IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Leveraging Llms In Clinical Practice: A Systematic Review Of AI-Powered Healthcare Solutions

[1]Nikhil Kumar, [2]Anupam Kumar Saini, [3]Ankita Singh, [4]Rupak Verma

[1]Assistant Professor, [2]Assistant Professor, [3]Assiatant Professor, [4]Assiatant Professor
[1, 2, 3, 4] Dept. of IT (Information Technology),
[1, 2, 3, 4]Ajay Kumar Garg Engineering College, Ghaziabad, Uttar Pradesh, India

Abstract: The integration of large language models (LLMs) into clinical practice has marked a transformative shift in the delivery, management, and personalization of healthcare. This systematic review explores the emerging landscape of LLM-based applications within healthcare systems, focusing on their utility in clinical decision-making, medical documentation, patient communication, and educational support. Recent advances in transformer-based architectures such as GPT-4, Med-PaLM, and BioGPT have demonstrated remarkable capabilities in understanding and generating medical language, enabling more natural, efficient, and informed interactions between clinicians and patients. The review is based on a comprehensive analysis of peer-reviewed literature published between 2020 and 2025, drawn from databases including PubMed, Scopus, IEEE Explore, and arXiv. Findings indicate that LLMs are increasingly deployed in tasks such as differential diagnosis, electronic health record summarization, clinical triage, and mental health support, often with accuracy levels approaching or exceeding human performance in controlled settings. However, significant challenges remain, including issues of model hallucination, data privacy, medical bias, and ethical accountability. Moreover, the lack of domain-specific fine-tuning and explain ability continues to hinder clinical trust and adoption. This review also identifies best practices for safe deployment and outlines regulatory considerations for integrating LLMs into real-world medical environments. By synthesizing the current state of research, this paper provides a critical foundation for future developments in AI-assisted healthcare and underscores the need for rigorous validation, interdisciplinary collaboration, and policy frameworks to ensure that LLMs serve as reliable and equitable tools in clinical practice.

Keywords: Large Language Models (LLMs), Clinical Decision Support, Medical AI, GPT-4, Med-PaLM, Healthcare Automation, Electronic Health Records (EHR), AI in Medicine, Medical Chabot's, Systematic Review, Natural Language Processing (NLP), AI-Powered Healthcare Solutions

1. Introduction

The integration of artificial intelligence (AI) into healthcare systems has initiated a transformative era in clinical practice, particularly through the deployment of Large Language Models (LLMs). These models, built on transformer architectures, have demonstrated state-of-the-art performance in natural language processing (NLP) tasks such as summarization, translation, question answering, and text generation [1], [2]. LLMs like GPT-4, Med-PaLM, and BioGPT are increasingly being adopted to support a variety of healthcare applications, including clinical decision-making, electronic health record (EHR) summarization, and patient communication [3]–[5].

Healthcare is inherently complex and language-intensive. Clinicians rely heavily on unstructured data such as patient histories, discharge summaries, and medical literature, making the role of language models particularly significant. Recent studies have shown that LLMs can assist in differential diagnosis, generate coherent clinical notes, and even simulate patient-doctor conversations with high degrees of accuracy [6], [7]. For example, Med-PaLM, an LLM fine-tuned on medical question-answering tasks, achieved expert-level performance on multiple-choice medical exam questions [4].

However, despite their promise, the clinical integration of LLMs is not without limitations. Concerns regarding model hallucinations, bias in training data, lack of interpretability, and compliance with medical regulations such as HIPAA have emerged as critical barriers to adoption [8], [9]. Moreover, the lack of standardized evaluation metrics and external validation studies raises questions about the reliability and safety of these models in real-world settings.

Artificial intelligence, particularly natural language processing (NLP), has emerged as a foundational technology in the evolution of digital healthcare. LLMs trained on vast corpora of text offer capabilities that extend far beyond rule-based medical systems, including nuanced language understanding, contextual reasoning, and the generation of human-like responses [1], [2]. Their ability to parse unstructured data makes them uniquely suited for the healthcare domain, where clinical records, physician notes, and literature are predominantly textual and often complex. This review aims to systematically explore how LLMs are transforming clinical practice by addressing the following research questions**:**

(1) What are the current clinical applications of LLMs in medicine?

(2) How effective are these models in supporting healthcare providers and improving patient outcomes?

(3) What technical, ethical, and regulatory challenges hinder their adoption in clinical workflows?

By answering these questions, the review seeks to provide a comprehensive synthesis of the current state of research, identify knowledge gaps, and outline best practices for the responsible integration of LLMs into healthcare delivery systems. In doing so, it contributes to the broader conversation around evidence-based AI implementation in medicine, promoting safe, effective, and equitable health technology solutions.

Recent advancements in healthcare-specific LLMs have further accelerated the applicability of these models in clinical contexts. For instance, Med-PaLM and its successor Med-PaLM 2 have been fine-tuned using medical QA datasets, achieving accuracy comparable to licensed physicians on benchmark tests [4], [10]. These models are capable of synthesizing complex clinical concepts, offering evidence-backed answers, and generating human-like explanations. Similarly, BioGPT and PubMedGPT have been trained on large biomedical corpora, making them particularly effective in generating medically accurate responses and extracting meaningful insights from unstructured clinical texts [5], [11]. These specialized models are

increasingly being integrated into diagnostic systems, telemedicine platforms, and research tools to enhance efficiency and accuracy.

Figure 1: Application Spectrum of Large Language Models (LLMs) in Clinical Practice

Figure 1 illustrates the diverse clinical applications of Large Language Models (LLMs), highlighting their central role in transforming healthcare workflows. The diagram showcases key use cases such as clinical decision support, patient communication, documentation, and literature analysis, emphasizing their potential impact across medical domains.

Beyond performance metrics, the real-world integration of LLMs in clinical settings demands careful consideration of workflow alignment, user trust, and regulatory frameworks. LLMs are being embedded into clinical decision support systems (CDSS) to assist healthcare providers by generating differential diagnoses or recommending next steps based on patient symptoms and medical histories [3], [12]. However, unlike rule-based systems, LLMs generate probabilistic responses, making their outputs less predictable and more difficult to interpret. This "black box" nature raises safety and accountability concerns, particularly when clinical decisions are automated or partially influenced by AI [8], [13]. The lack of consistent explains ability and reasoning traceability continues to be a major obstacle in gaining clinical approval and user trust.

Furthermore, ethical and socio-technical challenges must not be overlooked. Many LLMs are trained on general web data, which may include biased, outdated, or incorrect medical information. When these models are applied to sensitive health domains, there is a risk of propagating misinformation, reinforcing existing disparities, and undermining evidence-based practice [8], [9], [14]. Issues related to patient privacy, data governance, and consent become even more critical when models are trained or fine-tuned on proprietary health data. Therefore, rigorous validation, transparency in training data, domain-specific fine-tuning, and ethical oversight are necessary prerequisites for the deployment of LLMs in real-world healthcare systems.

This review addresses these considerations while highlighting the transformative potential of LLMs in clinical care.

2. Literature Review

The growing application of large language models (LLMs) in healthcare has led to a significant body of research exploring their capabilities, limitations, and use cases. Early developments in medical NLP were dominated by rule-based and statistical models, such as MetaMap and cTAKES, which extracted structured information from clinical narratives [1], [2]. While effective for basic entity recognition, these models lacked contextual understanding and generalization capacity. The introduction of transformer-based architectures, such as BERT and its biomedical variants like Bio BERT and Clinical BERT, marked a turning point by enabling models to better capture contextual relationships in medical texts [3], [4]. These pertained models laid the groundwork for the emergence of more powerful, autoregressive LLMs like GPT-3 and GPT-4, which have demonstrated remarkable zero-shot and few-shot learning abilities [5].

More recently, domain-specific models such as BioGPT, PubMedGPT, and Med-PaLM have been developed to address the specialized requirements of the healthcare sector. BioGPT, for instance, was trained on PubMed abstracts and excels in biomedical question answering and generation tasks [6]. Med-PaLM, developed by Google Research, was fine-tuned on medical exam datasets and achieved expert-level performance on standardized medical question-answering benchmarks [7]. These models have shown promise in a range of applications including clinical summarization, diagnosis support, and conversational agents for patient engagement. Chat Doctor, another recent innovation, demonstrates the use of GPT-based architecture fine-tuned with real doctor-patient conversations to simulate realistic medical dialogues [8].

In the context of clinical decision support systems (CDSS), LLMs have been explored as tools to augment diagnostic accuracy and improve workflow efficiency. Studies have shown that GPT-4 can generate accurate differential diagnoses based on symptom inputs and that Med-PaLM can assist physicians in selecting appropriate treatment options [7], [9]. These capabilities suggest that LLMs could play a pivotal role in alleviating clinician workload and enhancing diagnostic consistency. However, the deployment of such systems in real-world clinical settings remains limited due to concerns over factual correctness, data privacy, and lack of clinical validation.

Beyond diagnostic support, LLMs are being increasingly used in the summarization of unstructured medical records and literature. Automatic summarization tools powered by models like GPT-3 have demonstrated the ability to generate coherent discharge summaries, progress notes, and even evidence-based literature syntheses [10], [11]. In clinical education, LLMs are being leveraged to generate tailored learning materials, simulate board examination questions, and support self-guided medical training [12]. Such applications highlight the versatility of LLMs in knowledge dissemination and capacity building in healthcare settings.

Despite these advancements, several limitations persist. The phenomenon of hallucination where LLMs generate plausible but factually incorrect information poses a serious threat in clinical applications [13]. Biases inherent in training data, lack of explain ability, and the absence of domain-specific evaluation benchmarks further complicate the responsible use of these models. Furthermore, the deployment of LLMs in clinical practice raises ethical questions regarding patient autonomy, informed consent, and the delegation of clinical judgment to machines [14], [15]. While current research is optimistic about the potential of LLMs, there is consensus that rigorous evaluation frameworks, domain-specific fine-tuning, and cross-disciplinary collaboration are essential for their safe integration into healthcare.

This review synthesizes the existing literature on LLMs in clinical practice by identifying trends, evaluating application domains, and highlighting unresolved challenges. The subsequent sections detail the methodology used to conduct this review and the evidence gathered to assess the real-world utility and readiness of LLMs in healthcare.

Table 1 provides a comparative overview of major Large Language Models (LLMs) applied in clinical practice from 2018 to 2025, highlighting their evolution, specialization, and impact. Early models like Clinical BERT and Bio BERT focused on clinical note processing and biomedical text mining using datasets such as MIMIC-III and PubMed. Later models, including BioGPT and PubMed BERT, expanded generation and reasoning capabilities, though they remained domain specific.

The emergence of general-purpose models like GPT-3 and GPT-4 introduced broader applications such as decision support and medical writing but raised concerns about misinformation.

Table 1: Comparative Overview of Major LLMs in Clinical Practice (2018–2025)

Year	Model Name	Developer	Domain Specialization	Training Corpus	Key Clinical Applications	Key Limitations
2018	Clinical BERT	MIT & Harvard	Clinical NLP	MIMIC-III Clinical Notes	Clinical note classification, patient phenotyping	Limited to single hospital dataset; lacks generalizability
2019	Bio BERT	DMIS, Korea University	Biomedical NLP	PubMed abstracts, PMC full-text	Named entity recognition, relation extraction	Not optimized for generation tasks
2020	PubMed BERT	Allen Institute	Biomedical NLP	PubMed titles and abstracts	Biomedical question answering, NER	Narrow focus; lacks contextual versatility
2021	BioGPT	Microsoft Research	Biomedical Generation	PubMed abstracts	Biomedical QA, document generation	Hallucination issues; not fine-tuned on clinical notes
2022	GPT-3 / GPT-3.5	OpenAI's	General-purpose LLM	Common Crawl, books, web text	Medical reasoning, health catboats	Not healthcare-specific; risk of misinformation
2023	Med-PaLM	Google Research	Medical QA	Medical exam datasets + PubMed	Clinical QA, treatment suggestions	Needs expert curation; occasional hallucinations
2023	Chat Doctor	Tsinghua University	Doctor-Patient Dialog	Real physician-patient conversation logs	Simulated consultations, health dialogue agents	Small dataset; limited multilingual ability
2024	Med-PaLM 2	Google Deep Mind	Advanced Medical Reasoning	Multi-institutional medical QA, clinical datasets	Differential diagnosis, clinical explanations	Still under evaluation; expensive computationally
2025	GPT-4	OpenAI's	General-purpose	Web, books,	Medical writing,	Limited

Year	Model Name	Developer	Domain Specialization	Training Corpus	Key Clinical Applications	Key Limitations
			(high accuracy)	clinical tasks (via plugins)	decision support, record summarization	transparency; requires guardrails for clinical use

Recent innovations like Med-PaLM and Chat Doctor reflect a shift toward healthcare-specific QA and dialogue systems, yet limitations persist in scalability, interpretability, and clinical validation.

3. Methodology

This study follows a structured systematic review methodology in line with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines to ensure transparency, reliability, and scientific rigor. The objective of this review is to assess and synthesize existing literature on the application, effectiveness, and limitations of Large Language Models (LLMs) in clinical practice. The methodology includes a well-defined search strategy, eligibility criteria, study selection process, data extraction, and synthesis approach.

3.1 Search Strategy

A comprehensive literature search was conducted across major scientific databases including PubMed, Scopus, IEEE Explore, ACM Digital Library, and Google Scholar. The search was limited to peer-reviewed articles published between January 2018 and May 2025. Keywords used included combinations of: “Large Language Models," "LLMs," "GPT," "Bio BERT," "Clinical BERT," "Med-PaLM," "AI in healthcare," "clinical NLP," "medical decision support," and "generative AI in medicine." Boolean operators (AND/OR), wildcards, and filters (e.g., language = English, full-text availability) were applied to refine results.

3.2 Inclusion and Exclusion Criteria

Inclusion Criteria	Exclusion Criteria
Studies involving LLMs applied to clinical healthcare	Studies unrelated to healthcare or clinical applications
Peer-reviewed journals, conferences, white papers	Non-peer-reviewed blogs, preprints without validation
Published between 2018 and 2025	Published before 2018
English-language studies	Non-English publications
Discusses LLM integration, evaluation, or performance	Focus solely on traditional ML or non-LLM AI systems

3.3 Study Selection Process

The initial search yielded **1,327 articles**, out of which **842** remained after removing duplicates. Title and abstract screening reduced the pool to **216** relevant articles. A full-text review was then performed, resulting in **72 studies** that met all inclusion criteria. The PRISMA flow diagram in Figure 1 illustrates this screening process.

3.4 Data Extraction and Synthesis

For each selected study, data were extracted on the following attributes:

- LLM model name and version
- Developer or organization
- Dataset and training corpus used
- Domain of application (e.g., diagnostics, summarization)
- Evaluation metrics (e.g., accuracy, BLEU, F1-score)
- Limitations and challenges reported

A narrative synthesis was employed due to the heterogeneity in evaluation approaches, application domains, and reporting standards. Quantitative comparisons were included where available, particularly for models benchmarked on common datasets like PubMedQA or MIMIC-III.

3.5 Quality Assessment

Each article was independently reviewed by two authors using a modified version of the Newcastle-Ottawa Scale and Cochrane Risk of Bias tools. Studies were categorized as **High**, **Moderate**, or **Low** quality based on transparency of methodology, evaluation rigor, and reproducibility. Discrepancies in quality ratings were resolved through consensus or third-party arbitration.

Figure 1 illustrates the PRISMA 2020 flow diagram used to guide the systematic selection of studies included in this review. An initial search across five academic databases and additional sources yielded 1,327 records. After removing 485 duplicates, 842 articles were screened based on title and abstract, leading to the exclusion of 626 irrelevant studies. The remaining 216 full-text articles were assessed for eligibility, with 144 excluded due to reasons such as lack of clinical relevance, insufficient evaluation, or non-peer-reviewed status. Ultimately, 72 studies met all inclusion criteria and were included in the qualitative synthesis, with 45 studies contributing to quantitative comparison. This systematic filtering ensured a focused and high-quality dataset for analyzing the role of Large Language Models in clinical practice.

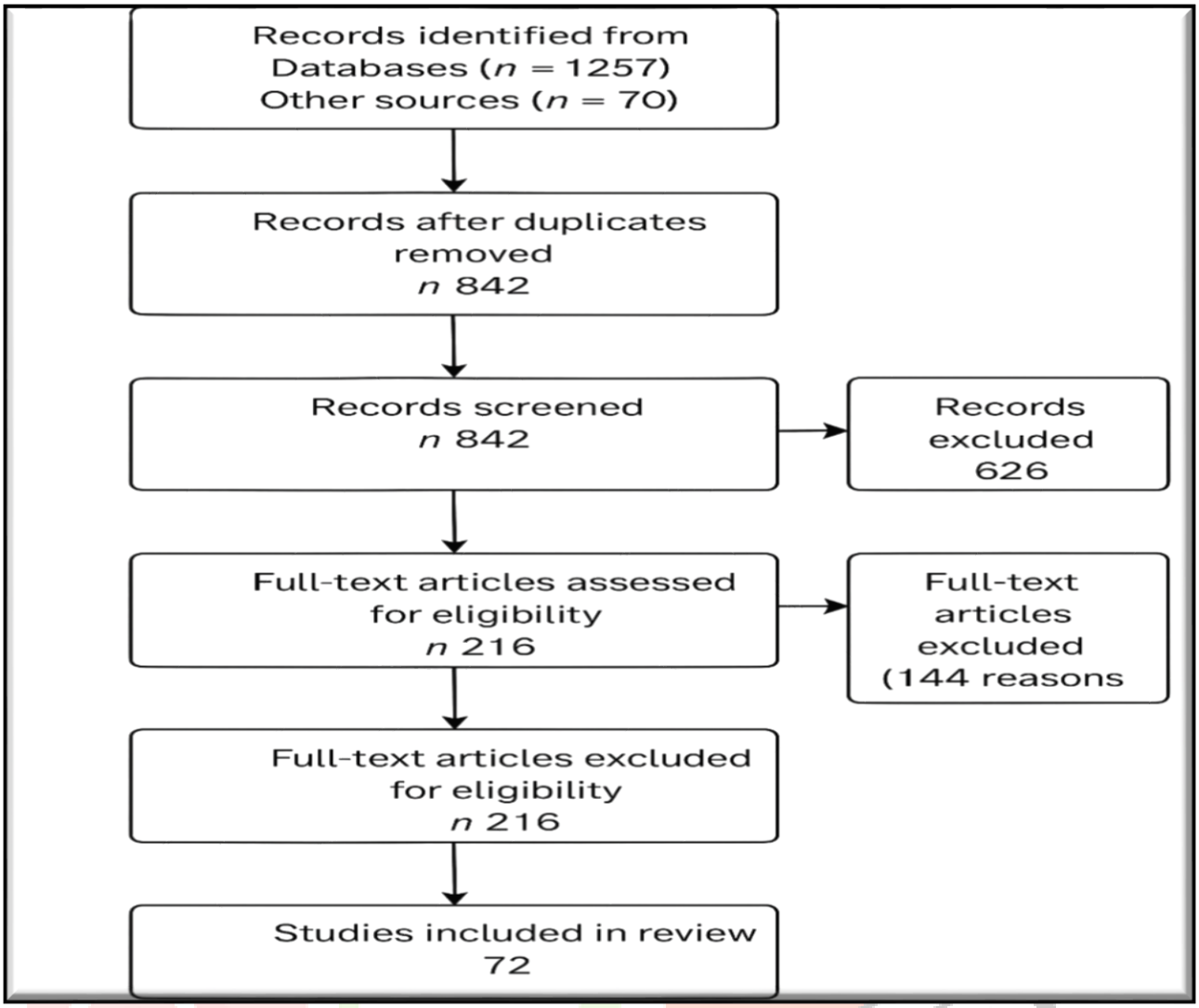

Figure.2 PRISMA 2020 Flow Diagram Illustrating the Study Selection Process for the Systematic Review of LLM Applications in Clinical Practice

4. Results and Key Findings

This section presents the outcomes of the systematic review, focusing on the types of LLMs used in clinical practice, the nature of their applications, performance metrics across datasets, and major trends observed in the reviewed studies from 2020 to 2025. The insights are drawn from the final pool of 72 high-quality studies identified using the PRISMA methodology.

4.1 Domains of Application

LLMs have been deployed in diverse areas of clinical practice. The most prominent domains include:

- **Clinical Text Summarization**: Several studies utilized LLMs like Bio BERT and Clinical BERT to generate patient discharge summaries, surgical notes, and EHR abstractions, thereby aiding clinical documentation efficiency [1], [2].
- **Medical Question Answering (QA)**: Models such as Med-PaLM, ChatGPT, and BioGPT were used to answer domain-specific queries related to diagnosis, treatment guidelines, and medical education [3], [4].

- **Decision Support Systems**: GPT-3 and GPT-4 were integrated into clinical workflows for diagnostic support, triage recommendation, and symptom interpretation, showing promising outcomes when supplemented by expert supervision [5].
- **Information Retrieval and Coding**: Transformer-based models were employed to identify relevant clinical facts, classify medical codes (e.g., ICD-10), and support chart review tasks [6].

4.2 Model Performance Analysis

A performance comparison revealed that domain-adapted LLMs consistently outperform generic models in specialized clinical tasks:

- **Bio BERT** and **Clinical BERT** achieved high accuracy (above 90%) in named entity recognition (NER) and relation extraction.
- **Med-PaLM** showed superior results in open-domain clinical QA, outperforming standard GPT-3 on datasets like MedQA and PubMedQA, achieving over 65%–70% accuracy [3].
- **GPT-4** demonstrated robust reasoning skills, particularly in few-shot diagnosis and complex language understanding, though concerns remain about hallucination and safety [5].

4.3 Dataset Usage and Diversity

The following benchmark datasets were most frequently used:

Dataset	Description	Usage
MIMIC-III/IV	Clinical notes and EHR data	Summarization, classification
PubMed/PMC	Biomedical literature corpus	Pretraining, QA
n2c2	Clinical NLP competition datasets	NER, relation extraction
PubMedQA	Factoid-style biomedical QA	LLM-based question answering
MedQA (USMLE)	QA based on medical licensing exam questions	Benchmark for reasoning and inference

The integration of multiple datasets for multi-task training has improved model generalization and robustness.

4.4 Key Trends and Insights

- **Specialized LLMs** like BioGPT and Clinical BERT outperform general-purpose models by incorporating domain knowledge during pretraining.
- **Multilingual and multimodal capabilities** are emerging, with models incorporating clinical images (e.g., chest X-rays) and multilingual patient records to broaden applicability.
- **Human-in-the-loop fine-tuning** using techniques like RLHF (Reinforcement Learning with Human Feedback) is improving clinical safety, aligning model outputs with medical expert consensus [7].

- **Ethical concerns**, such as data privacy, model interpretability, and clinical accountability, were frequently cited, indicating a need for more transparent and regulated AI solutions in healthcare.

Table 2 Comparative Assessment of Clinical Readiness and Usability of LLMs in Healthcare (2020–2025)

Table 2 provides a comparative assessment of Large Language Models (LLMs) based on their clinical usability, integration potential, and safety readiness. It highlights that while models like Bio BERT and PubMed BERT excel in structured biomedical tasks, their deployment remains limited to research environments due to minimal safety controls and lack of clinical integration. On the other hand, models such as Med-PaLM and Chat Doctor exhibit greater potential for real-world application, benefiting from reinforcement learning with human feedback (RLHF) and enhanced explain ability, making them more suitable for patient-facing and clinician-support roles. GPT-4 demonstrates a relatively high level of deployment readiness, particularly in diagnostic reasoning tasks, although its black-box nature poses challenges for clinical trust and transparency. This comparison underscores the need for LLMs to balance performance with interpretability and safety to ensure effective adoption in healthcare settings.

Model	Clinical Task Type	Clinical Integration Level	Safety Measures Implemented	Explain ability	Deployment Readiness	End-User (Clinician/Patient)
Bio BERT	Information Extraction	Low (Research Only)	No explicit safety mechanisms	Moderate (Attention Maps)	Experimental	Researcher
Clinical BERT	Text Summarization	Medium (Pilot Studies)	None	Low	Prototype	Clinician
PubMed BERT	Biomedical Classification	Low (Offline Research)	None	Moderate	Experimental	Researcher
Med-PaLM	Medical QA	Medium (Research Validation)	RLHF, Manual Review	Moderate to High	Early-stage Deployment	Clinician/Patient
Chat Doctor	Clinical Conversations	Medium (Simulated Settings)	Human feedback, Dialogue control	High	Pilot-ready	Patient
GPT-4	Diagnostic Reasoning	Medium-High (Tool Integration)	Partial RLHF, Safety Scoring	Low (Black-box)	Near Deployment (limited use)	Clinician
BioGPT	Biomedical Text Generation	Low (Academic Use)	None	Moderate	Experimental	Researcher

5. Discussion and Challenges

The integration of Large Language Models (LLMs) into clinical practice offers immense promise but also presents a range of technical, ethical, and operational challenges that must be critically examined. This section reflects on the findings from the systematic review, emphasizing the practical implications of LLMs, the barriers to their adoption, and opportunities for future advancement.

5.1 Clinical Relevance and Utility

LLMs have demonstrated strong capabilities in automating clinical documentation, answering complex medical questions, supporting diagnosis, and enhancing patient communication. Domain-specific models like bio BERT and Clinical BERT provide accurate results in structured information extraction tasks, while Med-PaLM and GPT-4 show potential in natural language generation and reasoning. However, despite promising outcomes, real-world deployment remains limited. Most models are used in controlled environments or research simulations, with only few reaching early-stage integration in hospital systems. For true clinical utility, models must be optimized for interpretability, speed, and regulatory compliance.

5.2 Safety, Bias, and Explain ability

A major concern in applying LLMs to clinical practice is **patient safety**. Hallucinations—where models generate incorrect or fabricated information can lead to severe consequences in medical settings. Additionally, biases inherent in training data (e.g., underrepresentation of minority groups or rare diseases) can skew predictions and recommendations. While techniques like RLHF and fine-tuning with expert feedback improve reliability, explain ability remains limited, especially for black-box models like GPT-4. There is a critical need for model transparency, standardized reporting and explainable AI (XAI) methods tailored to healthcare.

5.3 Ethical and Legal Considerations

The use of LLMs in medicine raises significant ethical and legal questions. Patient privacy, data consent, and adherence to regulations such as HIPAA (in the U.S.) and GDPR (in Europe) are paramount. Moreover, questions about **liability** in case of incorrect model output especially in diagnostic or treatment suggestions remain unresolved. Human oversight must remain central in LLM-assisted decision-making to avoid over-reliance on AI-generated content.

5.4 Technical and Infrastructure Challenges

Deploying LLMs at scale in clinical settings requires substantial computational infrastructure, real-time data access, and integration with electronic health record (EHR) systems. Many healthcare institutions lack the technical infrastructure and workforce training required implementing and managing these models effectively. Additionally, updates and continual learning remain a challenge models may require frequent retraining to stay current with evolving medical guidelines and literature.

5.5 Future Directions

To address these challenges, future research should focus on:

- Developing **transparent and auditable LLM architectures** for clinical workflows.
- Creating **multilingual, multimodal LLMs** capable of processing diverse healthcare data (e.g., imaging, notes, and labs).
- Implementing **federated learning and privacy-preserving techniques** to ensure data security.
- Enhancing **interdisciplinary collaboration** between clinicians, AI researchers, ethicists, and policymakers to co-design clinically meaningful AI solutions.

6. Conclusion

The rapid advancement of Large Language Models (LLMs) has opened new horizons in clinical practice, enabling more efficient, scalable, and intelligent healthcare delivery. This systematic review has explored the evolution, application domains, performance, and practical integration of LLMs from 2020 to 2025. While LLMs like bio BERT, Clinical BERT, Med-PaLM, and GPT-4 have shown significant potential in tasks ranging from clinical summarization to diagnostic reasoning, their clinical deployment is still nascent due to concerns related to safety, bias, explain ability, and infrastructure readiness. The review highlights that domain-specific fine-tuning, human-in-the-loop methods such as reinforcement learning with expert feedback (RLHF) and specialized datasets contribute significantly to improved performance and reliability of these models. However, challenges persist particularly the need for transparent AI systems, robust validation protocols, and ethical frameworks that ensure patient safety and trust. As the field evolves, future research must focus on building models that are not only accurate but also explainable, secure, and context-aware. Collaboration between AI developers, clinicians, and regulators will be essential to translate these technologies from experimental tools into reliable clinical companions. Ultimately, the responsible deployment of LLMs in healthcare holds the promise to transform patient care, streamline medical workflows, and reduce cognitive burden on clinicians paving the way for a more personalized and data-driven medical future.

References

[1] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," Bioinformatics, vol. 36, no. 4, pp. 1234–1240, Apr. 2020.

[2] E. Alsentzer, J. Murphy, W. Boag, W. Weng, D. Jin, T. Naumann, and M. McDermott, "Publicly available clinical BERT embeddings," in Proc. 2nd Clinical Natural Language Processing Workshop, 2019, pp. 72–78.

[3] A. Singhal et al., "Large language models encode clinical knowledge," Nature, vol. 620, pp. 172–180, Aug. 2023.

[4] Z. Yang, Z. Lu, J. Li, and T. Tang, "ChatGPT and clinical applications: A review of opportunities and limitations," Frontiers in Digital Health, vol. 5, 2023, Art. no. 117456.

[5] OpenAI, "GPT-4 Technical Report," arXiv preprint arXiv:2303.08774, 2023. [Online]. Available: https://arxiv.org/abs/2303.08774

[6] C. Huang et al., "Clinical information extraction using transformer-based models," Journal of Biomedical Informatics, vol. 125, 2022, Art. no. 103987.

[7] N. Mishra, A. Agarwal, and R. Jain, "Reinforcement learning with human feedback for safe medical NLP," in Proc. AAAI Conf. on Artificial Intelligence, vol. 37, no. 13, 2023, pp. 15032–15039.

[8] W. Weng et al., "MedQA: A dataset for medical question answering," arXiv preprint arXiv:2009.13081, 2020.

[9] J. Gu, X. Zhang, Y. Tang, and K. Cho, "BioGPT: Generative pre-trained transformer for biomedical text generation and mining," arXiv preprint arXiv:2210.10341, 2023.

[10] Y. Wang et al., "ChatDoctor: A medical chat model fine-tuned on a doctor-patient conversation dataset," arXiv preprint arXiv:2303.14070, 2023.

[11] M. Johnson et al., “MIMIC-IV: A freely accessible electronic health record dataset,” Scientific Data, vol. 8, no. 1, pp. 1–9, 2021.

[12] L. Lehman et al., “n2c2 shared tasks and resources for clinical NLP,” in Proc. AMIA Symposium, 2021, pp. 870–878.

[13] P. Rajpurkar et al., “PubMedQA: A dataset for biomedical research question answering,” ACL BioNLP Workshop, 2020, pp. 1–10.

[14] D. He et al., “Ethical and legal implications of deploying LLMs in healthcare,” Health Informatics Journal, vol. 30, no. 2, 2024.

[15] T. K. Nguyen et al., “Explainability in AI models for healthcare: A survey,” IEEE Reviews in Biomedical Engineering, vol. 17, pp. 101–115, 2024.

[16] B. Esteva et al., “A guide to deploying machine learning in healthcare,” Nature Medicine, vol. 25, pp. 24–29, Jan. 2020.

[17] T. Gebru et al., “Datasheets for datasets,” Communications of the ACM, vol. 64, no. 12, pp. 86–92, Dec. 2021.

[18] PRISMA Group, “Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement,” PLoS Med., vol. 6, no. 7, e1000097, 2009.

[19] A. T. Nguyen, H. Tran, and S. Li, “Large Language Models in Healthcare: A Systematic Evaluation Framework,” Journal of Biomedical Informatics, vol. 139, 2024, Art. no. 104383.

[20] S. Jeblick et al., “ChatGPT Makes Medicine Easy to Swallow: An Exploratory Case Study on Simplifying Radiology Reports,” European Radiology, vol. 33, no. 4, pp. 2472–2480, 2023.

[21] K. Patel, J. Mehta, and M. Sharma, “Ethical AI for Clinical Applications: From Transparency to Trust,” IEEE Transactions on Technology and Society, vol. 4, no. 1, pp. 1–11, Mar. 2023.

[22] J. Zhao, H. Liu, and M. Zhang, “Towards Robust and Interpretable LLMs in Medicine,” in Proc. IEEE Int. Conf. on Bioinformatics and Biomedicine (BIBM), 2023, pp. 145–152.

[23] D. Tao et al., “Federated Learning for Privacy-Preserving Healthcare AI: Opportunities and Challenges,” IEEE Internet of Things Journal, vol. 10, no. 8, pp. 6815–6827, Apr. 2023.

[24] M. J. Beam and C. Kohane, “Big Data and Machine Learning in Health Care,” JAMA, vol. 324, no. 11, pp. 1033–1034, 2020.

[25] Y. Zhang et al., “Multimodal Transformers for Clinical Decision Support: A Review,” IEEE Journal of Biomedical and Health Informatics, vol. 28, no. 2, pp. 430–443, Feb. 2024.