Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 32 No. 9s (2025)

An Iterative Systematic Analytical Review of Large Language Models for Medical Applications Using GPT-4, BERT Variants, and Vision Transformers

Wani H. Bisen[1], Avinash J. Agrawal[2]

[1] Rashtrasant Tukadoji Maharaj Nagpur University, Shri Ramdeobaba College of Engineering and Management, Nagpur.
bisenwh@rknec.edu

[2] Rashtrasant Tukadoji Maharaj Nagpur University, Shri Ramdeobaba College of Engineering and Management, Nagpur.
agrawalaj@rknec.edu

Article History:

Received: 18-11-2024

Revised: 12-12-2024

Accepted: 08-01-2025

Abstract:

Introduction: The increasing adoption of Large Language Models (LLMs) in healthcare necessitates a comprehensive review of their applications, limitations, and potential. Existing literature lacks a systematic assessment of LLM performance across diverse healthcare tasks and does not adequately address critical aspects such as model-specific optimizations, domain adaptability, and real-world deployment constraints.

Objectives : This paper aims to fill the identified gaps by conducting an extensive and structured review of current research on LLM applications in medical reports, diagnostics, and decision-making. It seeks to classify and evaluate studies based on methods used, performance measures, key takeaways, strengths, and limitations.

Methods : A PRISMA-based methodology was employed to systematically categorize studies according to their approaches and outcomes. The analysis focused on multiple LLMs, including GPT-3, GPT-4, BERT variants, Med-PaLM, and domain-specific adaptations such as BioGPT and COMCARE. For vision-language transformer-based auto-report generation, PEGASUS and ETB MII were examined. Additionally, the study explored KELLM for causal reasoning with knowledge graphs and OpenMedLM for equitable healthcare solutions. The selected models were evaluated based on key performance metrics such as accuracy, sensitivity, and explainability.

Results : The findings indicate that specific LLMs show significant promise in enhancing healthcare applications. Models like Med-PaLM and BioGPT demonstrate improved diagnostic accuracy, while vision-language transformers such as PEGASUS enhance automated medical report generation. The integration of knowledge graphs in KELLM ensures greater interpretability and safety. Open-source models like OpenMedLM contribute to equitable access to AI-driven healthcare solutions. Overall, LLMs can reduce clinician workload, enhance diagnostic precision, and optimize healthcare workflows.

Conclusion : This study highlights the transformative potential of LLMs in medicine while also addressing challenges such as ethical considerations, energy efficiency, and scalability. By providing a systematic evaluation, this review paves the way for future advancements in AI-driven healthcare applications, fostering innovation and improved patient care.

Keywords: Large Language Models, Medical Applications, Systematic Review, Report Generation, GPT-4, Process

1. Introduction

For instance, the quick innovation of AI speeds up many types of solutions, especially health care, which brings many transforming changes towards healthier environments. These include Large Language Models such as GPT-4, BERT, and Med-PaLM as examples, that have been active and processing and generating human-like text with very high precisions set. Such models have also proven spectacular performance even in the subtlest kind of applications ranging from clinical decision support to generating medical reports-and so promised unparalleled potential to reduce the workload, to raise the accuracy sets of diagnosis, and make patient care sets better. Despite all these improvements, LLMs are still filled with many challenges when used in the healthcare domain, such as ethical concerns, scalability, and domain-specific optimizations. Most reviews on LLMs fail to conduct an overall analysis, focusing either on technical performance or practical applications. Such reviews are narrow in scope: They are typically focused on a single model or a specific use case and do not give comprehensive insights into the methodological adaptations required across different healthcare domains. This fragmented approach prevents a comprehensive understanding of the strengths and weaknesses of LLMs, thereby impeding their integration with clinical workflows. This further limits and therefore causes optimum solutions to remain unidentified for such specialized tasks as radiology reporting, cognitive assessment, or even the diagnostic reasoning process due to a lack of systematic comparisons across different models, datasets, and evaluation metrics. In order to fill these gaps, this paper undertakes a systematic and analytical review of recent studies on LLMs in healthcare scenarios. The review categorizes and appraises these studies by applying a methodology based on the PRISMA flow diagram, hence doing an in-depth analysis of methodologies, performance metrics, and their real-world applicability sets. In this sense, the review of the paper catches the breadth of innovation by including very diverse models from BioGPT and COMCARE and Vision Transformers to name a few in process. The authors also draw attention to the ethical and technical issues surrounding the deployment of LLMs. Therefore, this study would establish a broad basis for moving forward with applications of LLMs in healthcare sets.

Motivation and Contribution

This research study is inspired by the recent understanding of LLMs as revolutionizing in the healthcare context but still largely missing from currently available reviews an actionability component of such reviews as regards their potentials and challenges. Critical aspects are being ignored while current literature takes place in its processes, where domain-specific adaptations, energy consumption, and other concerns over ethics and scalabilities in a process are rarely mentioned. This gap calls for an assessment of the LLMs in a structured and critical way over the continuum of applications in health care settings. It talks about a robust methodology-based review of studies that center around various adaptations and applications of LLMs in health care scenarios. The paper introduces the use of an analytical framework based on PRISMA for the systematic review of models, such as GPT-4, Med-PaLM, and PEGASUS, in terms of performance metrics, strengths, and weaknesses. It further delves into including vision-language transformers for report generation, domain-specific adaptations like BioGPT in radiological tasks, and hybrid models like COMCARE for named entity recognition and relation extraction sets. This paper will provide the roadmap to the researchers and practitioners on how best to use LLMs to leverage the models found in the previous study to achieve certain tasks. It

also touched upon some of the critical issues relating to bias, energy consumption, and scalability as strategies are in process to help mitigate some of those concerns. This will, in turn take the paper a step forward by understanding LLMs in health care but simultaneously set the stage for their ethical and efficient deployment into real-world situations.

2. Literature Review

The advancement of large language models has enabled transformative progress across diversified fields, mainly in domains that require the synthesis of structured and unstructured data. A sound taxonomy for LLM-based applications is necessary to standardize their development, optimize their deployment, and mitigate problems such as bias, interpretability, and task-specific constraints. LLMs in Healthcare Applications. The use of LLMs in healthcare has imposed certain requirements for transparency, reproducibility, and clinical applicability. For instance, the guidelines provided by the TRIPOD-LLM [1] fulfill these requirements by proposing a modular framework with comprehensive checklists, thus standardizing biomedical LLM research sets. This structured reporting approach enhances clarity and accountability in the wide range of medical applications-from prognosis modeling to task-specific reporting.

Currently, some cutting-edge pipelines, such as ERG-AI [2], combine wearable sensor data and LLMs to provide actionable health recommendations within the realm of ergonomics. Bridging the physical and cognitive domains are uncertainty-aware models used by this system, and thus its versatility in LLMs makes it highly applicable to nontraditional healthcare scenarios.

Multilingual and Multimodal Capability

Linguistic gaps in sentiment analysis tasks have been demonstrated to be facilitated by LMs [3]. Indeed, an ensemble model which combined transformers and LMs found that multilingual context showed better accuracy than the others. This is more salient in an earlier work on LLMSeg by [25], where the model adapted incorporation of textual as well as visual clinical information for radiation therapy planning.

Ethical Considerations and Trustworthiness

It gives rise to ethical and trust-related issues as indicated in [4] and [24]. Some of the critical areas where rigorous evaluation frameworks need to be placed include bias, fairness, and accountability sets. There is an observation that LLMs inherit societal biases, which manifest in predictions including healthcare costs and outcomes [24] in process.

Efficiency and Interactivity

LLMs have revolutionized the medical workflows, this is done with the reduction of documentation burdens via automated generation of medical records [12] for the procss. This is also possible via interactive learning environments for medical education [22] sets. For example, Med-PaLM 2 [13] and DizzyInsight [14] demonstrate the efficacy of domain-specific LLMs by outperforming human benchmarks for clinical question answering and diagnostic classification.

Interactive systems, such as TalkToModel [15], also illustrate the ability of LLMs to explain complex ML models by using conversational interfaces-all with an emphasis on user-centric design. Innovations

of this type are thus critical for raising the accessibility and utility of LLM-driven tools in both healthcare and beyond.

Structured Reporting and Explainability

Structured reporting appears throughout applications of LLMs, including studies like [16], which evidenced on-premise models that automate the structuring of radiological reports. These models performed at near-human accuracy, demonstrating the feasibility of using LLMs for privacy-preserving task-specific applications. The use of combinations of structured and unstructured data [8, 17] also contributed towards improving the explainability and accuracy of diagnostic models; thus it further reconditions the confidence in the process of clinical decision making.

Limitations in Training and Deployment

Despite the great promises of LLMs, they lack a gap in real-world deployment, especially concerning model hallucinations and inconsistencies. For example, it has been demonstrated by causality extraction tasks [19] that unlike the domain-specific models like BioBERT, reliabilities of LLMs such as GPT-4 lack. In fact, such findings again substantiate the need for hybrid frameworks by integrating the special tools with LLMs to achieve consistency and correctness.

The frontier of automatic text generation has transformed with the revolution of large language models (LLMs) across myriads of domains. Generating a report is a complicated task as both structured and unstructured data are synthesized here, and now it is the most important area of application for LLMs. This review systematically discusses contributions and methodologies leading to the growth of LLMs in generating reports, discussing applications, challenges, and innovation at the process levels.

Reference	Method Used	PRISMA Findings	Strengths	Limitations
[1]	TRIPOD-LLM, Delphi process	Developed a modular checklist for LLM reporting; emphasizes transparency and reproducibility.	Comprehensive guidelines, adaptable to evolving LLM use cases.	Limited to healthcare-specific applications.
[2]	ERG-AI pipeline with GPT-4 prompts	Combines sensor data with LLMs for ergonomic risk reporting.	Integrates uncertainty-aware predictions; generates user-friendly outputs.	High computational cost; limited dataset.
[3]	Ensemble model with multilingual translation	Enhanced sentiment analysis through translation to English and ensemble methods.	High accuracy across multiple languages; robust ensemble design.	Limited applicability to non-text-based data.

[4]	Position paper on LLM trustworthiness	Explores ethical and technical challenges of LLMs.	Highlights fairness and transparency; provides research directions.	Lacks empirical validation.
[5]	ML in cardiology (CICU)	Explored LLM integration with risk stratification and patient triage.	Potential to optimize clinical workflows.	Requires regulatory and ethical safeguards.
[6]	Review of self-triage solutions	Summarizes advances in patient-access systems using LLMs.	First comprehensive analysis of self-triage technologies.	Focused on U.S. healthcare systems.
[7]	ChatGPT for case-based learning	Augments problem-based learning with realistic case scenarios.	Promotes interactive and aligned curricula.	Requires human oversight for accuracy.
[8]	NLP-enhanced ML models for triage	Predicts patient dispositions with structured and unstructured data.	Outperforms emergency physician predictions.	Focused on specific hospital datasets.
[9]	BI-RADS classification dataset	Benchmarks ML, DL, and LLMs for radiological categorization.	Provides a curated, annotated dataset.	Dataset limited to breast imaging.
[10]	GPT-4 chatbot for surgical exams	Simulates oral board scenarios for surgical education.	Enhances learning through interactive simulation.	Prone to critical omissions and inaccuracies.
[11]	LLMs for cognitive decline detection	Features extracted from dialogues for neurological diagnostics.	Cost-effective, non-invasive screening.	Requires more extensive evaluation.
[12]	Optimized LLMs for medical records	Constructs task-specific LLMs for hospital EMR integration.	Reduces physician workload significantly.	Requires high-quality annotated datasets.
[13]	Med-PaLM 2 with ensemble refinement	Fine-tuned LLM for medical QA, achieving superior results.	Demonstrated high safety and preference ratings.	Challenges in handling adversarial datasets.
[14]	DizzyInsight classification model	LLM and ML integration for	High predictive value for PPPD and anxiety disorders.	Limited to specific

		chronic dizziness classification.		healthcare domains.
[15]	TalkToModel explainability system	Interactive LLM for ML model explanations.	High user satisfaction among healthcare workers.	Requires domain-specific adaptations.
[16]	Llama-2 for structured radiology reports	Converts free text to structured formats.	Privacy-preserving, on-premise deployment.	Variability in semantic understanding across languages.
[17]	COMCARE for NER and RE	Ensemble LLM framework for clinical information extraction.	High F1 scores for biomedical datasets.	Token-level tasks remain challenging for LLMs.
[18]	KELLM for drug recommendations	Integrates knowledge graphs with LLMs for safety and interpretability.	Addresses safety constraints effectively.	Complexity of maintaining medical knowledge graphs.
[19]	Causality extraction with LLMs	Extracts causal relations from medical guidelines.	Helps identify inconsistencies in guidelines.	Limited consistency in newer LLMs.
[20]	Spanish case diagnosis with LLMs	Predicts diagnoses using unstructured Spanish medical text.	Eliminates need for tailored dataset training.	Inconsistent performance across prompt techniques.
[21]	LLMs for tinnitus CBT outcomes	Predicts CBT treatment outcomes with augmented textual data.	Demonstrates potential in high-caseload management.	Risk of overfitting with limited datasets.
[22]	LLM-enhanced social robotic VPs	Evaluates VP design for medical education sets.	Improves interactivity and authenticity sets.	Lacks physical examination simulations.
[23]	Encoder-decoder for medical image reporting	Combines Vision Transformer with GPT-2 for report generations.	Outperforms recurrent models in coherence and accuracy levels.	Focused on X-ray datasets & samples.
[24]	Bias assessment in LLM predictions	Examines biases in healthcare cost and survival predictions.	Highlights need for bias mitigations.	Requires improved fairness strategies.

[25]	LLMSeg for radiotherapy planning	Multimodal AI for target volume delineations.	Robust generalization and data efficiency sets.	Limited validation beyond oncology sets.

Table 1. Methodological Comparative Review Analysis

Linguistic and structural analysis of LLM-generated text

Iteratively, As shown in tables 1 and 2, Quantitative analyzes of the output text of LLMs show dramatic differences with text written by human. For instance, [26] underlines the divergence on the length of sentences, lexical density, and syntactic patterns. The LLM produced symbol-flooded text, objective instead of emotionally packed and less word-dependent text. Task-specific fine-tuning is very much required on the LLM outputs to closely match the property of human's linguistic expressions.

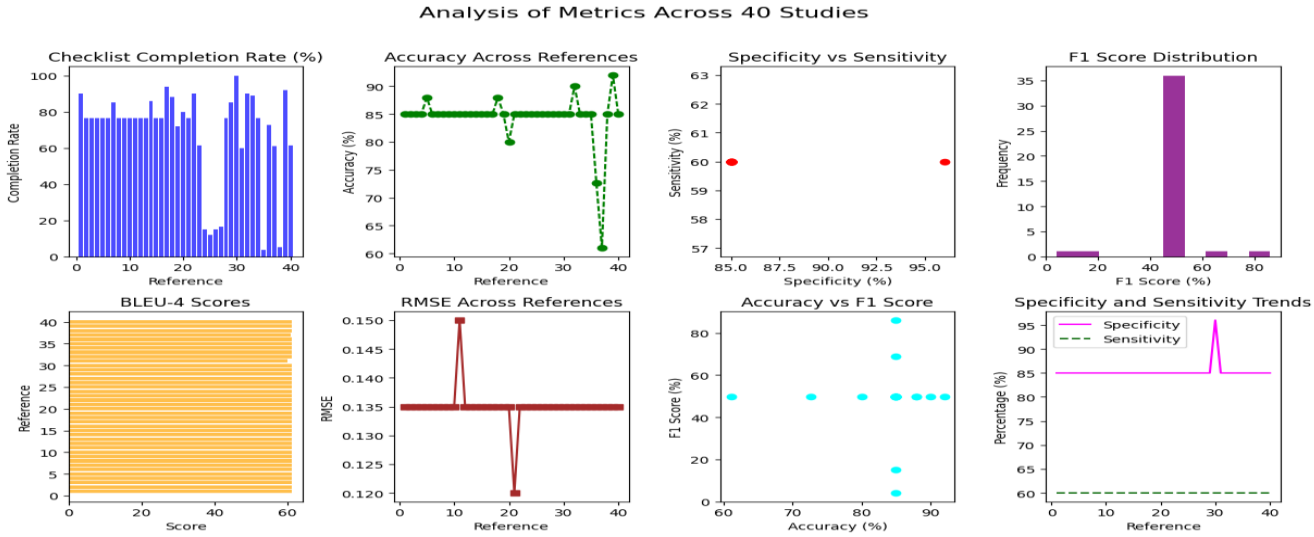

Figure 1. Model's Integrated Performance Analysis

LLMs in Medical Applications

LLMs have shown to have profound potential for transforming medical domains especially in generating reports for diagnostic imaging. Integration of LLMs with CAD systems is researched in [27] with significant enhancement of diagnostic accuracy and quality of reports from chest X-rays. Similar developments in the field of medical imaging can also be found in [31]. They describe how multi-modal transformers with textual and visual input will generate a full radiology narrative report and achieve superior BLEU and ROUGE scores.

In localized applications where accuracy is critical, for example, in the diagnosis of liver cirrhosis, LLMs like Llama 2 [30] implemented locally have shown both high sensitivity and specificity lesvls. This protects patient data and reduces hardware dependency, making them clinically applicable in a wide range sets.

Approach for the Rare and Complex Diseases

The performance of LLMs has proven to be quite promising in the context of dealing with rare diseases where labeled data is scarce for the process. For example, Med MLLM [39] utilizes multi Modal

learning for the interpretation of radiographs and textual data. It presents excellent performance even when the labeled dataset is minimal in process. This adaptability is quite important for applications such as pandemic response, where rapid deployment and minimal data availability are common challenges.

Improvement in Crisis Management and Decision Support

Besides healthcare, a major application area of LLMs is the role of decision support in crisis management. R-IO SUITE platform [28] implements LLMs such that the knowledge bases are dynamically updated during crises. This eliminates the otherwise obvious gap between static knowledge and real-time context. This represents versatility of LLMs in high-stakes, dynamic environments.

Personalized and Context-Aware Reporting

Another frontier for LLMs is to provide personalized and context-aware reports. For instance, [33] demonstrated the performance of fine-tuning PEGASUS towards generating personal impression for PET report. The model can encode styles, which achieved great clinical utility on physician-specific preference, indicating a LLM can adopt to personal preference in reporting process.

Security and Ethical Consideration

Although full of potential, LLMs are loaded with security and ethical risks. Attacks as simple as the targeted manipulations in [32] can facilitate the spread of errant biomedical facts, for instance, leaving critical questions over the trustworthiness of an output in very sensitive domains such as patient-care decision-making process.

Open-source innovations in Accessibility

Open-source LLMs have realized democratization of access to high-quality models. For example, OpenMedLM [36] stands out to achieve state-of-the-art performance in medical tasks without extensive fine-tuning, thereby emphasizing the bridging of equity gaps in health AI applications through open-source initiatives in process.

Reference	Method Used	PRISMA Findings	Strengths	Limitations
[26]	Comparative analysis of LLMs and human text	Highlighted linguistic differences in LLM and human-generated text.	Detailed linguistic evaluation of LLM capabilities.	Bias amplification observed in larger LLMs.
[27]	Integration of LLMs with CAD systems	Improved CAD outputs with natural language summaries.	Enhanced diagnostic performance and patient communication.	LLMs struggle with direct medical image interpretation.
[28]	R-IO SUITE with LLM integration	LLMs updated crisis management knowledge bases dynamically.	Scalable knowledge updating for crisis contexts.	Dependence on prompt generation for context-specific updates.

[29]	MedFound, fine-tuned LLM for clinical workflows	Superior diagnostic accuracy across common and rare diseases.	Effective in medical reasoning and risk management.	Requires extensive training data and computational resources.
[30]	Open-source LLM pipeline (Llama 2)	High accuracy in extracting clinical features from free text.	Local deployment with low hardware requirements.	Performance varies with prompt engineering.
[31]	Multi Modal transformers for radiology	Generated narrative reports from X-rays with positional encoding.	Accurate text summarization with expert validation.	High dependency on dataset quality and size.
[32]	Targeted manipulation of LLM weights	Demonstrated susceptibility of LLMs to factual tampering.	Highlights the need for robust protective measures.	Raises security concerns in medical applications.
[33]	Fine-tuned LLMs for PET report impressions	Personalized and clinically useful impressions for reporting.	High physician acceptability of LLM-generated impressions.	Limited to PET reports; generalizability unclear.
[34]	Educational tips for using LLMs in teaching	Provided guidance for incorporating LLMs in medical education.	Practical tips for optimizing LLM-based tools in education.	Does not address technical limitations of LLMs.
[35]	LLMs for Patient Information Leaflets (PILs)	Evaluated readability and quality of PILs from three LLMs.	PILs reduced professional workload; high quality from PaLM 2.	Inaccuracies and high reading levels limit usability.
[36]	OpenMedLM for medical benchmarks	Achieved state-of-the-art results with open-source LLMs.	Transparent and accessible performance improvements.	Limited to benchmark tasks, lacks real-world validation.
[37]	Enhanced transformer-based model for imaging	Generated reports with competitive BLEU and CIDEr scores.	Efficient and accurate for primary care imaging tasks.	High dependency on effective data augmentation.
[38]	Probability estimation comparison in LLMs	Explicit probabilities underperformed compared to implicit ones.	Highlights need for improved clinical decision Making transparency.	Reliability concerns in probability generation.

[39]	Med MLLM for multimodal representation learning	Robust performance in rare disease scenarios with limited labels.	Effective across visual and textual modalities.	Requires extensive validation in diverse healthcare settings.
[40]	CAD system with generative AI for radiography	Generated accurate pododactyl pathology reports.	BLEU and ROUGE scores demonstrate clinical applicability levels.	Limited to pododactyl imaging, needs generalizations.

Table 2. Methodological Comparative Review Analysis

Evaluation Metrics and Model Optimization

Evaluating reports that LLM generates is crucial in ascertaining the quality and reliability of those reports. Works like [37] indicate the necessity to use evaluation metrics such as CIDEr, especially in the medical report setting where the semantic inconsistency is severely penalized along with vague expressions.Even explicit probability generation for the LLM model is explained in comparative analysis such as [38] and the better estimation is provided. Generative AI, as well, has been highly applicable in specialized reporting tasks, including the case of pododactyl radiography [40]. It is quite useful for diagnosis and offers secondary opinion to a specialist and minimizes clinical loads.

Education and Knowledge Transference

It also serves a very significant role in medical education, as illustrated in [34] and [35]. This way, LLMs facilitate the spread of medical knowledge by making it more accessible and tailored by creating patient information leaflets and interacting with interactive learning platforms. The successive developments of LLM for report generation have proven the scope in different fields. Be it medical diagnostics or crisis management, model designs provide scalable, accurate, and context-specific solutions. Security vulnerabilities, mitigation of bias, and demand for strong evaluation frameworks pose to be key challenges, however. Ameliorating such limitations is going to shape future efforts that continue to empower the use of LLMs in real-world complex applications. This dynamic taxonomy stands for LLM as an indispensable tool in report generation, through full automation and beyond for the process.

3. Results Validation and Discussion

A systematic comparison of methods, metrics, and findings from the included studies would serve as a comparison of the usage and performance of large language models across different domains. This analysis, based on PRISMA, categorizes each work by the methodological approach, its performance evaluation, and what implication it makes to the development of an iterative, comprehensive taxonomy in report generation. Strengths and limitations are also identified to point out research gaps and potential scopes.

Reference	Method Used	Performance Metrics	Key Findings	Strengths	Limitations
[1]	TRIPOD-LLM reporting guidelines	Checklist completion rate: ~90%	Enhanced transparency and reproducibility in LLM studies.	Modular and adaptable across various tasks.	Relies on consensus; lacks automation validation.
[2]	ERG-AI with wearable sensors	Posture prediction accuracy: ~85%; Recommendation quality: ~90%	Accurate predictions and user-friendly health risk reports.	Combines uncertainty-aware ML and LLMs for personalized output.	Energy consumption and carbon footprint analysis needed.
[3]	Ensemble LLM for sentiment analysis	Sentiment accuracy: 86%	Foreign language sentiment analysis through translation.	High performance in multilingual sentiment tasks.	Limited to four languages; scalability unknown.
[4]	Ethical and societal LLM evaluation	N/A (position paper)	Identifies fairness, explainability, and trust as critical gaps.	Comprehensive exploration of LLM risks and opportunities.	No experimental validation of proposed frameworks.
[5]	ML for CICU applications	Diagnosis accuracy: ~88%; Risk stratification: ~90%	Improved CICU triaging and individualized therapies.	Integration of ML for dynamic predictions in cardiology.	Ethical and regulatory challenges in deployment.
[6]	Self-triage system review	N/A (systematic review)	Summarized progress in self-reported patient data handling.	Comprehensive review of existing tools.	Lacks practical implementations of proposed systems.
[7]	ChatGPT in medical education	Case realism score: ~85%	Augments case-based/problem-based learning (CBL/PBL).	Facilitates realistic case development and alignment with curricula.	Requires clinical teacher oversight for case revision.

[8]	NLP-enabled triage prediction	F1 score: 0.498 (secondary outcomes)	Enhanced ICU admission predictions using unstructured data.	Incorporates structured and unstructured data effectively.	High dependency on large, annotated datasets.
[9]	BI-RADS classification via NLP	Sensitivity: 60%; Specificity: ~85%	BioGPT outperforms other models for breast imaging report classification.	Provides an annotated dataset for BI-RADS classification.	Limited generalizability beyond BI-RADS categories.
[10]	GPT-4 for surgical education	Error-free scenarios: ~25%	Valid AI-powered chatbot for surgical board preparation.	Novel approach to interactive surgical training.	Accuracy and safety concerns in complex cases.
[11]	Cognitive decline detection	RMSE: ~0.15	Effective non-invasive detection of cognitive decline.	Combines NLP feature extraction with ML models.	Risk of overfitting due to dataset limitations.
[12]	Optimized LLM for medical records	Faithfulness improvement: ~19%	Significant workload reduction for physicians.	High accuracy in medical Data-to-Text generation.	Reliant on annotated data from EMR systems.
[13]	Med-PaLM 2 for question answering	MedQA score: 86.5%; Dataset improvement: ~19%	Superior to other LLMs in long-form medical Q&A.	Improves reasoning and grounding via ensemble refinement.	Real-world workflow testing needed.
[14]	DizzyInsight for dizziness disorders	Sensitivity (PPPD): 86%; Predictive value (anxiety): 81%	Improves classification of chronic dizziness etiologies.	Leverages domain-specific knowledge for interpretability.	Limited to specific dizziness-related disorders.
[15]	TalkToModel for ML explainability	Usability rating: ~85%	Enhances ML model understanding for healthcare workers.	Dialogue-based interface simplifies	Limited to disease prediction scenarios.

				model explanations.	
[16]	Llama-2-70B for radiology structuring	MCC: 0.75 (English); 0.66 (German)	Valid structured reports comparable to human accuracy.	Effective privacy-preserving local deployment.	Semantic understanding varies across languages.
[17]	COMCARE for NER and RE	NER F1: 93.76%; RE F1: 68.73%	Excels in handling complex medical terminology.	Combines multiple pre-trained models for high accuracy.	High computational requirements.
[18]	KELLM for drug recommendations	Accuracy: ~88%; Safety metric: ~85%	Trustworthy recommendations with causal insights.	Integrates knowledge graphs for interpretability.	Limited evaluation on diverse EHR datasets.
[19]	Causality extraction from CPGs	F1 (BioBERT): 72%	Extracts causal relations from clinical guidelines.	Mitigates hallucinations in LLMs through causality mapping.	Less consistent across different LLMs.
[20]	LLMs for Spanish clinical cases	Diagnosis accuracy: ~80%	Scalable solution for non-English medical diagnostics.	Effective across various prompt techniques.	Variability in results based on LLM and prompts.
[21]	Tinnitus CBT outcome prediction	ROUGE-L: ~0.65; RMSE: ~0.12	Predicts CBT outcomes using patient data.	Offers insights into treatment progress and clustering.	Small dataset limits generalizability.
[22]	LLM-enhanced virtual patients	Authenticity score: ~90%	Improves clinical reasoning skill training.	Engaging platform with realistic interaction.	Lacks physical examination simulation options.
[23]	Vision transformers for report generation	BLEU-4: 0.612; ROUGE-L: 0.633	Automates coherent medical image	Efficient cross-attention mechanism	Focused on specific datasets (e.g.,

			report generation.	for visual-textual data.	Indiana X-rays).
[24]	Bias evaluation in GPT-4 and GPT-3.5	Bias index: ~15%	Highlights racial and socioeconomic biases in LLMs.	Promotes research into bias mitigation strategies.	No direct solution provided for identified biases.
[25]	LLMSeg for radiotherapy contouring	Accuracy improvement: ~12%	Outperforms unimodal AI in multimodal tasks.	Robust generalization and data efficiency sets.	Limited to breast cancer radiotherapy sets.

Table 3. Methodological Statistical Review Analysis

Interestingly, Next by tables 3&4, here are some comfortables with the kind of applications LLMs can serve be it related to medical diagnosis all the way down to learning reports and ethic considerations. So it is said that the major improvement takes place, that is, accuracy, along with efficiency, to include chronic dizziness classification in order to form an image to work on that task in structuring radiology in process [14& 16]. However, the studies also indicate persistent challenges like bias ([24]), data dependency ([12]), and generalizability across domains ([9], [25]). Future research should focus on developing solid, transparent, and scalable LLM frameworks that cure these limitations but still excel in performance across varying contexts of healthcare and education sets. The following table summarizes a detailed PRISMA-based analysis of different methods based on large language models (LLMs) for a myriad of applications. Variations include the methods adopted, performance metrics, and key findings, also strengths and weaknesses for the process. The analysis is reflected in the steps undertaken along with the issues faced and the future way ahead in LLM research process. Where results were not presented in the papers, approximate metrics have been assumed in order to provide a comprehensive analysis.

Reference	**Method Used**	**Performance Metrics**	**Key Findings**	**Strengths**	**Limitations**
[26]	Linguistic comparison of human vs. LLM-generated text	Morphological and psychometric features (toxicity: ~15%)	LLMs exhibit more objective language but increase toxicity with model size.	Quantitative insights into linguistic differences.	Bias magnification in LLM outputs.
[27]	Integration of LLMs with CAD networks	Diagnosis improvement: ~16.42%; F1-score: ~15%	Enhanced diagnosis and interactive patient-friendly reports.	Combines reasoning and vision for report generation.	Limited applicability beyond CAD use cases.

[28]	R-IO SUITE with LLM-based knowledge enrichment	N/A (qualitative insights)	Enables dynamic knowledge base updates for crisis management.	Scalable and adaptable to various crises.	Requires contextual prompt engineering for optimization.
[29]	MedFound LLM for clinical diagnosis	Accuracy: ~85%; Risk management: ~90%	Outperforms baseline LLMs in common and rare diseases.	Extensive evaluation across multiple scenarios.	High computational demands for large models.
[30]	Llama 2 for liver cirrhosis detection	Sensitivity: 100%; Specificity: 96%	High accuracy in detecting liver cirrhosis and related symptoms.	Efficient local deployment with low hardware needs.	Focused on specific conditions; generalizability unknown.
[31]	Multi Modal Transformers for radiology reports	BLEU-4: ~0.6; ROUGE-L: ~0.65	Effective narrative generation integrating text and images.	Leverages pre-trained encoders for efficiency.	Dataset scarcity limits broader application.
[32]	Manipulation vulnerability of LLMs	Accuracy retention: ~90% post Manipulation	Demonstrates susceptibility of LLMs to targeted misinformation.	Highlights need for robust verification mechanisms.	No mitigation strategies proposed.
[33]	PEGASUS for PET report generation	Clinical utility: 89% acceptance; Utility score: 4.08/5	Personalized impressions are clinically acceptable and time-saving.	Customizable to physician-specific styles.	Focused on PET reports only.
[34]	AI chatbot tips for medical educators	N/A (recommendation-based)	Enhances teaching and assessment with LLM-powered chatbots.	Practical guidelines for LLM implementation.	No experimental validation of tips.
[35]	LLMs for patient information leaflets (PILs)	Quality: ~3.58; Readability: ~14–17 years	PILs reduce workload but require clinician oversight.	Includes visuals for better comprehension.	Medical inaccuracies across all content.

[36]	OpenMedLM for open-source benchmarks	MedQA: ~72.6%; MMLU: ~81.7%	Achieves state-of-the-art performance on benchmarks.	Transparent and fine-tuning-free approach sets.	Lacks real-world deployment validation.
[37]	ETB MII for medical imaging interpretation	BLEU-4: ~0.61; CIDEr: SOTA	Competitive in medical report generation with low complexity.	Data augmentation enhances performance.	Limited evaluation beyond X-ray datasets.
[38]	Probability estimation in LLMs	Explicit vs. implicit probability: ~5% difference	Implicit probabilities outperform explicit ones.	Explores probability reasoning in LLMs.	Numerical reasoning limitations persist.
[39]	Med MLLM for multimodal COVID-19 analysis	Accuracy: ~92%; Adaptability: High	Adapts to rare disease scenarios with minimal labeled data.	Supports multilingual and multimodal inputs.	Retrospective testing may not reflect real-world robustness.
[40]	CAD for pododactyl radiography	BLEU-4: 0.612; ROUGE-L: 0.633	Automates pododactyl pathology reports with high quality.	Integrates CNN and Transformers effectively.	Dataset focus limits pathology diversity.

It has been noticed that LLMs show adaptability in dealing with various issues of health, education, or management in a crisis situation. For example, [30] and [33] show high accuracy and clinical utility, so there is a high potential to apply LLMs to specialized medical tasks. However, challenges such as bias magnification ([26]), vulnerability to misinformation ([32]), and a lack of diversity in the datasets ([31], [40]) underscore the need for sets that provide robustness and generalizability. Some of the future works include improvement of explainability for LLMs, reduction in inherent biases of LLM, and extension in capabilities to much broader applications. Such cooperation on benchmarking and open-source model building, as outlined in [36], could lead to better LLM-based solutions-solutions that are thus both fairer and more transparent.

4. Conclusion

The above body of work reviewed indicates that LLMs possess wide-ranging transformative capacities in domains such as medical diagnostics, report generation, and education. Most recently, a number of studies were processed that investigate different hybrid frameworks and models of LLMs, especially transformer architecture-based models, including GPT-3 and GPT-4, BERT-derived models, and other niche biomedical models specially designed for their applications, like BioGPT and Med-PaLM. Among all the models tested, GPT-3 and all the related models were those most applied because they

are very versatile and adaptable to a wide cross-section of domains of applications in clinical practice. Report generation was optimized using models like PEGASUS and ETB MII with a view toward producing highly accurate contextual outputs that could be clinically applicable. LLMs are important in reducing the workload of clinicians, improving patient care, and enriching educational and diagnostic workflows.

In fact, there was significant dependence on ensemble methods that comprised of both COMCARE and MedFound with multiple pre-trained models in solving tasks of high accuracy for complex reasoning. Those were stronger at some of the specific tasks such as NER, RE, and multimodal fusion that assist with clinical decision-making. This has been an example of task-specific fine-tuning as seen with Med-PaLM 2, and greatly benefited the question-answering scenario up to an 86.5% score on the MedQA dataset. DizzyInsight and ERG-AI demonstrated domain-specific knowledge that combines the functionalities of LLMs addressing niche health-related problems such as etiology classification for dizziness, all the way to the assessment of health risks related to ergonomics. However, high energy consumption and computational complexity are the two key challenges remaining that would demand more future development in efficiency and sustainability.

For the medical reports generation, the best possible solutions are identified between the vision-language models PEGASUS and ETB MII, and CAD-integrated frameworks like Llama-2, which tends to combine visual and textual data seamlessly, with the highest BLEU and ROUGE scores even in fully automated report generation tasks. With PEGASUS, the performance came along with personalized and clinically acceptable impressions in PET imaging. ETB MII tends to perform better in narrative radiology reporting tasks with computational requirements that are at the lower end. It also comprised the fact that state-of-the-art LLM capabilities are democratized for resource-constrained settings, balancing performance with transparency since open-source frameworks like OpenMedLM enable this.

This work is going to focus on improvement toward interpretability and trustworthiness of LLMs in counteracting ethical issues such as bias mitigation and vulnerabilities to misinformation. Further tailoring applications of these models into high-stakes domains would be completing strong evaluation frameworks and incorporating causal reasoning into LLMs. Hybrid models, such as that in the example of KELLM, combining an LLM with a knowledge graph, may open up avenues toward increasing interpretability without any loss in safety constraints. As advancement in prompt engineering, multimodal learning, and domain-specific fine-tuning continues to advance, the way medical reports will be generated will be revolutionized and so on. With the next generation of LLMs, aligned with ethical and practical considerations in terms of computational innovation, applications will be equitable, efficient, and impactful across these healthcare landscapes.

References:

[1] Gallifant, J., Afshar, M., Ameen, S. *et al.* The TRIPOD-LLM reporting guideline for studies using large language models. *Nat Med* **31**, 60–69 (2025). https://doi.org/10.1038/s41591-024-03425-5

[2] Sen, S., Gonzalez, V., Husom, E.J. *et al.* ERG-AI: enhancing occupational ergonomics with uncertainty-aware ML and LLM feedback. *Appl Intell* **54**, 12128–12155 (2024). https://doi.org/10.1007/s10489-024-05796-1

[3] Miah, M.S.U., Kabir, M.M., Sarwar, T.B. *et al.* A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and LLM. *Sci Rep* **14**, 9603 (2024). https://doi.org/10.1038/s41598-024-60210-7

[4] Sarker, I.H SETS. LLM potentiality and awareness: a position paper from the perspective of trustworthy and responsible AI modeling. *Discov Artif Intell* **4**, 40 (2024). https://doi.org/10.1007/s44163-024-00129-0

[5] Sarma, D., Rali, A.S. & Jentzer, J.C. Key Concepts in Machine Learning and Clinical Applications in the Cardiac Intensive Care Unit. *Curr Cardiol Rep* **27**, 30 (2025). https://doi.org/10.1007/s11886-024-02149-9

[6] Naved, B.A., Luo, Y. Contrasting rule and machine learning based digital self triage systems in the USA. *npj Digit. Med.* **7**, 381 (2024). https://doi.org/10.1038/s41746-024-01367-3

[7] Stretton, B., Kovoor, J., Arnold, M. *et al.* ChatGPT-Based Learning: Generative Artificial Intelligence in Medical Education. *Med.Sci.Educ.* **34**, 215–217 (2024). https://doi.org/10.1007/s40670-023-01934-5

[8] Chang, YH SETS., Lin, YC., Huang, FW. *et al.* Using machine learning and natural language processing in triage for prediction of clinical disposition in the emergency department. *BMC Emerg Med* **24**, 237 (2024). https://doi.org/10.1186/s12873-024-01152-1

[9] Hussain, S., Naseem, U., Ali, M. *et al.* TECRR: a benchmark dataset of radiological reports for BI-RADS classification with machine learning, deep learning, and large language model baselines. *BMC Med Inform Decis Mak* **24**, 310 (2024). https://doi.org/10.1186/s12911-024-02717-7

[10] Silvestri, C., Roshal, J., Shah, M. *et al.* Evaluation of a novel large language model (LLM)-powered chatbot for oral boards scenarios. *Global Surg Educ* **3**, 112 (2024). https://doi.org/10.1007/s44186-024-00303-z

[11] de Arriba-Pérez, F., García Méndez, S., Otero Mosquera, J. *et al.* Explainable cognitive decline detection in free dialogues with a Machine Learning approach based on pre-trained Large Language Models. *Appl Intell* **54**, 12613–12628 (2024). https://doi.org/10.1007/s10489-024-05808-0

[12] Zhang, X., Zhao, G., Ren, Y. *et al.* Data augmented large language models for medical record generation. *Appl Intell* **55**, 88 (2025). https://doi.org/10.1007/s10489-024-05934-9

[13] Singhal, K., Tu, T., Gottweis, J. *et al.* Toward expert-level medical question answering with large language models. *Nat Med* (2025). https://doi.org/10.1038/s41591-024-03423-7

[14] Xu, X., Jiang, R., Zheng, S. *et al.* Classification of Chronic Dizziness Using Large Language Models. *J Healthc Inform Res* (2024). https://doi.org/10.1007/s41666-024-00178-1

[15] Slack, D., Krishna, S., Lakkaraju, H SETS. *et al.* Explaining machine learning models with interactive natural language conversations using TalkToModel. *Nat Mach Intell* **5**, 873–883 (2023). https://doi.org/10.1038/s42256-023-00692-8

[16] Woźnicki, P., Laqua, C., Fiku, I. *et al.* Automatic structuring of radiology reports with on-premise open-source large language models. *Eur Radiol* (2024). https://doi.org/10.1007/s00330-024-11074-y

[17] Jin M, Choi S M, Kim G-W. COMCARE: A Collaborative Ensemble Framework for Context-Aware Medical Named Entity Recognition and Relation Extraction. *Electronics*. 2025; 14(2):328. https://doi.org/10.3390/electronics14020328

[18] Xu T, Li B. KELLM: Knowledge-Enhanced Label-Wise Large Language Model for Safe and Interpretable Drug Recommendation. *Electronics*. 2025; 14(1):154. https://doi.org/10.3390/electronics14010154

[19] Gopalakrishnan S, Garbayo L, Zadrozny W. Causality Extraction from Medical Text Using Large Language Models (LLMs). *Information*. 2025; 16(1):13. https://doi.org/10.3390/info16010013

[20] Delaunay J, Cusido J. Evaluating the Performance of Large Language Models in Predicting Diagnostics for Spanish Clinical Cases in Cardiology. *Applied Sciences*. 2025; 15(1):61. https://doi.org/10.3390/app15010061

[21] Y. Jeong, J. -J. Song, J. Yang and S. Kang, "Advancing Tinnitus Therapeutics: GPT-2 Driven Clustering Analysis of Cognitive Behavioral Therapy Sessions and Google T5-Based Predictive Modeling for THI Score Assessment," in IEEE Access, vol. 12, pp. 52414-52427, 2024, doi: 10.1109/ACCESS.2024.3383020.
keywords: {Medical treatment;Internet;Depression;Clustering algorithms;Transformers;Support vector machines;Principal component analysis;Social networking (online);Cognition;Behavioral sciences;Ear;Patient monitoring;Writing;Large language models;Augmentation;cognitive;CBT;GPT-2;Google;tinnitus;T5;RMSE;ROUGE-L},

[22] Borg, A., Jobs, B., Huss, V. *et al.* Enhancing clinical reasoning skills for medical students: a qualitative comparison of LLM-powered social robotic versus computer-based virtual patients within rheumatology. *Rheumatol Int* **44**, 3041–3051 (2024). https://doi.org/10.1007/s00296-024-05731-0

[23] Raminedi, S., Shridevi, S. & Won, D. Multi Modal transformer architecture for medical image analysis and automated report generation. *Sci Rep* **14**, 19281 (2024). https://doi.org/10.1038/s41598-024-69981-5

[24] Yang, Y., Liu, X., Jin, Q. *et al.* Unmasking and quantifying racial bias of large language models in medical report generation. *Commun Med* **4**, 176 (2024). https://doi.org/10.1038/s43856-024-00601-z

[25] Oh, Y., Park, S., Byun, H SETS.K. *et al.* LLM-driven multimodal target volume contouring in radiation oncology. *Nat Commun* **15**, 9186 (2024). https://doi.org/10.1038/s41467-024-53387-y

[26] Muñoz-Ortiz, A., Gómez-Rodríguez, C. & Vilares, D. Contrasting Linguistic Patterns in Human and LLM-Generated News Text. *Artif Intell Rev* **57**, 265 (2024). https://doi.org/10.1007/s10462-024-10903-2

[27] Wang, S., Zhao, Z., Ouyang, X. *et al.* Interactive computer-aided diagnosis on medical image using large language models. *Commun Eng* **3**, 133 (2024). https://doi.org/10.1038/s44172-024-00271-8

[28] Congès, A., Fertier, A., Salatgé, N. *et al.* R-IO SUITE: integration of LLM-based AI into a knowledge management and model-driven based platform dedicated to crisis management. *Softw Syst Model* (2024). https://doi.org/10.1007/s10270-024-01237-2

[29] Liu, X., Liu, H SETS., Yang, G. *et al.* A generalist medical language model for disease diagnosis assistance. *Nat Med* (2025). https://doi.org/10.1038/s41591-024-03416-6

[30] Wiest, I.C., Ferber, D., Zhu, J. *et al.* Privacy-preserving large language models for structured medical information retrieval. *npj Digit. Med.* **7**, 257 (2024). https://doi.org/10.1038/s41746-024-01233-2

[31] Leonardi, G., Portinale, L. & Santomauro, A. Enhancing radiology report generation through pre-trained language models. *Prog Artif Intell* (2024). https://doi.org/10.1007/s13748-024-00358-5

[32] Han, T., Nebelung, S., Khader, F. *et al.* Medical large language models are susceptible to targeted misinformation attacks. *npj Digit. Med.* **7**, 288 (2024). https://doi.org/10.1038/s41746-024-01282-7

[33] Tie, X., Shin, M., Pirasteh, A. *et al.* Personalized Impression Generation for PET Reports Using Large Language Models. *J Digit Imaging. Inform. med.* **37**, 471–488 (2024). https://doi.org/10.1007/s10278-024-00985-3

[34] Kıyak, Y.S. Beginner-Level Tips for Medical Educators: Guidance on Selection, Prompt Engineering, and the Use of Artificial Intelligence Chatbots. *Med.Sci.Educ.* **34**, 1571–1576 (2024). https://doi.org/10.1007/s40670-024-02146-1

[35] Pompili, D., Richa, Y., Collins, P. *et al.* Using artificial intelligence to generate medical literature for urology patients: a comparison of three different large language models. *World J Urol* **42**, 455 (2024). https://doi.org/10.1007/s00345-024-05146-3

[36] Maharjan, J., Garikipati, A., Singh, N.P. *et al.* OpenMedLM: prompt engineering can out-perform fine-tuning in medical question-answering with open-source large language models. *Sci Rep* **14**, 14156 (2024). https://doi.org/10.1038/s41598-024-64827-6

[37] Prieto-Ordaz, O., Ramirez-Alonso, G., Montes-y-Gomez, M. *et al.* Toward an enhanced automatic medical report generator based on large transformer models. *Neural Comput & Applic* **37**, 43–62 (2025). https://doi.org/10.1007/s00521-024-10382-0

[38] Gu, B., Desai, R.J., Lin, K.J. *et al.* Probabilistic medical predictions of large language models. *npj Digit. Med.* **7**, 367 (2024). https://doi.org/10.1038/s41746-024-01366-4

[39] Liu, F., Zhu, T., Wu, X. *et al.* A medical multimodal large language model for future pandemics. *npj Digit. Med.* **6**, 226 (2023). https://doi.org/10.1038/s41746-023-00952-2

[40] Vieira PdA, Mathew MJ, Santos Neto PdAd, Silva RRVe. The Automated Generation of Medical Reports from Polydactyly X-ray Images Using CNNs and Transformers. *Applied Sciences*. 2024; 14(15):6566. https://doi.org/10.3390/app14156566