Archives available at journals.mriindia.com

International Journal on Advanced Computer Engineering and Communication Technology

ISSN: 2278-5140

Volume 14 Issue 03s, 2025

An Iterative Comprehensive Evaluation of Large Language and Vision Models in Medical AI: Benchmarks, Adaptability, and Deployment Challenges

[1]Wani H. Bisen, [2]Avinash J. Agrawal
[1]Rashtrasant Tukdoji Maharaj Nagpur University,
Shri Ramdeobaba College of Engineering and Management, Nagpur, India
[2]Ramdeobaba University Nagpur, Shri Ramdeobaba College of Engineering and Management,
Nagpur, India.
Email: [1]wanib.26@gmail.com, [2]agrawalaj@rknec.edu
9075260902. 9422830245

Peer Review Information

Submission: 05 Nov 2025

Revision: 25 Nov 2025

Acceptance: 17 Dec 2025

Keywords

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

Abstract

PRISMA principles provide a thorough analysis of current advances in large language models (LLMs) and multimodal transformers for medical applications. As LLMs like GPT-4, BioGPT, Med-PaLM, and hybrid frameworks like COMCARE enter clinical processes, thorough synthesis is essential to increase performance, methodological adaptability, and implementation practicality in many healthcare situations. Their creativity in medical report writing, decision support, and diagnosis is notable, but the literature has not established a cohesive taxonomy that evaluates these models by uniform metrics, domain-specific generalizability, and ethical acceptability. Over 40 studies examined radiology report production, clinical question responding, cognitive assessment, and causal reasoning. After testing vision-language transformer architectures like PEGASUS and ETB MII for automated imaging-based reporting, graph-based reasoning was used to evaluate drug safety and interpretability of knowledge-integrated models like KELLM. As needed, BLEU, ROUGE, F1 score, CIDEr, and qualitative evaluations were used. Domain- adapted and hybrid models improve diagnostic accuracy, task- specific explainability, and clinician workload differently. Model illusion, biases, hostile manipulation, and resource-intensive fine- tuning persist. The report recommends strong benchmarking, public evaluation standards, and ethical frameworks for LLMs in high-stakes medical applications. This study defines LLMs' therapeutic utility and recommends infrastructure, ethics, and technology for safe and successful integration. This effort prepares scalable, interpretable, and equitable medical AI systems.

INTRODUCTION

AI speeds many solutions, including health care, making settings healthier. Big Language Models Like GPT-4 in process. BERT and Med-PaLM accurately generate human- like text. In subtle applications like clinical decision assistance and medical report production, such models have shown impressive performance, promising to reduce burden, enhance diagnosis accuracy, and improve patient care. Despite these advances,

healthcare LLMs confront ethical, scalability, and domain-specific optimization concerns. LLM exams usually evaluate technical performance or practical applications without much analysis. Such assessments only cover one model or use case and do not address methodological adjustments needed across healthcare fields.

This fragmented strategy prevents LLM integration with clinical operations by preventing full knowledge of their strengths and weaknesses. Without rigorous comparisons of models, datasets, and evaluation criteria, ideal solutions for radiology reporting, cognitive assessment, and diagnostic reasoning cannot be found. This thorough and analytical assessment of contemporary healthcare LLM studies fills these gaps. This study is identified and evaluated using the PRISMA flow diagram, analyzing methods, performance measurements, and real-world application. The article reviews BioGPT, COMCARE, and Vision Transformers models in progress, illustrating innovation's range. LLM deployment creates ethical and technical difficulties, authors say. Thus, this study would prepare healthcare for LLMs.

MOTIVATION AND CONTRIBUTION

The recent discovery that LLMs are altering healthcare but missing actionability in current studies of their potentials and challenges prompted this study process. Domain-specific adaptations, energy use, ethics, and scalability are ignored in previous research sets. Methodical and comprehensive LLM evaluation across health care applications is needed to address this gap. A rigorous methodology-based examination of health care LLM adaption and implementation studies is discussed for the process.

The study evaluates GPT-4, Med-PaLM, and PEGASUS models for performance, strengths, and weaknesses using PRISMA. We study vision-language transformers for report production, domain-specific adaptations like BioGPT for radiological operations, and hybrid models like COMCARE for named entity recognition and relation extraction sets. This publication will help academics and practitioners use LLMs to apply models from the earlier study sets. It addressed prejudice, energy utilization, and scalability while developing solutions. This will clarify LLMs in health care and enable their ethical and effective application in real life scenarios.

LITERATURE REVIEW

Fields that synthesize structured and unstructured data have been altered by large language models. LLM-based applications need a good taxonomy to standardize development, optimize deployment, and decrease bias, interpretability, and task-specific restrictions

A. LLMs in Healthcare Applications

LLMs in healthcare must be transparent, reproducible, and clinically applicable. TRIPOD-LLM [1] modularizes biomedical LLM research sets and includes detailed checklists. Structured reporting enhances clarity and accountability in medical applications including prognostic modeling and task- specific reporting. Modern pipelines like ERG-AI [2] promote ergonomics-related health using wearable sensor data and LLMs. This system bridges physical and cognitive realms using uncertainty-aware models, making it adaptable in LLMs and ideal for unconventional healthcare sets.

B. Multilingual and Multimodal Capability

LMs facilitate sentiment analysis linguistic gaps [3]. Multilingual context outperformed others in a transformer- LM ensemble model. This is relevant to LLMSeg [25], which adapted textual and visual clinical information for radiation therapy planning.

C. Ethical Considerations and Trustworthiness

It causes ethical and trust issues [4] and [24]. Evaluating bias, fairness, and accountability requires rigorous frameworks. LLMs inherit societal biases that impact healthcare cost and outcome forecasts [24].

D. Efficiency and Interactivity

Automating medical record production with LLMs has altered medical workflows [12]. Also possible with interactive medical education learning environments [22]. Domain- specific LLMs like Med-PaLM 2 [13] and DizzyInsight [14] outperform humans in clinical question answering and diagnostic classification sets. LLMs can explain complex ML models utilizing conversational interfaces and user-centric design, as shown by TalkToModel [15]. Thus, such changes are necessary to make LLM-driven technology more accessible and relevant in healthcare and beyond.

E. Structured Reporting and Explainability

On-premise LLM-automated radiological report structuring was shown in [16] for the process. These models demonstrated that LLMs can perform privacy-preserving task-specific tasks with near human accuracy sets. Combining structured and unstructured data [8, 17] increased diagnostic model explainability and accuracy, boosting clinical decision- making confidence sets.

F. Limitations in Training and Deployment

LLMs have no reality gaps in model hallucinations and discrepancies, notwithstanding their assurances. Similar to

domain-specific models like BioBERT, LLMs like GPT-4 lack reliability in causality extraction [19] sets. These findings again suggest hybrid frameworks that combine particular tools with LLMs for consistency and accuracy. Large language models (LLMs) have transformed automatic text output in several fields. LLMs are used for reporting, which is problematic because they combine structured and unstructure data samples. LLM growth in report generation, applicability, barriers, and process innovation are examined in process.

Table 1: Methodological Comparative Review Analysis

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 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 chronic dizziness classification.	High predictive value for PPPD and anxiety disorders.	Limited to specific 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.	Improves interactivity and authenticity.	Lacks physical examination simulations.
[23]	Encoder-decoder for medical image reporting	Combines Vision Transformer with GPT-2 for report generation.	Outperforms recurrent models in coherence and accuracy levels.	Focused on X-ray datasets.
[24]	Bias assessment in LLM predictions	Examines biases in healthcare cost and survival predictions.	Highlights need for bias mitigation.	Requires improved fairness strategies.
[25]	LLMSeg for radiotherapy planning	Multimodal AI for target volume delineations.	Robust generalization and data efficiency.	Limited validation beyond oncology.

G. Linguistic and structural analysis of LLM-generated text

Diagnostic imaging reports can be transformed by LLMs. CAD-LLM integration enhances chest X-ray diagnostic accuracy and quality [27]. [31] advances medical imaging similarly. They explain how multi-modal transformers with textual and visual input produce high-BLEU and ROUGE radiologist narrative reports.

Llama 2 [30] is a local LLM with high sensitivity and specificity for liver cirrhosis diagnosis. They protect patient data and eliminate hardware reliance, making them therapeutically relevant in many settings.

H. Approach for the Rare and Complex Diseases

LLMs show potential in rare illnesses with scant labeled data. Med MLLM [39] translates text and radiographs using multi- modal learning. Even

with a minimal labeled dataset, it works.
Pandemic response requires versatility due to rapid deployment and limited data samples.

I. Improvement in Crisis Management and Decision Support

LLMs support crisis management decisions outside of healthcare. R-IO SUITE [28] updates crisis knowledge bases with LLMs. Real-time context and static knowledge are integrated. LLMs adapt to high-stakes situations.

J. Personalized and Context-Aware Reporting

Customized, context-aware reports are another LLM frontier. Fine-tuning PEGASUS to generate personal PET report impressions worked [33]. The model may encode styles that provide clinical utility on physician-specific preference, allowing LLMs to report personally for the process.

K. Security and Ethical Consideration

Although full of potential, LLMs are loaded with security and LLMs pose security and ethical problems despite their potential. Simple attacks like [32]'s targeted alterations can propagate misleading biomedical data, creating concerns about patient-care decision-making output integrity sets.

L. Open-source innovations in Accessibility

Open-source LLMs make high-quality models accessible. OpenMedLM [36] performs state-of-the-art medical tasks without fine-tuning, demonstrating how open-source projects eliminate equity gaps in health AI applications.

Table 2: Methodological Comparative Review Analysis

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 in 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 summarization with expert validation.	High dependency on dataset quality and size.
[32]	Targeted manipulation of LLM weights	Demonstrated the 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 LLM-based tools in medical education.	Practical tips for optimizing LLM implementation.	Does not address technical limitations of LLMs.
[35]	LLMs for Patient Information Leaflets (PILs)	Evaluated readability and quality of PILs from three LLMs.	Reduced professional workload; high quality outputs.	Inaccuracies and high reading levels limit usability.
[36]	OpenMedLM for	Achieved state-of-the-	Transparent and	Limited to

	medical benchmarks	art results with open-source LLMs.	accessible performance improvements.	benchmark tasks, lacks real-world validation.
[37]	Enhanced transformer-based model for imaging	Generated reports with competitive BLEU and CIDEr scores.	Efficient reports for primary care imaging tasks.	High dependency on effective primary data augmentation.
[38]	Probability estimation under comparison in LLMs	Explicit probabilities underperformed compared to implicit ones.	Highlights need for improved clinical decision 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.	Limited to pododactyl imaging, needs generalizations.

M. Evaluation Metrics and Model Optimization

Quality and reliability of LLM reports must be assessed. Because medical reports penalize semantic inconsistencies and unclear phrases, evaluation metrics like CIDEr are needed. Comparative analysis clarifies LLM model probability generating and improves estimation. Pododactyl radiography reporting benefits from generative AI [40]. Diagnostics, secondary opinion to specialists, and clinical burden reduction are its benefits.

N. Education and Knowledge Transference

As illustrated in [34] and [35], medical education depends on it. LLMs make medical knowledge more accessible and customizable by creating patient information packets and interactive learning platforms.

Multiple LLM report generation developments have proved its Versatility In Process. Scalable, accurate, and context- specific medical diagnosis and crisis management models. However, security, bias reduction, and good evaluation frameworks are key challenges. LLMs in complex real-world applications will depend on improving such constraints. From complete automation to beyond, this dynamic taxonomy shows LLM as an essential report generating tool sets.

RESULTS VALIDATIONS AND DISCUSSION

A complete comparison of approaches, measures, and outcomes from the included studies would compare large language model usage and performance across domains. This PRISMA-based study classifies each study by method, performance, and impact on an iterative, comprehensive report taxonomy. Study gaps and scopes are revealed by strengths and shortcomings. Tables 3 and 4 highlight LLM usage in medical diagnostics, learning reports, and ethics. Accuracy and efficiency improve when persistent dizziness is categorized to provide a radiological structural image [14& 16]. Bias, data dependence, and domain generalizability are ongoing challenges ([24], [12], [9], [25]).

Table 3: Methodological Statistical Review Analysis

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%	Combines posture predictions and user-friendly health risk reports.	Integrates uncertainty-aware ML and LLMs for personalized output.	Energy consumption and carbon footprint analysis needed.
[3]	Ensemble LLM	Sentiment	High	Robust	Limited to

	for sentiment analysis	accuracy: 86%	performance in multilingual sentiment tasks.	ensemble design.	four languages; scalability unknown.
[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.
[9]	BI-RADS classification via NLP	Sensitivity: 60%; Specificity: ~85%	BioGPT outperforms other models for breast imaging report classification.	Provides annotated dataset for BI-RADS classification.	Limited generalizability beyond BI-RADS categories.
[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.
[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 recommendation	Accuracy: ~88%; Safety metric: ~85%	Trustworthy recommendations with causal insights.	Integrates knowledge graphs for interpretability.	Limited evaluation on diverse EHR datasets.

Future research should develop reliable, transparent, and scalable LLM systems that overcome these limitations and work well in healthcare and education. The following table examines large language model (LLM)-based techniques for numerous applications using PRISMA. Methods, metrics, significant results, and process strengths and weaknesses vary. An examination of LLM research steps, problems, and future objectives is presented. Where articles did not provide findings, approximation measures were utilized for comprehensive analysis.

Table 4: Methodological Statistical Review 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.

[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]	MultiModal 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.
[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.
[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.	Lacks real-world deployment validation.
[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.

LLMs handle health, education, and management crises successfully. Due to their accuracy and clinical utility, LLMs can be utilized for specialized medical tasks [30] and [33]. Robust and generalizable sets are essential due to bias magnification ([26]), misinformation vulnerability ([32]), and dataset diversity concerns ([31], [40]) for different scenarios. Future LLM work involves explainability, bias reduction, and application expansions. Benchmarking and open-source model design could make LLM-based solutions fairer and more transparent [36] for the process.

CONCLUSION

Previous research showed LLMs can change medical diagnoses, report writing, and teaching. BERT-derived models, transformer architecture-based models like GPT-3 and GPT-4, and specialist biomedical models like BioGPT and Med-PaLM were studied in recent LLM hybrid framework investigations.

The most often used models were GPT-3 and similar models because they are versatile and adaptable to many clinical applications. PEGASUS and ETB MII optimized report production delivered correct contextual outputs for therapeutic application. LLMs provide better patient care, minimize physician workload, and improve education and diagnosis.

Ensemble techniques like COMCARE and MedFound with numerous pre-trained models were essential for accurate reasoning on complex tasks. These excelled in clinical decision-making tasks like NER, RE, and multimodal fusion. Like Med-PaLM 2, task-specific fine-tuning boosted MedQA question-answering to 86.5%. DizzyInsight and ERG-AI used LLM functionalities to address dizzy etiology categorization and ergonomics health risk assessment. Energy consumption and computing complexity remain important issues that require efficiency and sustainability improvements.

Vision-language models PEGASUS and ETB MII and CAD- integrated frameworks like Llama-2 seamlessly incorporate visual and textual data in fully automated report generating tasks and

have the greatest BLEU and ROUGE scores for medical report generation. PEGASUS produced customized, clinically acceptable PET imaging. ETB MII excels in narrative radiology reporting with modest computational requirements. Open-source frameworks like OpenMedLM democratize state-of-the-art LLM capabilities for resource- constrained environments, balancing performance and transparency in process.

This effort will improve LLM interpretation and credibility to combat Ethical difficulties include prejudice mitigation and misinformation vulnerability. Strong LLM evaluation procedures and causal reasoning would further customize these models to high-stakes contexts. KELLM, a knowledge graph-LLM hybrid, may increase interpretability without compromising safety. Quick engineering, multimodal learning, and domain-specific fine-tuning will change medical reports. Next-generation LLMs with ethical and practical computational innovation will provide equitable, efficient, and effective healthcare applications.

ACKNOWLEDGMENT

The authors would like to acknowledge the support of Shri Ramdeobaba College of Engineering and Management, Nagpur for providing the necessary facilities. Appreciation is extended to colleagues and peers for their helpful discussions and suggestions.

REFERENCES

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

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/s10489-024-05808-0

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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