuning) yielded lower F1 scores ranging from 69.6% to 91.3%. In contrast, the fine-tuned GPT-2 achieved a superior F1 score of 93.0%.[50]	Fine-tuning improved performance
Information Extraction	Identifying ED visits due to symptomatic kidney stones	Did not mention	ChatGPT	Did not mention	The fine-tuned GPT-3.5 achieved a macro-F1 of 0.82, clearly surpassing the pre-trained GPT-3.5, which achieved a macro-F1 of 0.79, representing an absolute increase of 0.03 in macro-F1 score.[14]	Fine-tuning improved performance
	Extracting essential lung cancer staging descriptors from surgical pathology reports	LoRA	Llama-2-7B, Mistral-7B	Did not mention	The Deductive Mistral-7B model achieved a macro F1 score above 0.9900 for the majority of descriptors. In contrast, non-fine-tuned (general-purpose) models such as Llama-2-7B and Mistral-7B struggled, often scoring below 0.7000 on the same tasks.[51]	Fine-tuning improved performance
	Extracting detailed sleep-related information from the clinical notes of Alzheimer's disease (AD) patients	LoRA	Llama 2	Did not mention	Llama 2-SFT consistently exhibited higher performance metrics, for instance, achieving an F1 score of 0.96 for "night wakings" compared to only 0.39 by Llama 2-CoT (Chain-of-Thought) without fine-tuning.[52]	Fine-tuning improved performance and outperformed CoT prompting
	Automated extraction of social support (SS) and social isolation (SI) information from clinical psychiatry notes	LoRA	T5	Did not mention	The zero-shot FLAN-T5-XL models yielded notably lower macro averaged F1-scores around 0.49–0.52, while the fine-tuned version improved these scores substantially to 0.65 (Mount Sinai Health System) and 0.82 (Weill Cornell Medicine). This represents absolute F1-score improvements of roughly 0.13–0.33.[53]	Fine-tuning with synthetic, clinically relevant examples improved performance
Summarization	Clinical text summarization	QLoRA	FLAN-T5	One NVIDIA Quadro RTX 8000	Fine-tuning improved FLAN-T5's MEDCON score from ~26 to ~29 on the Open-i dataset and boosted BERTScore by ~5 points (e.g., from ~85.0 to ~89.5) on the MeQSum task. ROUGE-L scores also increased by 3–4 points.[26] The performance of Med-Alpaca was numerically worse than Alpaca in 21 out of 24 experimental configurations (87.5%) across four NLP evaluation metrics (BLEU, ROUGE-L, BERTScore, and MEDCON).	Fine-tuning open-source models performed better than the original model with few shot prompting but still performed worse than GPT-4 with few shot prompting. Fine-tuning can hurt performance when it is done for domain adaptation without aligning it to the specific task.
	Generating layman-friendly summarization of hospital discharge summary	QLoRA	LLaMA 2	Did not mention	Fine-tuning outperformed both non-fine-tuned prompting strategies across all quantitative and qualitative measures, with a ~2.2x increase in BLEU score (0.1438 vs. 0.0640), a ~38% improvement in ROUGE-L F1 (0.3877 vs. 0.2818), a ~33% increase in precision (0.3264 vs. 0.2462), a ~52% increase in ROUGE-L recall (0.4773 vs. 0.3143), and a notable boost in human rating, rising from 6.13 to 8.90 out of 10 compared to zero-shot prompting.[54]	Fine-tuning outperformed zero-shot prompting and soft prompting
	Generation of concise and accurate summaries for knee Magnetic Resonance Imaging (MRI) radiology reports	Did not mention	T5	One NVIDIA GeForce RTX 3070 GPU	The fine-tuned T5 model achieved a Rouge-L of 0.638 and Bleu-1 of 0.652, far outperforming its zero-shot version (Rouge-L: 0.182, Bleu-1: 0.247)[55]	Fine-tuning improved performance

	Automated summarization of clinical conversations, specifically emergency department (ED) consultation calls	Did not mention	BART	One A100 NVIDIA GPU with 40 GB of VRAM),	Fine-tuned BART-Large-CNN achieved significantly higher performance metrics (ROUGE-1 F1=0.49, ROUGE-2 F1=0.23, ROUGE-L F1=0.35) compared to its zero-shot counterpart (ROUGE-1 F1=0.23, ROUGE-2 F1=0.07, ROUGE-L F1=0.16). Similar substantial improvements due to fine-tuning were observed in the other models (T5-small, T5-base, PEGASUS-PubMed), with T5-base improving its ROUGE-1 F1 from 0.20 (zero-shot) to 0.37 (fine-tuned), demonstrating nearly double performance improvement.[56]	Fine-tuning improved performance
Documentation	De-identification of Protected Health Information (PHI) in medical records	Did not mention	ChatGPT-3.5, GPT-4, PaLM	Did not mention	zero-shot GPT-3.5 with brief prompts reached an accuracy of 94.38%, whereas fine-tuned GPT-3.5 with detailed prompts achieved 99.20% accuracy, marking an absolute improvement of nearly 5%. Similarly, fine-tuning the PaLM model notably increased accuracy from 85.26% (zero-shot, brief prompts) to 98.36% (fine-tuned, detailed prompts), representing an impressive absolute improvement of approximately 13.1%.[57]	Fine-tuning improved performance. It is ten times cheaper to fine-tune the GPT-3.5 model than using the GPT-4 model directly, although there is a small single time fine-tuning charge associated and also limited engineering efforts required.

Table 4. The Effectiveness of Multimodal Data Fusion towards Performance Improvement of Applying Generative LLMs to EHR in Clinical Tasks

Clinical Task Category	Detailed Clinical Task	Data Fusion Method[58]	Involved Modality	Quantitative Performance Summary	Notable Findings
Clinical Decision Support	Disease diagnosis	Early	Structured Data; Text	Multimodal data fusion in the EHR-KnowGen model improved Micro F1 by 12.8%–14.6%, Macro F1 by 13.6%–14.5%, and accuracy by 3.2%–13.5% compared to single-modality baselines using only medical notes. For instance, Micro F1 increased from ~0.259 (BERT) and ~0.263 (GPT2) to ~0.297 with multimodal integration.[59]	Multimodal data fusion improved performance
	Diagnosing and classifying pressure injuries	Late	Text; Image	Multimodal integration in the YOLOv7-based system for pressure injury classification led to a high average F1 score of 0.9238, outperforming prior unimodal models such as MobileNetV2 (0.60–0.86), Faster R-CNN (0.6956), and SE-ResNext101 (0.7910). This represents a 7.1%–53.9% relative improvement, highlighting the clear quantitative advantage of the multimodal approach.	Multimodal data fusion improved performance
	Interpret radiological images for various clinical tasks, including diagnosis, formulation of treatment plans, and identification of abnormal findings.	Early	Text; Image	The multimodal data fusion method employed involves ChatGPT-4V integrating textual patient histories with visual radiographic data to perform clinical reasoning. The paper quantitatively emphasizes the improvements achieved through multimodal integration. For example, diagnostic accuracy for USMLE-style questions improved significantly when multimodal fusion was used (textual patient history + images: 77.01% accuracy) compared to using images alone (19.54% accuracy). Thus, multimodal fusion increased accuracy by 57.47 percentage points, representing approximately 294% relative improvement, with statistical significance ($P < 0.0001$).[60]	Multimodal data fusion improved performance
	Recognizing and diagnosing pathological conditions in radiological images	Intermediate	Text; Image	Quantitative analysis revealed substantial performance differences between multimodal GPT-4V (image + text) and single-modal GPT-4 (text only): For primary diagnoses identification, GPT-4V achieved an accuracy of 27%, significantly outperforming GPT-4's accuracy of 18%. Including differential diagnoses increased GPT-4V's accuracy to 35%, compared to GPT-4's 28%. In answering complex radiological multiple-choice questions, GPT-4V reached an accuracy of 64%, whereas GPT-4 (text only) obtained an accuracy of 47%. This represents a statistically significant performance increase ($P < .001$) of 17 percentage points, or approximately a 36.2% relative improvement, clearly demonstrating the quantitative benefits of multimodal integration. Across various radiological subspecialties, GPT-4V consistently outperformed GPT-4, with significantly higher accuracy in 15 out of 16 evaluated subspecialties. For instance, GPT-4V had notably higher performance in complex domains such as gastrointestinal imaging, neuroradiology, and musculoskeletal imaging, among others. The only exception was cardiac imaging, where GPT-4V performed worse (36% accuracy) compared to GPT-4 (57% accuracy). In some subspecialties like breast imaging and obstetrics/gynecology, both models performed equally.[61]	Multimodal data fusion does not necessarily improve performance.
	Classifying cases as COVID-19 or non-COVID-19 and further distinguishing COVID-19 variants like Omicron from previous strains.	Intermediate	Text; Image	For the COVID-19 classification task, in retrospective studies, the Med-MLLM achieved AUC scores of 95.3% for image-only, 93.8% for text-only, and 95.9% for image-text fusion. In prospective studies, the AUC scores were 64.8% for image-only, 72.9% for text-only, and 78.2% for image-text fusion.[62]	Multimodal data fusion enables accurate rare disease diagnosis
	Predicting patient survival or death outcomes following hospitalization due to COVID-19	Intermediate	Text; Image	For the COVID-19 prognosis task, in retrospective studies, the Med-MLLM achieved AUC scores of 81.2% for image-only, 88.3% for text-only, and 92.0% for image-text fusion. The results were the same in prospective studies, with AUC scores of 81.2% for image-only, 88.3% for text-only, and 92.0% for image-text fusion.[62]	Multimodal data fusion enables accurate rare disease diagnosis
Information Extraction	Detecting and resolving laterality errors in radiology reports	Early	Text; Image	The multimodal G-AI prototype (ATARI), when utilizing text alone, already demonstrated high accuracy, correctly identifying the absence of laterality mismatches (NLP false-positives) with an accuracy of 97.4% (115 out of 118). However, integrating both image and text queries (multimodal fusion), ATARI's accuracy further improved to 98.3% (116 out of 118). Similarly, using image query alone, ATARI also achieved an accuracy of 98.3% (116 out of 118), emphasizing that the visual modality substantially complements and enhances textual analysis.[42]	Multimodal data fusion improved performance
Documentation	Automatically generating medical reports from radiology images.	Intermediate	Text; Image	For the COVID-19 reporting task, their proposed Med-MLLM surpassed earlier methods by 5.5%, 3.6%, and 4.6% in terms of Precision, Recall, and F1 scores, respectively.[62]	Multimodal data fusion improved the performance of the documentation task.

Table 5. The Effectiveness of Ensemble of The LLM and Traditional AI towards Performance Improvement of Applying Generative LLMs to EHR in Clinical Tasks

Clinical Task Category	**Detailed Clinical Task**	**Ensemble Method**	**Performance Summary**	**Notable Findings**
Clinical Decision Support	Detecting early signs of cognitive decline (such as mild cognitive impairment and dementia) from clinical notes	Majority vote	The ensemble model combining GPT-4, DNN, and XGBoost via majority voting achieved the highest performance, with a precision of 90.2%, recall of 94.2%, and F1 score of 92.1%. Compared to the best single model (XGBoost, F1: 85.4%), the ensemble improved F1 by 6.7 percentage points, demonstrating statistically significant gains by leveraging the complementary strengths of individual models.[21]	Ensemble improved performance
	Predicting patient admissions from the emergency department	Adding ML prediction results to the LLM's prompt	Enhancing GPT-4 with RAG-based few-shot prompts improved accuracy from 77.5% to 81.3%, AUC from 0.79 to 0.82, and AUPRC from 0.48 to 0.56. Further integrating traditional ML model probabilities into prompts boosted performance to 83.1% accuracy, 0.87 AUC, and 0.71 AUPRC[46]	Ensemble improved performance

References

1. 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
2. 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
3. Beltagy I, Peters ME, Cohan A. Longformer: The Long-Document Transformer. Published online December 2, 2020. doi:10.48550/arXiv.2004.05150
4. 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
5. 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
6. OpenAI, Achiam J, Adler S, et al. GPT-4 Technical Report. Published online December 18, 2023. doi:10.48550/arXiv.2303.08774
7. 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
8. 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
9. Bedi S, Liu Y, Orr-Ewing L, et al. Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review. *JAMA*. 2025;333(4):319-328. doi:10.1001/jama.2024.21700
10. 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
11. 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
12. Zhang J, Sun K, Jagadeesh A, et al. The potential and pitfalls of using a large language model such as ChatGPT, GPT-4, or LLaMA as a clinical assistant. *J Am Med Inform Assoc JAMIA*. 2024;31(9):1884-1891. doi:10.1093/jamia/ocae184
13. 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
14. Bejan CA, Reed AM, Mikula M, et al. Large Language Models Improve the Identification of Emergency Department Visits for Symptomatic Kidney Stones. *MedRxiv Prepr Serv Health Sci*. Published online August 13, 2024:2024.08.12.24311870. doi:10.1101/2024.08.12.24311870
15. Shah-Mohammadi F, Finkelstein J. Extraction of Substance Use Information From Clinical Notes: Generative Pretrained Transformer-Based Investigation. *JMIR Med Inform*. 2024;12:e56243. doi:10.2196/56243
16. Shifai N, van Doorn R, Malvehy J, Sangers TE. Can ChatGPT vision diagnose melanoma? An exploratory diagnostic accuracy study. *J Am Acad Dermatol*. 2024;90(5):1057-1059. doi:10.1016/j.jaad.2023.12.062
17. 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
18. Williams CYK, Miao BY, Kornblith AE, Butte AJ. Evaluating the use of large language models to provide clinical recommendations in the Emergency Department. *Nat Commun*. 2024;15(1):8236. doi:10.1038/s41467-024-52415-1
19. Gupta S, Basu A, Nievas M, et al. PRISM: Patient Records Interpretation for Semantic clinical trial Matching system using large language models. *NPJ Digit Med*. 2024;7(1):305. doi:10.1038/s41746-024-01274-7
20. Zheng NS, Keloth VK, You K, et al. Detection of Gastrointestinal Bleeding With Large Language Models to Aid Quality Improvement and Appropriate Reimbursement. *Gastroenterology*. Published online September 18, 2024:S0016-5085(24)05467-2. doi:10.1053/j.gastro.2024.09.014
21. Du X, Novoa-Laurentiev J, Plasek JM, et al. Enhancing early detection of cognitive decline in the elderly: a comparative study utilizing large language models in clinical notes. *EBioMedicine*. 2024;109:105401. doi:10.1016/j.ebiom.2024.105401
22. Ge J, Li M, Delk MB, Lai JC. A Comparison of a Large Language Model vs Manual Chart Review for the Extraction of Data Elements From the Electronic Health Record. *Gastroenterology*. 2024;166(4):707-709.e3. doi:10.1053/j.gastro.2023.12.019

23. 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
24. Lehnen NC, Dorn F, Wiest IC, et al. Data Extraction from Free-Text Reports on Mechanical Thrombectomy in Acute Ischemic Stroke Using ChatGPT: A Retrospective Analysis. *Radiology*. 2024;311(1):e232741. doi:10.1148/radiol.232741
25. 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
26. 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
27. Zhang L, Liu M, Wang L, et al. Constructing a Large Language Model to Generate Impressions from Findings in Radiology Reports. *Radiology*. 2024;312(3):e240885. doi:10.1148/radiol.240885
28. 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
29. Steimetz E, Minkowitz J, Gabutan EC, et al. Use of Artificial Intelligence Chatbots in Interpretation of Pathology Reports. *JAMA Netw Open*. 2024;7(5):e2412767. doi:10.1001/jamanetworkopen.2024.12767
30. 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
31. Kumar AT, Wang C, Dong A, Rose J. Generation of Backward-Looking Complex Reflections for a Motivational Interviewing-Based Smoking Cessation Chatbot Using GPT-4: Algorithm Development and Validation. *JMIR Ment Health*. 2024;11:e53778. doi:10.2196/53778
32. 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
33. Sahoo SS, Plasek JM, Xu H, et al. Large language models for biomedicine: foundations, opportunities, challenges, and best practices. *J Am Med Inform Assoc JAMIA*. 2024;31(9):2114-2124. doi:10.1093/jamia/ocae074
34. Kim TT, Makutonin M, Sirous R, Javan R. Optimizing Large Language Models in Radiology and Mitigating Pitfalls: Prompt Engineering and Fine-tuning. *Radiogr Rev Publ Radiol Soc N Am Inc*. 2025;45(4):e240073. doi:10.1148/rg.240073
35. AlSaad R, Abd-Alrazaq A, Boughorbel S, et al. Multimodal Large Language Models in Health Care: Applications, Challenges, and Future Outlook. *J Med Internet Res*. 2024;26:e59505. doi:10.2196/59505
36. Maharjan J, Garikipati A, Singh NP, et al. OpenMedLM: prompt engineering can out-perform fine-tuning in medical question-answering with open-source large language models. *Sci Rep*. 2024;14(1):14156. doi:10.1038/s41598-024-64827-6
37. Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. Large Language Models in Medicine: The Potentials and Pitfalls: A Narrative Review. *Ann Intern Med*. 2024;177(2):210-220. doi:10.7326/M23-2772
38. Nagarajan R, Kondo M, Salas F, et al. Economics and Equity of Large Language Models: Health Care Perspective. *J Med Internet Res*. 2024;26:e64226. doi:10.2196/64226
39. Feng J, Phillips RV, Malenica I, et al. Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare. *NPJ Digit Med*. 2022;5(1):66. doi:10.1038/s41746-022-00611-y
40. 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
41. Bhayana R, Nanda B, Dehkharghanian T, et al. Large Language Models for Automated Synoptic Reports and Resectability Categorization in Pancreatic Cancer. *Radiology*. 2024;311(3):e233117. doi:10.1148/radiol.233117
42. Kathait AS, Garza-Frias E, Sikka T, et al. Assessing Laterality Errors in Radiology: Comparing Generative Artificial Intelligence and Natural Language Processing. *J Am Coll Radiol JACR*. 2024;21(10):1575-1582. doi:10.1016/j.jacr.2024.06.014
43. Gertz RJ, Dratsch T, Bunck AC, et al. Potential of GPT-4 for Detecting Errors in Radiology Reports: Implications for Reporting Accuracy. *Radiology*. 2024;311(1):e232714. doi:10.1148/radiol.232714

44. 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
45. 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
46. Glicksberg BS, Timsina P, Patel D, et al. Evaluating the accuracy of a state-of-the-art large language model for prediction of admissions from the emergency room. *J Am Med Inform Assoc JAMIA*. 2024;31(9):1921-1928. doi:10.1093/jamia/ocae103
47. 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
48. 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
49. 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
50. Cheligeer K, Wu G, Laws A, et al. Validation of large language models for detecting pathologic complete response in breast cancer using population-based pathology reports. *BMC Med Inform Decis Mak*. 2024;24(1):283. doi:10.1186/s12911-024-02677-y
51. Cho H, Yoo S, Kim B, et al. Extracting lung cancer staging descriptors from pathology reports: A generative language model approach. *J Biomed Inform*. 2024;157:104720. doi:10.1016/j.jbi.2024.104720
52. Sivarajkumar S, Tam TYC, Mohammad HA, et al. Extraction of sleep information from clinical notes of Alzheimer's disease patients using natural language processing. *J Am Med Inform Assoc JAMIA*. 2024;31(10):2217-2227. doi:10.1093/jamia/ocae177
53. Patra BG, Lepow LA, Kasi Reddy Jagadeesh Kumar P, et al. Extracting social support and social isolation information from clinical psychiatry notes: comparing a rule-based natural language processing system and a large language model. *J Am Med Inform Assoc JAMIA*. 2025;32(1):218-226. doi:10.1093/jamia/ocae260
54. Goswami J, Prajapati KK, Saha A, Saha AK. Parameter-efficient fine-tuning large language model approach for hospital discharge paper summarization. *Appl Soft Comput*. 2024;157:111531. doi:10.1016/j.asoc.2024.111531
55. López-Úbeda P, Martín-Noguerol T, Díaz-Angulo C, Luna A. Evaluation of large language models performance against humans for summarizing MRI knee radiology reports: A feasibility study. *Int J Med Inf*. 2024;187:105443. doi:10.1016/j.ijmedinf.2024.105443
56. Sezgin E, Sirrianni JW, Kranz K. Evaluation of a Digital Scribe: Conversation Summarization for Emergency Department Consultation Calls. *Appl Clin Inform*. 2024;15(3):600-611. doi:10.1055/a-2327-4121
57. Yashwanth YS, Shettar R. Zero and Few Short Learning Using Large Language Models for De-Identification of Medical Records. *IEEE Access*. 2024;12:110385-110393. doi:10.1109/ACCESS.2024.3439680
58. Mataraso SJ, Espinosa CA, Seong D, et al. A machine learning approach to leveraging electronic health records for enhanced omics analysis. *Nat Mach Intell*. Published online January 16, 2025:1-14. doi:10.1038/s42256-024-00974-9
59. 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
60. Zhu L, Mou W, Lai Y, et al. Step into the era of large multimodal models: a pilot study on ChatGPT-4V(ision)'s ability to interpret radiological images. *Int J Surg Lond Engl*. 2024;110(7):4096-4102. doi:10.1097/JS9.0000000000001359
61. Busch F, Han T, Makowski MR, Truhn D, Bressem KK, Adams L. Integrating Text and Image Analysis: Exploring GPT-4V's Capabilities in Advanced Radiological Applications Across Subspecialties. *J Med Internet Res*. 2024;26:e54948. doi:10.2196/54948
62. 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