British Journal of Multidisciplinary and Advanced Studies,7(2),1-30, 2026

Computer Science

Print ISSN: 2517-276X

Online ISSN: 2517-2778

https://bjmas.org/index.php/bjmas/index

Publication of the European Centre for Research Training and Development -UK

The Evolution of Small Language Models in Healthcare: A Narrative-Evolutionary Literature Review

Saurabh Raj Sangwan[1], Akshi Kumar [2]*

[1]Department of Computer Science Engineering, Maharishi Markandeshwar Engineering College, Maharishi Markandeshwar (Deemed to be University), Mullana-Ambala, Haryana, India

[2]School of Computing, Goldsmiths, University of London, London, United Kingdom

doi: https://doi.org/10.37745/bjmas.0524 Published March 05, 2026

Citation: The Evolution of Small Language Models in Healthcare: A Narrative-Evolutionary Literature Review, *British Journal of Multidisciplinary and Advanced Studies*,7(2),1-30

Abstract: *Small Language Models (SLMs) are emerging as transformative tools in healthcare AI, offering an optimal balance between task performance, privacy, and computational efficiency. Unlike Large Language Models (LLMs), which require substantial infrastructure and raise privacy concerns, SLMs are designed for domain-specific applications in real-time, resource-constrained clinical settings. This narrative-evolutionary review traces the development, capabilities, and deployment of SLMs across key healthcare domains, including clinical documentation, decision support, mental health, and telehealth. It outlines enabling technologies such as model compression, federated learning, and edge deployment that facilitate secure and efficient operation within sensitive healthcare infrastructures. To assess their clinical utility, we introduce a domain-relevance matrix and evaluate SLMs using both conventional and deployment-aware metrics, such as latency, memory usage, and compliance with data regulations. The review also identifies critical research gaps in benchmarking practices, data diversity, and explainability. We emphasize the importance of collaborative data ecosystems, standardized evaluation protocols, and socio-technical governance to ensure safe, multimodal, and regulation-compliant integration of SLMs in healthcare workflows. Overall, our findings position SLMs as a scalable, transparent, and context-aware solution for advancing real-world medical NLP, especially in scenarios demanding adaptability, trustworthiness, and decentralized AI performance.*

Keywords: Small Language Models (SLMs); Healthcare AI; Clinical Decision Support; Medical NLP

INTRODUCTION

Artificial Intelligence (AI) has significantly transformed healthcare, enhancing diagnostics, patient care, clinical decision-making, and operational efficiency [1]. AI-driven tools now enable faster and more accurate diagnoses, automated documentation, and personalized treatment. One of the most transformative applications of AI is Natural Language Processing (NLP), which facilitates the analysis of electronic health records (EHRs), radiology reports,

clinical notes, and patient-provider communications [2]. These technologies have streamlined clinical workflows and accelerated biomedical research, drug discovery, and AI-powered engagement through medical chatbots and virtual assistants.

Large Language Models (LLMs) such as GPT-4, BioBERT, Med-PaLM, and ChatGPT have demonstrated advanced capabilities in contextual reasoning and biomedical knowledge extraction [3, 4]. Trained on large-scale datasets comprising clinical trials, scientific literature, and healthcare dialogues, LLMs support tasks including disease diagnosis, summarization of research findings, outcome prediction, and administrative task automation. Their ability to generate coherent, human-like responses has made them valuable for decision support in clinical environments. However, despite these benefits, their deployment in real-world healthcare faces significant limitations [5–7], including high computational demands, limited interpretability, privacy vulnerabilities, and reliance on cloud infrastructure. These factors challenge their compliance with regulations like HIPAA and GDPR and complicate their integration in high-risk medical workflows, where trust and transparency are essential.

To mitigate these challenges, researchers are turning to Small Language Models (SLMs) compact, efficient NLP models tailored for healthcare environments [8, 9]. More formally, SLMs denote lightweight neural architectures tailored for domain-specific language tasks, engineered to operate under constraints of computation, latency, and data privacy. Designed for strong task-specific performance with minimal computational overhead, SLMs are well-suited for decentralized, real-time, and regulation-sensitive deployments such as rural hospitals, mobile health applications, and wearable medical devices. Unlike LLMs, they can run locally, reducing latency and data exposure, and enhancing compliance with regulatory standards. Their explainable nature and ethical design further support adoption in high-stakes healthcare contexts. Importantly, SLMs are gaining traction in diverse clinical applications, including structured documentation, radiology report summarization, clinical triage, and AI-based medical assistance. Their adaptability allows deployment in low-connectivity areas, broadening access to AI-enabled care. Additionally, SLMs align with emerging demands for privacy-preserving, energy-efficient, and transparent AI systems. This shift reflects a growing recognition of the need for AI models that are not only powerful but also socially responsible and deployable.

In spite of the progress, the current literature remains limited in its systematic examination of SLMs. Whilst much of the focus has been on LLMs in healthcare [10–13], few studies have explored the technical evolution, deployment readiness, or domain-specific impact of SLMs. There is also a lack of clarity on how SLMs may offer more feasible and ethical solutions to longstanding challenges in AI-driven healthcare. This literature review aims to fill that gap by examining the development, applications, and implications of SLMs in healthcare AI. The study is guided by the following research questions:

- What key technological advancements have enabled the development of SLMs for healthcare applications?
- How do SLMs compare to LLMs in computational efficiency, privacy preservation, and real-world adaptability?

- What challenges remain in implementing SLMs in clinical settings, and how can future research address them?

This review is structured as follows: Section 2 outlines the fundamentals of Small Language Models (SLMs) and their distinctions from LLMs. Section 3 traces the evolution of AI in healthcare. Section 4 examines real-world SLM applications in clinical settings. Section 5 compares SLMs and LLMs, highlighting research gaps. Section 6 evaluates SLM performance on efficiency and privacy. Section 7 introduces a domain-relevance matrix to assess impact. Section 8 discusses key limitations and the future directions. Finally, section 9 concludes with synthesized insights and recommendations.

Methodology: Narrative-Evolutionary Approach

This review employs a Narrative-Evolutionary approach, a hybrid methodology that combines historical analysis with an evolutionary lens to trace the development of Small Language Models (SLMs) in healthcare. Grounded in narrative synthesis frameworks in health research [14] and models of technological evolution [15], this method is mainly suited to fast-moving domains where terminology, use cases, and design paradigms evolve rapidly. Unlike traditional scoping or systematic reviews that organize studies thematically or by outcome, the Narrative-Evolutionary lens offers a temporal perspective, mapping developmental momentum, surfacing inflection points, and aligning fragmented research under the emerging SLM umbrella. This is essential for understanding not just what changed in medical NLP, but how and why.

Rather than presenting isolated findings, this approach constructs a coherent, chronologically structured narrative that highlights key transitions, technological breakthroughs, and contextual constraints shaping SLM adoption. It emphasizes continuity and causality, linking past advancements to present capabilities and future directions. Importantly, prior to 2024, many domain-specific lightweight models were not explicitly labelled as "SLMs," hindering conceptual consolidation. This review retrospectively unifies them under the SLM framework, clarifying their cumulative impact on medical AI. Though narrative in structure, the methodology maintains scholarly rigour through a transparent, multi-step process of literature sourcing, selection, and synthesis.:

- **Framing the Review Scope and Research Questions:** The review began by identifying core research questions focused on the emergence, differentiation, and application of SLMs in healthcare. This ensured conceptual alignment with the evolutionary perspective, emphasizing longitudinal trends, turning points, and gaps.
- **Search Strategy Development:** To ensure comprehensiveness, this review adopted a structured yet flexible literature retrieval strategy. Within the healthcare domain, databases such as PubMed, IEEE Xplore, ACM Digital Library, and arXiv were systematically searched. The search terms evolved in line with the maturation of the field, beginning with keywords like *"lightweight transformer," "domain-specific NLP," "clinical BERT,"* and *"on-device NLP."* As the terminology gained clarity, mostly in 2023–2024, more targeted terms such as *"Small Language Models (SLMs)"* were included to capture emerging works explicitly addressing compact, domain-specific models. This iterative approach allowed the review to reflect both historical

Print ISSN: 2517-276X
Online ISSN: 2517-2778
https://bjmas.org/index.php/bjmas/index
Publication of the European Centre for Research Training and Development -UK

progression and the evolving language of the field, capturing foundational contributions as well as recent empirical advancements.

- **Chronological Curation and Thematic Mapping:** Retrieved articles were not only categorized by topic but mapped along a temporal axis to establish a historical trajectory. Each model or contribution was positioned within its respective developmental wave (e.g., rule-based era, statistical NLP, deep learning, transformer compression).
- **Identification of Technological and Conceptual Inflection Points:** Key turning points were identified based on shifts in methodology (e.g., introduction of knowledge distillation), application settings (e.g., move from lab to edge deployment), and emerging concerns (e.g., interpretability, privacy, cost). These helped frame the review as an evolving narrative.
- **Narrative Synthesis:** The selected works were synthesized into a cohesive storyline, focusing not just on technical progression, but on contextual drivers such as clinical needs, regulatory environments, and deployment challenges. This narrative lens helped surface broader patterns and causal relationships not evident in traditional thematic reviews.
- **Integration of Recent Empirical and Preprint Literature (2023–2024):** Given the recency of the SLM discourse, especially as an explicit term, newer preprints and grey literature (e.g., SSRN, arXiv, conference proceedings) were incorporated to ensure coverage of the latest benchmarks, use cases, and conceptual shifts.
- **Conceptual Unification Under the 'SLM' Framework:** Studies not originally labelled as SLMs were evaluated against definitional criteria (e.g., parameter size, deployment feasibility, domain specificity) and, if fitting, included to reflect the field's latent but evolving identity. This ensured terminological consistency and analytical depth.

To ensure inclusion quality, studies had to meet at least one of the following criteria: present empirical findings (e.g., benchmark scores), offer theoretical insights into model design, or demonstrate real-world healthcare applications. Studies were excluded if they lacked reproducibility, relied on speculative claims without technical detail, or focused solely on general-purpose LLMs. Thematic derivation followed an inductive, multi-phase coding process: after arranging studies chronologically, recurring patterns, such as modular fine-tuning, privacy-preserving inference, and domain adaptation were identified and grouped into meta-themes aligned with key inflection points. This preserved both conceptual coherence and the temporal integrity essential to the Narrative-Evolutionary framework.

In addition, four key design principles shaped the structure of this review. First, tracing the historical development of medical NLP that is from early expert systems to transformer-based models clarifies the emergence of SLMs as essential, resource-efficient alternatives to LLMs. Second, technological and regulatory turning points (e.g., knowledge distillation, model compression, domain-specific fine-tuning, and privacy-preserving AI) highlight how SLMs gained traction in healthcare. Third, the review adopts a structured, chronological narrative to make this evolution accessible to interdisciplinary readers. Finally, by reflecting on past challenges (e.g., inefficiencies, regulatory barriers) and projecting future directions (e.g.,

federated learning, on-device inference, multimodal integration), the review offers a coherent roadmap for advancing ethically grounded, context-aware SLM development in clinical settings.

Chronological Evolution of Language Models in Healthcare

The development of language models in healthcare has been shaped by decades of advancements in NLP, deep learning, and AI-driven clinical applications. From early rule-based expert systems to modern transformer-based models, each phase of evolution has contributed to making medical AI more efficient, scalable, and privacy-conscious. Fig. 1 illustrates the chronological evolution of language models in healthcare, tracing key advancements from early rule-based systems such as MYCIN (1970s) [16] to modern transformer-based models like Med-PaLM [17] and BioGPT [18]. It highlights the transition from statistical NLP to deep learning, emphasizing the growing focus on privacy-preserving AI solutions in clinical applications.

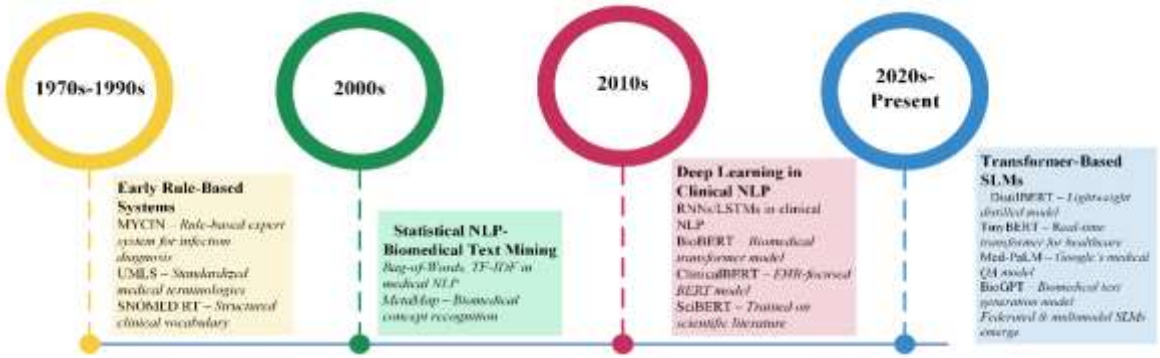

Fig.1. Language Model Development in Healthcare

Early Rule-Based Systems (1970s–1990s)

The earliest attempts at AI in healthcare relied on rule-based expert systems, where medical knowledge was represented through decision trees and manually crafted rules [19]. MYCIN [16] was among the first AI-driven medical systems, utilizing rule-based logic to assist in diagnosing bacterial infections and recommending antibiotic treatments. These systems were limited in adapting to real-world variability but played a foundational role in structured medical knowledge representation and later NLP applications in healthcare. To standardize medical terminology and facilitate NLP-based text analysis, the Unified Medical Language System (UMLS) [20] was introduced, creating a structured knowledge framework for biomedical applications. By the late 1990s, systems like SNOMED RT laid the groundwork for more structured clinical documentation [21]. These efforts culminated in the 2002 introduction of SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms) [22], which further streamlined medical documentation, enabling structured clinical coding and automated NLP-driven interpretations in electronic health records (EHRs).

Emergence of Statistical NLP and Machine Learning (2000s)

The 2000s marked a transition from rule-based systems to data-driven NLP, leveraging probabilistic models for more flexible and context-aware medical text processing. Bag-of-

Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) became widely used for biomedical text mining [23], clinical document classification [24], and drug discovery applications [25]. Although foundational, these methods enabled AI to extract patterns from unstructured medical text, paving the way for probabilistic models that improved clinical text mining. A breakthrough came with MetaMap, developed by the National Library of Medicine (NLM) [26], which enabled automated concept recognition in medical literature, significantly improving disease classification, information retrieval, and medical text summarization.

Deep Learning Revolution in Clinical NLP (2010s)

The 2010s witnessed a paradigm shift in medical NLP, where models transitioned from statistical pattern matching to contextual text understanding, enabling superior entity recognition, clinical summarization, and automated risk assessment [27-29]. The introduction of Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTMs) allowed for better handling of complex medical language, enabling improved EHR processing [30, 31], disease prediction [32, 33], and clinical summarization [34, 35]. This momentum led to the development of domain-specific transformer models, including BioBERT [36], which was trained on biomedical literature, enhancing named entity recognition (NER), question answering (QA), and relation extraction in medical applications. Following this, ClinicalBERT [37] was introduced to process unstructured EHRs, enabling AI-driven patient risk prediction and automated medical summarization. Around the same time, SciBERT [38] was developed to advance biomedical research NLP tasks, improving scientific document retrieval and knowledge discovery in the life sciences.

Rise of Transformer-Based SLMs (2020s–Present)

The 2020s marked a major shift towards transformer-based language model, emphasizing efficiency, adaptability, and privacy-conscious AI deployment [39, 40]. With the introduction of model compression and knowledge distillation techniques, researchers could retain the capabilities of large transformer models though reducing computational costs [41, 42]. DistilBERT [43] pioneered this approach, making real-time clinical NLP processing feasible on low-resource devices [44]. This was further improved by TinyBERT [45], a highly optimized transformer designed for on-device AI inference [46], allowing real-time medical text processing in hospitals, telehealth, and mobile health applications. Recognizing the need for medical question-answering AI, Med-PaLM [17] was developed as a large-scale model for medical QA, reinforcing the need for smaller, privacy-compliant alternatives in real-world clinical applications. Similarly, BioGPT [18], a transformer-based model specialized in biomedical text generation, enhanced AI-assisted medical documentation, and literature summarization, helping physicians and researchers process large volumes of medical knowledge faster.

Thus, the evolution of language models in healthcare reflects a journey of continuous refinement, from the early rule-based expert systems to today's advanced data-driven approaches. A new category has recently gained prominence: Small Language Models (SLMs), which are compact, efficient models designed for targeted tasks and resource-constrained environments [47]. SLMs are now driving progress toward more specialized, real-time, and privacy-preserving AI applications [8]. With an increasing shift toward on-device AI

processing, these models enable low-latency, privacy-conscious deployments covering hospitals, telehealth platforms, and wearable health devices. As the field advances, addressing critical challenges such as domain-specific adaptation, explainability, and ethical compliance will be essential to ensure the safe, trustworthy, and effective integration of language models into clinical workflows.

Understanding Small Language Models (SLMs)

Small Language Models (SLMs) are a specialized subset of NLP models that prioritize efficiency, domain adaptation, and computational accessibility [48]. In contrast to Large Language Models (LLMs), which may comprise billions to trillions of parameters and demand substantial computational resources, SLMs operate with significantly fewer parameters. This compact architecture makes them ideal for real-time, edge-based, and privacy-sensitive applications in healthcare. SLMs are particularly well-suited to low-resource environments, offering scalable, high-performance NLP solutions without the need for large-scale infrastructure. Their growing adoption includes clinical decision support systems, medical chatbots, and biomedical text mining. To clarify their distinct value, Table 1 compares SLMs and LLMs across key dimensions.

Table 1. Comparative Analysis of SLMs vs. LLMs in Healthcare

Feature	**Small Language Models (SLMs)**	**Large Language Models (LLMs)**
Parameter Size	Typically, 1 million to 500 million parameters	Billions to trillions of parameters (e.g., GPT-4: ~1.5 trillion parameters)
Computational Efficiency	Optimized for low-power devices	Requires high-end GPUs/TPUs
Inference Speed	Faster due to compact size	Slower due to complex computations
Resource Requirements	Can run on local devices, mobile phones, or edge servers	Needs cloud-based or distributed computing
Privacy & Security	Enables on-premise data processing, better for GDPR/HIPAA compliance	Requires external cloud processing, increasing data security concerns
Domain-Specificity	Fine-tuned for clinical texts, radiology, genomics	General-purpose, may need extensive fine-tuning
Adaptability	Suitable for personalized healthcare AI	Challenging to personalize due to high data needs
Cost & Accessibility	Affordable & scalable for low-resource settings	Expensive to train & deploy
Explainability	Easier to interpret due to smaller size and modularity	Often opaque and harder to debug in clinical contexts

While LLMs demonstrate strong capabilities in biomedical literature processing, diagnostic assistance, and clinical report summarization, their integration into healthcare systems remains limited due to several critical drawbacks:

- **High computational demands**: Training and inference require substantial hardware investments.
- **Privacy concerns**: Cloud-based processing raises significant compliance risks under HIPAA and GDPR.

- **Latency issues**: Inference delays hinder their use in time-sensitive or emergency contexts.
- **Low interpretability**: The black-box nature of LLMs erodes clinician trust in their outputs.

Moreover, studies have reported instances where LLMs like ChatGPT hallucinate medical information, fabricate citations, or generate unsafe clinical suggestions [49, 50], making them unsuitable for deployment in regulated, high-stakes environments. These limitations underscore the growing appeal of SLMs models that are smaller, faster, privacy-preserving, and tailored for clinical workflows. As healthcare increasingly integrates AI tools, the demand for transparent, domain-specific, and resource-efficient solutions positions SLMs as a compelling alternative to their larger counterparts.

Key Characteristics of SLMs

The unique advantages of SLMs stem from their efficiency, adaptability, and compliance with privacy regulations, making them a viable alternative to traditional AI models.

Computational Efficiency and Edge Deployment

One of the standout benefits of SLMs is their low computational footprint, which makes them highly suitable for edge computing, mobile apps, and resource-constrained environments [51]. Unlike Large Language Models (LLMs) that require substantial cloud infrastructure, SLMs are optimized to run locally on smartphones, embedded systems, or on-device processors without the need for constant internet connectivity or server support. This enables real-time performance in latency-sensitive scenarios such as customer service chatbots, IoT devices, smart assistants, and industrial automation systems. Their compact architecture and use of model compression techniques make them ideal for scenarios where speed, power efficiency, and responsiveness are critical.

Domain-Specific Training and Adaptability

SLMs are often fine-tuned on curated, domain-specific datasets, allowing them to outperform general-purpose models in specialized tasks [52]. Whether applied in legal tech, finance, education, or manufacturing, SLMs can be trained to understand the terminology and context of a particular industry. This leads to more accurate, context-aware outputs without the computational overhead required to retrain or adapt large models. Their adaptability ensures better performance in specific use cases like contract summarization, sentiment analysis, technical support, or personalized education tools, making them an effective solution when precision matters most.

Real-Time Medical Decision Support

SLMs are built for speed. Their lightweight nature ensures rapid inference, enabling them to provide instant feedback, predictions, or recommendations in various time-sensitive scenarios [53]. This makes them valuable for use in fields like customer engagement, fraud detection, intelligent tutoring systems, or operational analytics. In environments where decisions must be made quickly and accurately, SLMs enable seamless integration with existing software and infrastructure, delivering fast and relevant information when it’s needed most.

Privacy Preservation and Compliance with Medical Regulations
In an era of increasing concern around data privacy and AI ethics, SLMs offer a clear advantage. Because they are compact and can be deployed on-premises or directly on user devices, they minimize the need to transmit sensitive data to external servers [54]. This approach strengthens user trust and simplifies compliance with data protection regulations such as GDPR, CCPA, and other industry-specific standards. By keeping data local, organizations maintain greater control over their AI systems, reduce the risk of breaches, and avoid over-reliance on third-party cloud services without sacrificing access to powerful language capabilities.

Energy Efficiency and Sustainability
SLMs align with the growing emphasis on sustainable AI. Their reduced parameter size and computational requirements lead to significantly lower energy consumption compared to LLMs. This not only minimizes the environmental impact of model training and deployment but also makes SLMs suitable for settings where energy is a limiting factor, such as rural clinics, mobile health units, or off-grid environments. As green AI gains traction, the low carbon footprint of SLMs enhances their viability for responsible deployment.

Modular Design and Ease of Integration
SLMs are often developed using modular architectures, making it easier to plug them into existing digital ecosystems without extensive refactoring. Their smaller size and compatibility with standard NLP pipelines or embedded APIs allow faster prototyping and integration. This is especially beneficial for startups or small-scale deployments in healthcare and legal tech where agility and time-to-market are crucial.

Together, these characteristics underscore the growing shift toward right-sized AI, where smaller, smarter, and more interpretable models are prioritized over brute-force computation

Key Technological Advancements Driving SLMs in Healthcare
The increasing relevance of SLMs in healthcare has been made possible through a series of significant breakthroughs in AI architecture and deployment. Advancements in transformer-based modelling, coupled with strategies like knowledge distillation and quantization, have significantly reduced the computational burden of language models without compromising performance. Domain-specific fine-tuning has further improved accuracy in clinical tasks. Furthermore, federated learning (privacy-preserving technique in which models are trained locally on institutional data and only model updates, not raw data are shared) and edge AI (deploying models directly on local devices like smartphones or hospital monitors to enable real-time, offline inference) capabilities have addressed pressing concerns around data privacy, latency, and infrastructure constraints. These advancements collectively empower SLMs to operate in real-time, resource-constrained, and privacy-sensitive healthcare environments, paving the way for more ethical, efficient, and accessible AI-driven medical applications.

Print ISSN: 2517-276X
Online ISSN: 2517-2778
https://bjmas.org/index.php/bjmas/index
Publication of the European Centre for Research Training and Development -UK

Transformer-Based Architectures and Efficient Model Scaling

The transformer model, introduced by Vaswani et al. [55], revolutionized NLP through its self-attention mechanism, allowing models to capture long-range dependencies more effectively than RNNs or CNNs. This breakthrough led to the creation of context-aware models suitable for complex biomedical tasks. However, traditional transformer-based models, such as BERT [56] and GPT-3 [57], were computationally expensive, making them impractical for real-time clinical use. The need for smaller and more efficient versions of transformers led to the development of distilled and optimized architectures, such as DistilBERT [43] and TinyBERT [45], which retained the core advantages of transformers while significantly reducing their size and computational footprint. These smaller models could now be deployed in low-resource environments, including mobile health applications and hospital IT systems, making real-time medical decision support feasible.

Model Compression and Knowledge Distillation

One of the most critical advancements in the development of SLMs has been the introduction of model compression techniques, which allow for reducing the size of large models without sacrificing accuracy, thus enabling their deployment in real-world healthcare environments. Knowledge distillation, introduced by Hinton et al. [58], plays a critical role in this process by allowing a smaller model (student model) to learn from a larger, pre-trained model (teacher model). This process transfers knowledge from complex, computationally heavy models to compact, efficient versions, making it possible to use high-performing AI models in real-world healthcare applications. Model pruning and quantization have further contributed to reducing memory and computational requirements, enabling SLMs to operate on embedded medical devices such as wearables, bedside monitoring systems, and mobile diagnostic tools. Recent advancements in quantization-aware training (QAT) [59] and low-rank matrix factorization [60] have improved the ability of compressed models to maintain high accuracy even when deployed in environments with constrained computational power.

These techniques have proven especially effective in biomedical NLP, where models such as BioDistilBERT and ClinicalTinyBERT enable fast and accurate processing of electronic health records (EHRs), clinical trial data, and radiology reports, ensuring adherence to privacy regulations [61]. Further advancements have led to compact models like BioTinyBERT, BioMobileBERT, and CompactBioBERT, developed through strategies such as knowledge distillation and continual learning on domain-specific corpora like PubMed [61]. By offering a strong trade-off between computational efficiency and performance, these models are well-suited for deployment in privacy-sensitive, real-world clinical environments and on low-resource or edge devices.

Domain-Specific Fine-Tuning for Medical NLP

A major limitation of general-purpose LLMs, such as GPT-4 and ChatGPT, is their lack of domain specificity when applied to biomedical and clinical text [62]. Medical language is highly technical, and many general NLP models fail to understand medical jargon, clinical abbreviations, and disease-specific terminology. The introduction of domain-specific pre-training and fine-tuning has been a game-changer in enhancing the relevance of NLP models in healthcare.

Pre-trained biomedical models such as BioBERT [36] and ClinicalBERT [37] have demonstrated that fine-tuning on domain-specific corpora significantly improves performance in medical text classification, question answering, and summarization tasks. The same principle applies to SLMs, which benefit from pre-training on smaller, highly curated medical datasets to improve their accuracy in real-world healthcare applications. By focusing on disease-specific literature, radiology reports, and pharmaceutical interactions, SLMs can be optimized to function as specialized AI assistants for medical professionals. Furthermore, contrastive learning [47] and multi-task fine-tuning [63] have enabled SLMs to be more contextually aware, allowing them to generate more reliable, interpretable, and bias-reduced medical outputs. These improvements have made SLMs effective for triaging patient symptoms, analyzing clinical notes, and assisting with evidence-based medical decision-making.

Federated Learning for Privacy-Preserving AI

A major challenge in healthcare AI is ensuring data privacy and compliance with regulations such as HIPAA (Health Insurance Portability and Accountability Act) and GDPR (General Data Protection Regulation). Traditional AI models require centralized datasets for training, raising concerns about patient data security and confidentiality.

Federated learning [64] has emerged as a key solution, allowing SLMs to be trained over multiple decentralized medical institutions without exposing sensitive patient data [65]. This approach enables hospitals and research centres to collaborate on AI model training in a privacy-preserving manner, ensuring that medical AI models become more robust and generalizable without compromising patient confidentiality. In federated learning settings, SLMs can be continuously improved on-site, within hospital networks, without the need for external cloud-based processing. This approach has benefited fields like radiology, pathology, and genomics, enabling models to learn from distributed, multi-institutional datasets without compromising data security.

Advancements in Edge AI for On-Device Processing

The integration of Edge AI with SLMs has further accelerated their adoption in healthcare by enabling real-time, low-latency AI processing on medical devices. Unlike cloud-based AI systems, which require internet connectivity and centralized servers, Edge AI allows language models to run directly on devices such as hospital servers, point-of-care testing tools, and wearable health monitors. Recent developments in hardware-optimized deep learning, such as Tensor Processing Units (TPUs) and low-power AI accelerators, have facilitated the deployment of efficient NLP models on portable healthcare devices [66]. These developments allow SLMs to analyze patient data in real time, providing instant guidance for clinical decision support without relying on cloud connectivity.

The development of SLMs for healthcare has been driven by multiple technological advancements, each playing a critical role in improving model efficiency, domain specificity, and privacy compliance. Fig. 2 visually represents the five key advancements that have enabled SLMs to become viable for real-world medical applications. These include transformer-based architectures, model compression and knowledge distillation, domain-specific fine-tuning, federated learning for privacy preservation, and Edge AI for on-device processing. Together,

these advancements have made it possible to deploy fast, lightweight, and privacy-conscious AI solutions in clinical environments.

Fig.2. Key technological enablers that have accelerated the deployment of SLMs in real-world healthcare settings.

SLMs are more than smaller alternatives to LLMs, they represent the future of practical, ethical, and scalable AI in healthcare. By offering real-time processing, privacy preservation, domain specialization, and cost efficiency, these models bridge the gap between AI development and real-world clinical adoption. Their ongoing evolution is set to power the next generation of AI solutions that emphasize efficiency, security, and patient-centred care throughout diverse healthcare environments. The following section outlines the key enablers driving the emergence of Small Language Models (SLMs) in healthcare, focusing on transformer architectures, model compression techniques, domain-specific fine-tuning, federated learning, and advances in edge AI.

Prior Research on Small Language Models (SLMs) in healthcare

The rapid advancement of AI in healthcare has spurred extensive research into the roles of both LLMs and SLMs in clinical tasks. LLMs like GPT-4 and Med-PaLM have demonstrated impressive capabilities specifically in medical education and knowledge synthesis but their deployment is limited by high computational demands, challenges in interpretability, and privacy concerns [67]. Recent systematic reviews [68, 69] have highlighted several limitations in how LLMs are currently evaluated pointing to a disproportionate focus on exam-based tasks, an underrepresentation of real-world clinical workflows (e.g., note-taking, triage, prescription generation), and insufficient attention to fairness, bias, and uncertainty. For example, Shool et al. [69] found that only 5% of LLM evaluations incorporated real patient care data, and 93.5% were based on general-domain models rather than clinical ones. Similarly, Bedi et al. [68] emphasize that evaluation dimensions like robustness and deployment feasibility remain underexplored, raising concerns about the practical readiness of LLMs in high-stakes healthcare contexts.

These limitations have catalyzed a growing shift toward the development and adoption of Small Language Models (SLMs). Designed with fewer parameters, domain specialization, and privacy-preserving capabilities, SLMs are increasingly recognized as more suitable for real-time, on-device healthcare applications. Although the term "SLM" has gained formal recognition only in recent years, many earlier models, often labelled "lightweight," "compressed," or "distilled", demonstrated the same core principles: reduced computational complexity, adaptability to limited-resource settings, and improved data compliance. As the field matured, especially post-2022, researchers began to collectively classify these models under the SLM umbrella, reflecting a unified design philosophy. What was once a dispersed set of advancements is now converging around a shared vision: developing efficient, task-specific language models that can function reliably in sensitive and resource-constrained clinical environments. This section synthesizes findings from key studies published between 2023 and 2025, offering an overview of how SLMs are being applied to various healthcare tasks. These works address a wide spectrum of applications, from clinical note summarization and question answering to named entity recognition (NER), triage support, and the development of explainable conversational agents. Together, they represent an evolving body of research that spans both academic inquiry and industry-led implementation efforts, illustrating the rising significance of SLMs in the healthcare AI ecosystem.

2023 marked significant strides in enhancing the practicality of Small Language Models (SLMs) for healthcare applications. Guo et al. [70] demonstrated that SLMs fine-tuned with synthetic data generated by LLMs could outperform few-shot GPT-4 on medical question answering benchmarks like PubMedQA. Their use of low-rank adaptation and generative augmentation highlighted the effectiveness of resource-efficient fine-tuning without increasing model size. In parallel, Hasan et al. [71] introduced a robust ensemble-distillation pipeline that distilled clinical outcome predictions into a lightweight DistilBERT model. Trained on MIMIC-III for mortality and length-of-stay tasks, the student model retained 97% of the ensemble's performance with a 70% reduction in parameters, offering a deployable solution for resource-constrained healthcare systems.

2024 brought an expansion into multilingualism, multimodality, and domain specialization. Rohanian et al. [61] explored compressing clinical language models through knowledge distillation (KD) to improve efficiency without sacrificing performance. DistilClinicalBERT, TinyClinicalBERT, and ClinicalMiniALBERT were derived from BioClinicalBERT using distinct KD strategies. Additionally, continual learning on MIMIC-III was used to improve the performance of compact models like BioDistilBERT and BioMobileBERT for clinical NLP tasks. Wang et al.'s Apollo framework [72] pushed multilingual training with models up to 7B parameters, covering six world languages and enabling proxy-tuning to improve larger LLMs. Building on reasoning capabilities, the Meerkat family incorporated chain-of-thought (CoT) fine-tuning using medical textbooks, leading Meerkat-7B to surpass the USMLE threshold comparable to GPT-4 while remaining open source [73]. pRAGe, developed by Buhnila et al., [74] tackled hallucination in medical text generation by integrating SLMs with external knowledge bases and prompt-tuned paraphrasing, showing effectiveness even in quantized models like BioMistral-7B-SLERP for French medical explanations Wang et al. [75] benchmarked five state-of-the-art SLMs, including TinyLlama and Phi-3-mini, for mobile

health event prediction, demonstrating that these models achieve comparable or superior performance to LLMs in tasks like stress, fatigue, and readiness prediction. Their deployment on devices such as the iPhone 15 Pro Max highlighted significant efficiency gains up to 15.5× faster inference and 9.9× improvement in first-token latency highlighting SLMs' suitability for privacy-preserving, real-time healthcare applications. Qu et al. [76] addressed privacy-sensitive deployment by applying targeted preprocessing to improve metastasis classification from unstructured EHR notes, using Gemma-2b and Gemma-7b models in secure, offline environments. To improve task-specific adaptability with minimal modification to pre-trained weights, Low-Rank Adaptation (LoRA) [77] was employed. Bjorkdahl et al. [78] added multimodal capability by integrating imaging, time-series, and text data using SLMs such as Phi-3-mini and Gemma-2B, paving the way for multitask disease risk prediction without task-specific optimization. DeviceBERT [79] further emphasized vocabulary adaptation in regulatory language processing, improving BioBERT's F1 score by over 13% for identifying medical device components from FDA recall summaries.

Domain-specific modelling also gained traction. Gwon et al. [80] introduced HeartBERT, a cardiology-focused SLM trained on curated PubMed data, showing that department-specific models outperformed general medical ones. Griewing et al. [81] developed BC-SLM, a breast cancer SLM aligned with German clinical guidelines. Demonstrating 86% concordance with tumour board decisions and surpassing ChatGPT-4 in transparency, BC-SLM illustrates the potential of SLMs to enhance clinical decision support, alongside improving explainability and data governance.

In the mental health domain, Diwakar and Raj [82] proposed a DistilBERT-based text classification approach to automate the diagnosis of mental health conditions such as anxiety, borderline personality disorder (BPD), and autism using data from online mental health communities.

Their lightweight model achieved 96% accuracy on a balanced dataset, showcasing the efficacy of SLMs in enabling fast, scalable, and non-invasive mental health screening.. Kumar et al. [83] introduced AST-D; a transformer-based summarization system trained on the curated *DepressiLex* dataset comprising recent research in depression detection. The study evaluated multiple lightweight transformer models, including BART, T5, PEGASUS, ProphetNet, and Longformer-Encoder-Decoder (LED) for their ability to generate concise and clinically useful summaries. Among these, LED emerged as the most effective for distilling complex and lengthy mental health literature into actionable findings, highlighting the capability of small, task-optimized models in evidence synthesis. A companion study by the same group [84] proposed a hybrid Biomedical Entity Linking (BEL) approach that integrates full-text search with compact embedding models such as BioBERT, MetaMap, FastText, and Llama, alongside a custom Depression Entity Relevance Ranker (DERR). This system achieved 84% accuracy and 95% Hits@5 in linking symptoms, medications, and comorbidities to structured knowledge bases like DSM-5 and UMLS. By emphasizing computational efficiency, domain-specific tuning, and improved clinical interpretability, both studies exemplify the practical application of Small Language Models (SLMs) in supporting mental health diagnostics, especially in resource-constrained settings.

2025 has further refined the role of SLMs in both structured reasoning and real-time, user-facing medical interfaces. Zong et al. [85] proposed *EvidenceMap*, a modular framework combining a 66M-parameter encoder-decoder with a 3B generator. Designed to mimic biomedical reasoning through evidence evaluation, logic modelling, and summarization, EvidenceMap outperformed an 8B retrieval-augmented LLM by nearly 20% in answer quality emphasizing the power of modular SLMs in high-fidelity medical QA. In a more patient-centric application, Magnini et al. [86] explored the feasibility of open-source SLMs as privacy-preserving components of medical assistant chatbots. By deploying lightweight models entirely on personal devices for hypertension management, their architecture avoided cloud reliance and maintained data sovereignty. While Gemini Pro 1.5 remained the benchmark leader in intent recognition, several open-source SLMs demonstrated competitive performance in empathetic and semantically accurate responses, positioning SLMs as viable tools in telemedicine and chronic condition self-management.

This trajectory of progress is reflected in Table 2, which synthesizes key contributions from recent research on Small Language Models (SLMs) in healthcare between 2023 and 2025. The table traces the evolution from early efforts in compact model development and distillation in 2023, to the emergence of multilingual, multimodal, and privacy-preserving applications in 2024, and further toward modular biomedical reasoning frameworks and patient-facing tools in 2025. This progression emphasises the increasing sophistication and specialization of SLMs, reinforcing their growing relevance and readiness for real-world clinical integration.

Table 2. Key Contributions of SLM-Based Healthcare Research (2023–2025)

Year	Study / Model	Focus Area	Key Contribution	Impact
2023	Guo et al. (SLM w/ Synthetic Data) [70]	Medical QA	Fine-tuning with GPT-generated synthetic data	Outperformed few-shot GPT-4 on PubMedQA with a smaller model
	Hasan et al. (Distilled DistilBERT) [71]	Clinical Outcome Prediction	Ensemble + distillation on MIMIC-III for mortality/LoS (Length of Stay)	97% of ensemble performance with 70% fewer parameters
	Rohanian et al. (BioDistilBERT etc.) [61]	Biomedical NLP	Knowledge distillation + continual learning on PubMed	Compact SLMs with 98% of BioBERT's performance
2024	Wang et al. (Apollo) [72]	Multilingual Medical NLP	Proxy-tuning + multilingual corpora (6 languages)	Generalizable SLMs for global populations
	Meerkat-7B [73]	Reasoning / Diagnostics	Textbook-driven chain-of-thought (CoT) fine-tuning	First open 7B model to pass USMLE threshold
	Buhnila et al. (pRAGe) [74]	Paraphrasing / Explainability	RAG + prompt tuning + French medical simplification	Accurate paraphrases using quantized models like BioMistral-7B-SLERP
	Wang et al. (TinyLlama) [75]	Mobile Health	Quantized SLMs for wearable signals + latency optimization	15.5× faster inference than GPT-4; real-time mobile health applications
	Qu et al. (Gemma-7B for EHR) [76]	Privacy-Preserving Healthcare	Regular-expression preprocessing + tuning on EHR notes	High metastasis classification in secure, local setups

2024	LoRA for Task-Specific Adaptation [77]	Efficient Fine-Tuning	Low-Rank Adaptation (LoRA) to adapt SLMs with minimal parameter updates	Enabled domain-specific tuning of SLMs with minimal computational overhead; critical for healthcare personalization and privacy
	Bjorkdahl et al. [78]	Multimodal Risk Prediction	Fusion of text, time-series, and imaging with Gemma/Phi-3 SLMs	Scalable multitask disease prediction (12 tasks)
	DeviceBERT [79]	Regulatory NER	Vocabulary enrichment for sub-domain device identification	+13% F1 score in low-data FDA contexts
	Gwon et al. (HeartBERT) [80]	Cardiology	Department-specific model built from scratch	Outperformed general medical models in cardiac QA tasks
	Griewing et al. (BC-SLM) [81]	Oncology Decision Support	Breast cancer SLM aligned with German clinical guidelines	86% concordance with tumour board; explainable and transparent
	Diwakar & Raj [82]	Mental Health Diagnosis	DistilBERT classifier for anxiety, BPD, autism	Achieved 96% accuracy on balanced mental health dataset
	Kumar et al. (AST-D) [83]	Mental Health Literature Summarization	Summarization of depression research using multiple SLMs	LED model most effective; reduces clinician cognitive load
	Kumar et al. (BEL for Depression) [84]	Biomedical Entity Linking	Hybrid of full-text search + embeddings + DERR ranker	84% accuracy, 95% Hits@5 for DSM-5/UMLS linking
2025	Zong et al. (EvidenceMap) [85]	Biomedical QA	Modular encoder-decoder + evidence modelling	19.9% better than 8B RAG model in answer quality
	Magnini et al. (Open-source SLM Chatbots) [86]	Telemedicine / Patient Self-Care	Fully local chatbot deployment on personal devices	Reliable intent detection + empathetic responses without cloud reliance

Fig. 3 presents a thematic mapping of these contributions, visually organizing the literature into six core application areas: Clinical Question Answering & Biomedical Reasoning, Clinical Outcome Prediction & Risk Stratification, Biomedical Entity Linking & Named Entity Recognition (NER), Real-Time, Privacy-Preserving & On-Device Healthcare, Multilingual & Domain-Specific SLMs, and Literature Summarization & Cognitive Support for Clinicians. This clustering reveals how SLM research has moved beyond proof-of-concept to address specific, high-impact tasks in both provider-facing and patient-facing contexts.

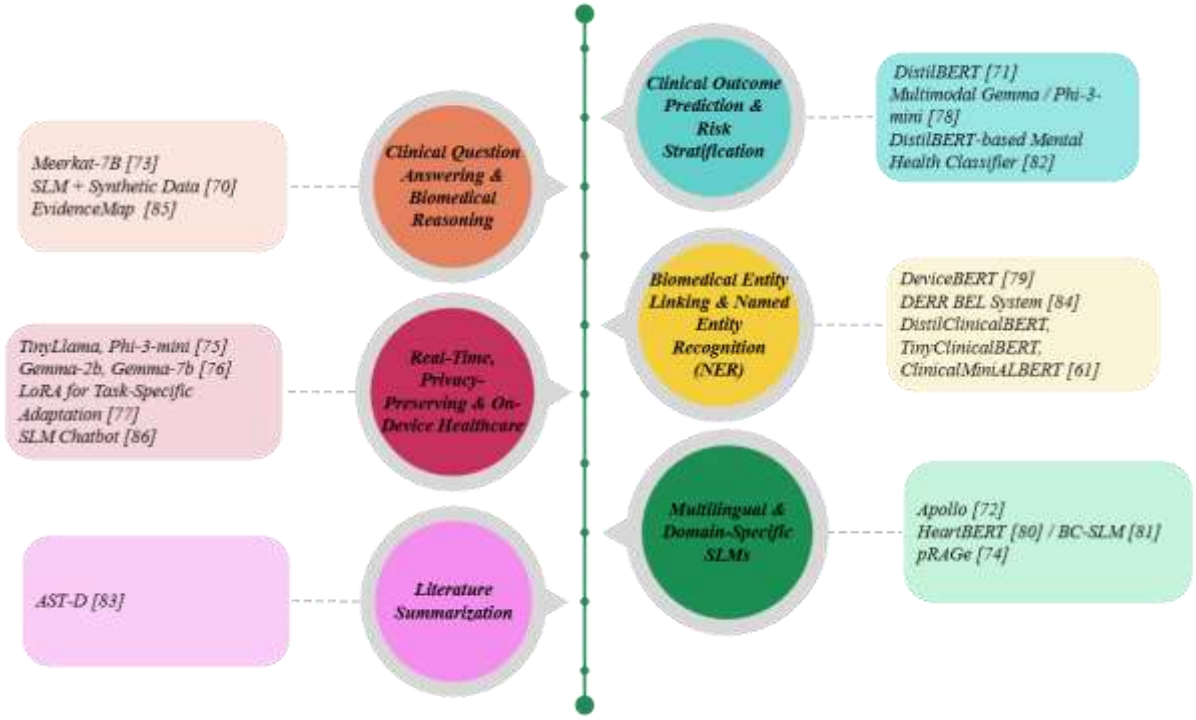

Fig.3. Thematic mapping of Small Language Model (SLM) applications in healthcare (2023–2025).

Key Applications of SLMs in Healthcare and Their Transformative Potential

Small Language Models (SLMs) are increasingly recognized as essential tools for enabling responsible, scalable, and context-sensitive artificial intelligence within diverse healthcare ecosystems. Unlike their larger counterparts, Large Language Models (LLMs), which often require significant computational resources and pose elevated risks to data privacy, SLMs offer a viable path toward real-time, privacy-conscious, and cost-effective AI deployment. Their compact architecture and fine-tuning adaptability position them as strategic assets in clinical environments constrained by limited infrastructure, strict regulatory requirements, and urgent decision-making scenarios. This section details six critical domains where SLMs are demonstrating transformative potential in modern healthcare delivery.

Clinical Documentation Automation

A foundational and significant application of SLMs is in the automation of clinical documentation processes. Physicians spend a substantial portion of their time recording clinical notes, entering diagnostic codes, and preparing discharge summaries, tasks that are essential but contribute significantly to administrative overload and clinician burnout. SLMs such as ClinicalMiniLM and DistilClinicalBERT, with their smaller parameter footprint and fine-tuning capabilities, are increasingly used for automating note generation, transcribing patient interactions, and summarizing lengthy consultations. Unlike LLM-based systems that typically depend on cloud-hosted infrastructure and remote data transfer, SLMs can be securely deployed on-premise within hospital intranets. This enables real-time transcription and summarization, with strict adherence to data governance standards such as HIPAA in the

United States and GDPR in the European Union. Future iterations may enable automatic extraction of key medical concepts from unstructured notes and facilitate standardized data formatting, improving the interoperability of electronic health records (EHRs) between systems and jurisdictions.

Clinical Decision Support

In time-sensitive clinical settings such as emergency departments, intensive care units (ICUs), and point-of-care rural clinics, rapid and reliable decision-making is critical. SLMs are well-positioned to serve as lightweight decision support systems that improve diagnostic accuracy without overwhelming the computational capacity of on-site infrastructure. Distilled models like ClinicalTinyBERT and DistilBERT, when fine-tuned on datasets such as MIMIC-III, have demonstrated high utility in predictive tasks including early detection of sepsis, abnormal lab result identification, patient triage prioritization, and recognition of drug interactions. Because SLMs can operate in low-latency environments and function offline, they are useful in clinical environments with limited or intermittent internet access. Furthermore, their ability to generalize within narrow clinical domains allows for integration into bedside dashboards, risk stratification tools, and hospital workflow systems, enabling scalable and explainable support for healthcare professionals.

Patient-Facing Conversational AI

SLMs are increasingly used to power patient-facing conversational agents that can operate in decentralized and privacy-sensitive contexts. In contrast to commercial LLM-based chatbots that often rely on internet-based APIs and data-sharing with third-party servers, SLMs provide a more secure, localized alternative suitable for healthcare environments. These conversational agents assist patients in a variety of ways, including symptom triage, medication adherence reminders, mental health self-assessments, and chronic disease check-ins. Models such as those developed by Magnini et al. [86] showcase how SLMs fine-tuned on intent recognition and dialogue datasets can maintain empathetic, context-aware conversations throughout multiple patient interactions, especially in settings where consistent clinical supervision is unavailable. Their reduced memory and power requirements allow them to be embedded within smartphones, tablets, or wearable health devices, thereby enabling equitable access in underserved, multilingual, or rural communities. Additionally, on-device processing ensures sensitive health conversations are not transmitted to external servers, strengthening patient trust and compliance with regional data protection mandates.

Remote Monitoring and Wearable Data Integration

SLMs are also gaining ground in the domain of remote patient monitoring, where continuous health data streams such as heart rate, oxygen saturation, sleep cycles, and stress markers are collected through wearable devices. By operating directly on edge devices, SLMs can perform lightweight, context-specific inferences that detect deviations from baseline health states and generate alerts for early intervention. Such deployment enables real-time anomaly detection without reliance on cloud infrastructure, offering advantages for patients with chronic illnesses who require ongoing monitoring. For instance, SLMs can detect early warning signs of atrial fibrillation, dehydration, or psychological distress, and respond with personalized recommendations or escalation triggers. Their compatibility with federated learning

Print ISSN: 2517-276X
Online ISSN: 2517-2778
https://bjmas.org/index.php/bjmas/index
Publication of the European Centre for Research Training and Development -UK

frameworks further supports decentralized model training directly on user devices, thereby preserving privacy. This unlocks a new layer of autonomy in mobile health (mHealth) applications and expands the clinical utility of wearables beyond fitness tracking to medically actionable information.

Medical Education and Clinical Training

In medical education, SLMs offer real-time, interactive learning tools that adapt to the evolving needs of medical students and junior clinicians. These models, when fine-tuned on specialty-specific corpora such as pharmacology guidelines or diagnostic protocols, can serve as bedside teaching assistants or digital preceptors. They provide concise explanations of diseases, drug mechanisms, and procedural steps, tailored to the learner's current clinical context or specialty area. Unlike static educational resources, SLMs can simulate dynamic clinical scenarios, support question-answering during ward rounds, and offer multilingual support in globally diverse training programs. Their low computational footprint allows deployment on personal laptops or institution-owned devices, facilitating learning without breaching institutional firewalls or requiring cloud access. As digital pedagogy continues to evolve, SLM-powered tools are expected to complement traditional curricula and democratize access to high-quality clinical education globally.

Biomedical Research and Knowledge Discovery

The ability of SLMs to ingest, process, and synthesize vast quantities of unstructured scientific text is proving transformative in biomedical research. Models like BioDistilBERT and EvidenceMap [85] exemplify the power of fine-tuned SLMs in extracting meaningful information from clinical trial reports, pharmacovigilance data, and biomedical literature repositories. These systems can identify complex relationships such as gene-disease associations, drug-event patterns, and longitudinal trends within treatment cohorts. SLMs also support literature summarization and evidence synthesis, enabling rapid creation of briefings or systematic review summaries that would otherwise require manual effort. Regulatory bodies and clinical researchers alike benefit from these capabilities when performing drug safety evaluations or formulating evidence-based guidelines. Furthermore, the integration of these SLM outputs into biomedical knowledge graphs facilitates real-time hypothesis generation and supports data-driven precision medicine at scale. Their reduced memory footprint ensures that such tools can be deployed in research environments with limited computational budgets, further enhancing accessibility and replicability.

To provide a high-level overview of these application domains, Fig. 4 presents a heatmap matrix illustrating the relevance and maturity of SLM adoption within core healthcare tasks. Cells are color-coded based on evidence from recent literature, where dark blue indicates mature applications with proven impact (score of 1), light blue indicates emerging use cases with growing interest (score of 0.5), and white denotes limited current applicability (score of 0). This matrix serves as a decision-support tool for clinicians, developers, and policymakers, aiding in the strategic prioritization of future SLM integration efforts.

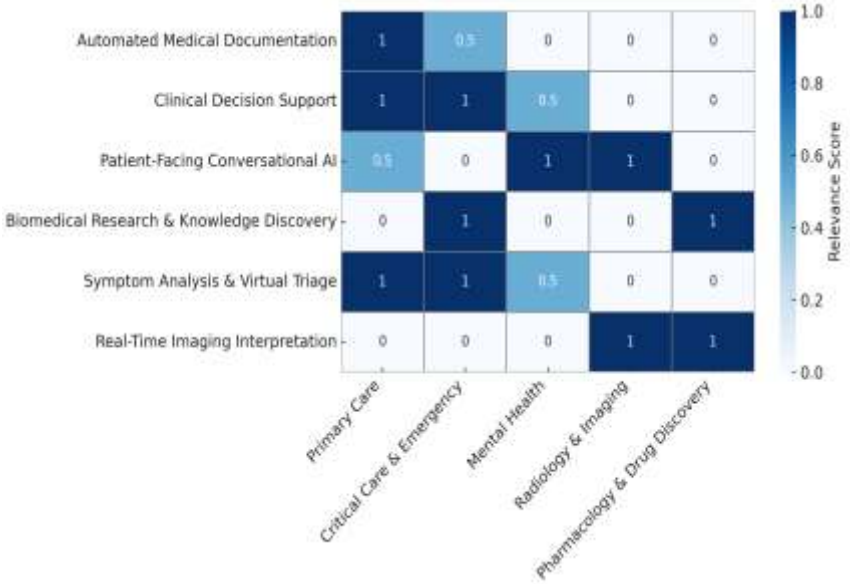

Fig. 4. SLM Applications Matrix within Healthcare Domains

Challenges in Implementing SLMs in Clinical Settings and Future Research Directions

This section examines the major challenges impeding the clinical implementation of SLMs and explores future research directions to improve their reliability, usability, and scalability in healthcare.

Model Interpretability and Trustworthiness

A critical barrier to SLM adoption in healthcare is the lack of interpretability and explainability in their decision-making process. Medical AI models must provide clinicians with clear, justifiable reasoning behind their outputs to ensure trust and clinical acceptance [87]. Unlike traditional rule-based expert systems, SLMs use deep learning architectures that operate as "black boxes", making it difficult for healthcare professionals to validate or understand their recommendations. To ensure trustworthy AI integration, future research must focus on developing explainable AI (XAI) techniques that can generate transparent, human-interpretable explanations for clinical predictions [88]. Methods such as attention visualization, saliency mapping, and counterfactual reasoning could be incorporated into SLM-based decision support systems to improve interpretability and physician confidence in AI recommendations.

Data Availability and Bias in Medical AI

SLMs are highly dependent on domain-specific training data, but access to large, high-quality medical datasets remains a challenge due to privacy restrictions, patient consent regulations, and institutional barriers [89]. Unlike LLMs, which are trained on massive publicly available corpora, SLMs require specialized clinical text, EHR records, and radiology reports to achieve domain-specific accuracy. However, imbalanced or biased training data can lead to systemic errors, disproportionately affecting underrepresented patient populations [90]. To address this, future research should focus on developing synthetic medical datasets that can simulate real-world clinical scenarios without exposing private patient data [91]. Additionally, federated learning frameworks can allow hospitals to train SLMs collaboratively without sharing raw

patient data, ensuring privacy, improving model robustness and generalization over diverse populations [92]. Recent work by Kumar et al. [93] applied Federated Transfer Learning (FTL) for privacy-preserved emotion recognition using biomarker-based EEG data, demonstrating that domain-specific models can be collaboratively trained using decentralized health data sources without compromising user privacy.

Integration with Clinical Workflows and Usability Concerns

For SLMs to be effective in real-world medical applications, they must be seamlessly integrated into electronic health records (EHR) systems, hospital IT infrastructure, and clinical decision support platforms [94]. However, many existing healthcare systems lack standardized APIs and interoperability frameworks, making it challenging to incorporate AI-powered tools into routine medical practice. In addition, healthcare professionals are often reluctant to adopt AI-based systems due to concerns about workflow disruption, increased cognitive load, and reliability issues [95]. To encourage SLM adoption, future research should focus on human-centered AI design, ensuring that models provide clinically meaningful outputs that support, rather than complicate the physician workflows. Further investment in natural language interfaces for AI-driven clinical documentation could help reduce administrative burdens on physicians, improving usability and adoption rates [96].

Regulatory and Compliance Challenges

The integration of SLMs in clinical practice is subject to strict regulatory scrutiny, given the high-stakes nature of medical AI. Compliance with frameworks such as HIPAA (Health Insurance Portability and Accountability Act), GDPR (General Data Protection Regulation), and FDA (Food and Drug Administration) guidelines remains a significant hurdle, as many AI models fall short on the explainability, transparency, and accountability standards required for safe clinical deployment [97, 98]. Although regulatory alignment is critical, it must also be supported by adherence to broader ethical principles that underpin trustworthy AI, namely fairness, transparency, privacy, accessibility, and compliance-readiness. Small Language Models (SLMs), by virtue of their lightweight, modular, and customizable nature, present unique opportunities to align more closely with these principles than conventional LLMs. This distinction is summarized in Table 3, which compares the ethical readiness of LLMs versus SLMs in healthcare AI.

Table 3. Ethical Considerations in LLM vs. SLM Deployment for Healthcare AI

Ethical Concern	LLMs	SLMs
Data Privacy	Cloud-based architectures risk patient data leakage	Enables on-device and federated learning setups to preserve privacy
Fairness	Often trained on general-domain, biased datasets	Can be fine-tuned on curated, demographically balanced clinical corpora
Explainability	Complex and opaque reasoning pathways	Smaller size and modular design enable clearer interpretability
Accessibility	Requires high-end GPUs and energy-intensive infrastructure	Runs on low-resource hardware; ideal for rural or mobile health setups
Compliance	Limited transparency for audit trails	Easier to integrate with regulatory auditing frameworks

Regulatory bodies require that AI-driven healthcare tools demonstrate clinical efficacy, patient safety, and algorithmic fairness before approval for real-world use [99]. Given the smaller scale and

targeted nature of SLMs, they are well-positioned to meet these demands, provided appropriate safeguards and validation mechanisms are in place. To accelerate responsible adoption, future research should focus on creating AI auditing frameworks that offer regulators standardized, transparent methods for evaluating model performance, bias, and interpretability. Additionally, cross-disciplinary collaboration among AI researchers, clinicians, and legal experts is essential to co-develop ethical governance models that support real-world SLM deployment in medicine [100].

Scalability and Real-World Validation

Most SLMs are trained and validated in controlled research environments, but their performance in real-world clinical settings remains an open challenge. Healthcare is inherently complex and dynamic, with diverse patient demographics, varying clinical protocols, and evolving medical knowledge [101]. AI models that perform well in research trials often struggle in real-world hospital deployments due to dataset shifts, model drift, and unexpected clinical edge cases [89]. To ensure that SLMs can be effectively scaled and validated, future research should prioritize prospective clinical trials and post-deployment monitoring frameworks that continuously assess model accuracy, safety, and usability in dynamic healthcare environments [102]. Additionally, adaptive learning techniques should be explored to allow SLMs to self-update in response to new clinical data, without compromising compliance with medical regulations. Finally, to further contextualize these challenges and the evolving readiness of SLMs in clinical environments, the following tables provide a complementary overview. Table 4 outlines the current deployment maturity of SLMs in various clinical use cases, whereas Table 5 maps specific challenges to emerging technical and organizational enablers.

Table 4. SLM Deployment Readiness Matrix

Dimension	**Current Status**	**Readiness Level**	**Examples / Notes**
Clinical Documentation Automation	Widely piloted	Moderate to High	Used in summarizing mental health records and automating structured note generation
Real-Time Clinical Decision Support	In early deployment	Emerging	Applied in mortality prediction, disease classification, and risk triage pipelines
Privacy-Preserving Local Inference	Technically feasible	High	Deployed on mobile devices using edge computing and federated learning strategies
Regulatory Approval Pathways	Lacking standardized criteria	Low	Alignment with clinical guidelines exists, but formal approvals remain limited
Explainability Tools for SLMs	Under research	Low	Efforts include modular reasoning, vocabulary tracing, and evidence-aware outputs
Cross-Institution Scalability	Experimentally validated	Emerging	Early trials with multilingual and federated frameworks ongoing
Multimodal SLM Integration	Very limited	Very Low	Few prototypes exist that integrate text with signals or imaging for diagnostics

Legend: High = Practically deployable | Emerging = Under active research/testing | Low = Needs foundational work | Very Low = Experimental or conceptual stage.

Table 5. Key Barriers and Emerging Enablers in SLM Deployment

Barrier	Underlying Cause	Potential Enabler
Lack of model interpretability	Transformer complexity (black-box reasoning)	Explainable AI techniques (e.g., attention maps, RAG)
Limited access to diverse datasets	Privacy laws, siloed health records	Synthetic data, federated learning, cross-site agreements
Clinician resistance to AI	Workflow burden, low trust	UX design, natural language interfaces, AI copilots
Regulation gaps for compact AI models	Lack of AI-specific clinical benchmarks	Development of SLM-specific validation frameworks
Computational scaling in diverse sites	Infrastructure disparity among hospitals	On-device inference, edge AI acceleration chips

In parallel, it is equally important to consider how different stakeholders are impacted by SLM adoption. Table 6 provides a stakeholder-centric view, summarizing the benefits and value SLMs can offer to key actors in the healthcare ecosystem including clinicians, patients, hospitals, and regulators.

Table 6. Stakeholder Impact Mapping for SLM Deployment in Healthcare

Stakeholder	Impact of SLM Integration
Clinicians	Reduces cognitive load via automated documentation, real-time triage support, and summarization tools.
Patients	Ensures data privacy and trust through on-device inference; improves access through empathetic chatbots and multilingual support.
Hospitals / IT Teams	Lowers deployment and maintenance costs; enables seamless integration with existing EHRs and supports infrastructure-light environments.
Regulators / Policymakers	Supports transparency and auditability via modular, explainable pipelines, aiding in compliance verification and ethical oversight.

These tables collectively highlight how SLMs are transitioning from experimental concepts to real-world viability, driven by targeted enablers that address current deployment bottlenecks and deliver measurable benefits to specific stakeholder groups.

Future Outlook: Toward Responsible, Adaptive, and Multimodal SLMs in Healthcare

The growing utility of Small Language Models (SLMs) in healthcare marks a significant shift toward scalable, domain-adaptable, and privacy-conscious AI systems. However, their sustained clinical impact will depend on coordinated progress across technical innovation, infrastructure readiness, regulatory alignment, and stakeholder trust. The next wave of SLM development is expected to be shaped by the following key trends:

- **Open-Source Medical SLMs and Benchmarking Initiatives:** The shift toward transparency, accessibility, and customization has spurred the development of open-source SLMs tailored for medical applications. However, for these models to be meaningfully compared and evaluated, there is a pressing need for standardized benchmarking platforms. These benchmarks should go beyond academic QA datasets and reflect clinically relevant tasks, such as diagnostic reasoning, longitudinal note summarization, symptom triage, and treatment planning. Also, to support global applicability, benchmarks must incorporate multilingual corpora, underrepresented healthcare settings, and noisy, real-world data that more closely resemble actual clinical workflows.
- **Instruction-Tuned SLMs for Clinician-AI Collaboration:** Instruction-tuning of SLMs using task-specific prompts aligned with real-world clinical responsibilities is emerging as a critical enabler of model usability and acceptance. Rather than requiring generic queries, these models should be optimized to interpret clinician intent through diverse interfaces, including voice,

text, and touch-based systems. Potential applications include generating discharge summaries, prioritizing lab results, flagging abnormalities, or responding to bedside queries during ward rounds. Emphasis on natural language interfaces, context-aware instructions, and structured outputs will be essential to reduce friction in human-AI collaboration at the point of care.

- **On-Device and Multimodal SLM Agents:** Future SLMs will be expected to handle not only clinical text but also structured data (e.g., vitals, EHR entries), sensor data (e.g., from wearables), and visual data (e.g., scans and reports). This shift will enable the emergence of multimodal agents that can process complex signals and deliver real-time, privacy-preserving intelligence directly on mobile or edge devices. In low-resource settings or remote monitoring scenarios, such models could support early intervention, chronic disease management, and mental health tracking without relying on continuous cloud connectivity, thus ensuring better data control and latency optimization.
- **Collaborative Data Ecosystems and Synthetic Data Generation:** Access to high-quality clinical datasets remains one of the primary bottlenecks in training and validating medical SLMs. Future success will depend on establishing multi-institutional consortia that facilitate collaborative data sharing through federated learning frameworks, protecting patient privacy and enabling model training across distributed datasets. Additionally, techniques for synthetic clinical data generation, including medical text simulation and controlled anonymization, must be advanced to mitigate data scarcity and reduce sampling biases. These approaches will be key in supporting SLM generalizability across hospitals, specialties, and patient demographics, in parallel with compliance to emerging regulations on data minimization and privacy-by-design.
- **Explainability and Clinician Trust Calibration:** As SLMs are embedded deeper into healthcare workflows, their explainability will become a non-negotiable requirement for clinical trust and adoption. Future work must integrate transparent reasoning modules such as attention heatmaps, counterfactual explanations, or SHAP-based visualizations to justify model decisions in lay clinical terms. This can enable actionable audit trails, promote shared decision-making, and assist with medico-legal documentation, particularly in high-stakes diagnostics or therapeutic planning
- **Continuous Learning and Model Adaptation in Clinical Environments:** Unlike static deployments, future SLMs must support lifelong learning via feedback loops from clinicians and evolving EHR data. Techniques like incremental fine-tuning, reinforcement learning with human feedback (RLHF), or prompt-based continual learning can ensure that SLMs remain up-to-date with emerging medical knowledge, drug interactions, or guidelines without catastrophic forgetting. Implementing secure, on-site adaptation pipelines will also reduce turnaround for domain shifts and local context sensitivity.
- **Socio-Technical Governance and Trustworthy AI Frameworks:** Beyond technical accuracy, the success of SLMs in healthcare will rely on embedding them within broader socio-technical governance structures. This includes AI assurance audits, bias assessments across demographics, and involving clinicians in participatory design. Development frameworks must align with evolving regulations (e.g., GDPR, HIPAA, EU AI Act, FDA's AI/ML Software as a Medical Device Guidance) to ensure safety, fairness, and accountability.
- **Expanded Evaluation Frameworks Differentiating SLMs from LLMs:** As SLMs move toward real-world deployment, future evaluation must extend beyond accuracy to include metrics attuned to their operational context. Unlike LLMs, which emphasize benchmark performance, SLMs demand assessment through latency, energy use, compliance score, explainability rank, and Clinical Alignment Score (CAS), reflecting adherence to medical guidelines. These metrics are vital for real-time, privacy-preserving, and resource-constrained

healthcare settings. Standardized leaderboards and tools capturing these dimensions will be key to enabling transparent, responsible adoption.

CONCLUSION

Small Language Models (SLMs) represent a critical shift in the evolution of AI in healthcare, offering an effective, ethical, and adaptable alternative to the dominant paradigm of Large Language Models (LLMs). As global health systems contend with challenges related to data privacy, infrastructure constraints, and the demand for domain-specific intelligence, SLMs are emerging as a viable solution. Their ability to operate on low-power devices, enable real-time inference, and align with regulatory frameworks makes them well-suited for deployment in diverse clinical environments, from tertiary hospitals to rural clinics. This review systematically examined the emergence and trajectory of SLMs through historical and technological lenses, identifying key enablers such as transformer miniaturization, model compression, domain-specific fine-tuning, federated learning, and Edge AI integration. We illustrated how SLMs are already transforming healthcare functions, including clinical documentation, diagnostic support, patient-facing conversational agents, and biomedical knowledge discovery, especially where responsiveness, data security, and cost-efficiency are critical.

To support broader adoption, several challenges must still be addressed: interpretability, robustness, clinical workflow integration, and compliance with evolving regulatory standards. The lack of standardized evaluation benchmarks and real-world validations continues to impede high-stakes deployment. Moving forward, we offer the following stakeholder-specific recommendations:

- *For developers:* Prioritize transparent and auditable model design, incorporate domain adaptation capabilities, and validate performance in real-world settings.
- *For health system leaders*: Invest in infrastructure for Edge AI deployment and staff training, especially in low-resource environments.
- *For regulators:* Establish context-aware benchmarks for SLM performance, including interpretability, equity, and data governance metrics.

SLMs are not merely smaller LLMs, they are foundational enablers of inclusive, efficient, and ethically aligned AI in healthcare. Beyond mapping their evolution, this review offers actionable recommendations. Theoretically, it clarifies the latent structure of the SLM landscape and illustrates the utility of narrative-evolutionary synthesis for interdisciplinary research. Practically, it highlights the relevance of SLMs in privacy-preserving, resource-constrained contexts offering scalable pathways for AI-driven care. These recommendations are intended to guide future model design, policy frameworks, and translational research within diverse healthcare domains.

Conflict of Interest Statement
The authors declare that there are no competing interests or financial relationships that could have influenced the work reported in this manuscript.

Data Availability Statement
This study did not generate or analyze any new datasets. All data referenced in the manuscript are publicly available in the cited literature. For reproducibility and transparency, any additional resources or supporting materials can be made available upon reasonable request to the corresponding author.

Use of Generative AI Statement

Generative AI tools were used to support the writing and editing of this manuscript. Specifically, AI-assisted drafting was employed for language refinement and structural coherence. All content has been critically reviewed and validated by the authors to ensure accuracy, originality, and scholarly integrity.

Funding Statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

1. Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, Aldairem A, Alrashed M, Bin Saleh K, Badreldin HA, Al Yami MS. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC medical education. 2023 Sep 22;23(1):689.
2. Singh GB, Kumar R, Ghosh RC, Punia A, Sharma N, Bhakuni P. Transforming Healthcare Systems With AI: A Deep Dive Into NLP Applications. InHarnessing AI and Machine Learning for Precision Wellness 2025 (pp. 111-130). IGI Global Scientific Publishing.
3. Raiaan MA, Mukta MS, Fatema K, Fahad NM, Sakib S, Mim MM, Ahmad J, Ali ME, Azam S. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. IEEE access. 2024 Feb 13;12:26839-74.
4. Yang R, Tan TF, Lu W, Thirunavukarasu AJ, Ting DS, Liu N. Large language models in health care: Development, applications, and challenges. Health Care Science. 2023 Aug;2(4):255-63.
5. Wang D, Zhang S. Large language models in medical and healthcare fields: applications, advances, and challenges. Artificial Intelligence Review. 2024 Sep 20;57(11):299.
6. Wang L, Wan Z, Ni C, Song Q, Li Y, Clayton E, Malin B, Yin Z. Applications and Concerns of ChatGPT and Other Conversational Large Language Models in Health Care: Systematic Review. Journal of Medical Internet Research. 2024 Nov 7;26:e22769.
7. Nassiri K, Akhloufi MA. Recent advances in large language models for healthcare. BioMedInformatics. 2024 Apr 16;4(2):1097-143.
8. Wang F, Zhang Z, Zhang X, Wu Z, Mo T, Lu Q, Wang W, Li R, Xu J, Tang X, He Q. A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness. arXiv preprint arXiv:2411.03350. 2024 Nov 4.
9. Zhang Q, Liu Z, Pan S. The Rise of Small Language Models. IEEE Intelligent Systems. 2025 Feb 20;40(1):30-7.
10. Nerella S, Bandyopadhyay S, Zhang J, Contreras M, Siegel S, Bumin A, Silva B, Sena J, Shickel B, Bihorac A, Khezeli K. Transformers and large language models in healthcare: A review. Artificial intelligence in medicine. 2024 Jun 5:102900.
11. Madan S, Lentzen M, Brandt J, Rueckert D, Hofmann-Apitius M, Fröhlich H. Transformer models in biomedicine. BMC Medical Informatics and Decision Making. 2024 Jul 29;24(1):214.
12. Thirunavukarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. Nature medicine. 2023 Aug;29(8):1930-40.
13. Wang L, Wan Z, Ni C, Song Q, Li Y, Clayton EW, Malin BA, Yin Z. A systematic review of chatgpt and other conversational large language models in healthcare. medRxiv. 2024 Apr 27.
14. Popay, J., Roberts, H., Sowden, A., Petticrew, M., Arai, L., Rodgers, M., Britten, N., Roen, K. and Duffy, S. (2006). Guidance on the conduct of narrative synthesis in systematic reviews. *A product from the ESRC methods programme Version. 2006 Apr 1;1(1):b92.*
15. Lyytinen, K., & Newman, M.. Explaining information systems change: a punctuated socio-technical change model. *European journal of information systems*, *2008 Dec 1;17(6):589-613.*
16. Shortliffe E, editor. Computer-based medical consultations: MYCIN. Elsevier; 2012 Dec 2.
17. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, Scales N, Tanwani A, Cole-Lewis H, Pfohl S, Payne P. Large language models encode clinical knowledge. arXiv preprint arXiv:2212.13138. 2022 Dec 26.

18. Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, Liu TY. BioGPT: generative pre-trained transformer for biomedical text generation and mining. Briefings in bioinformatics. 2022 Nov;23(6):bbac409.
19. Belciug, S., & Gorunescu, F. *Intelligent Decision Support Systems--A Journey to Smarter Healthcare* (pp. 130-137). Berlin/Heidelberg, Germany: Springer International Publishing, 2020
20. Selden C. Unified Medical Language System (UMLS): January 1986 Through December 1996: 280 Selected Citations. US Department of Health and Human Services, Public Health Service, National Institutes of Health, National Library of Medicine, Reference Section; 1997.
21. Spackman KA, Campbell KE, Côté RA. SNOMED RT: a reference terminology for health care. InProceedings of the AMIA annual fall symposium 1997 (p. 640).
22. Stearns MQ, Price C, Spackman KA, Wang AY. SNOMED clinical terms: overview of the development process and project status. InProceedings of the AMIA Symposium 2001 (p. 662).
23. Ahmed N, Dilmaç F, Alpkocak A. Classification of biomedical texts for cardiovascular diseases with deep neural network using a weighted feature representation method. InHealthcare 2020 Oct 10 (Vol. 8, No. 4, p. 392). MDPI.
24. Shao Y, Taylor S, Marshall N, Morioka C, Zeng-Treitler Q. Clinical text classification with word embedding features vs. bag-of-words features. In2018 IEEE International conference on big data (big data) 2018 Dec 10 (pp. 2874-2878). IEEE.
25. Youbi F, Settouti N. Analysis of machine learning and deep learning frameworks for opinion mining on drug reviews. The Computer Journal. 2022 Sep;65(9):2470-83.
26. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association. 2010 May 1;17(3):229-36.
27. Yang J, Liu Y, Qian M, Guan C, Yuan X.. Information extraction from electronic medical records using multitask recurrent neural network with contextual word embedding. Applied Sciences. 2019 Sep 4;9(18):3658.
28. Yadav S, Ramesh S, Saha S, Ekbal A. Relation extraction from biomedical and clinical text: Unified multitask learning framework. IEEE/ACM transactions on computational biology and bioinformatics. 2020 Aug 27;19(2):1105-16.
29. Chen YP, Chen YY, Lin JJ, Huang CH, Lai F. Modified bidirectional encoder representations from transformers extractive summarization model for hospital information systems based on character-level tokens (AlphaBERT): development and performance evaluation. JMIR medical informatics. 2020 Apr 29;8(4):e17787.
30. Beeksma M, Verberne S, van den Bosch A, Das E, Hendrickx I, Groenewoud S. Predicting life expectancy with a long short-term memory recurrent neural network using electronic medical records. BMC medical informatics and decision making. 2019 Dec;19:1-5.
31. Zhu R, Tu X, Huang J. Using deep learning based natural language processing techniques for clinical decision-making with EHRs. Deep learning techniques for biomedical and health informatics. 2019:257-95.
32. Lin YW, Zhou Y, Faghri F, Shaw MJ, Campbell RH. Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory. PloS one. 2019 Jul 8;14(7):e0218942.
33. Lipton ZC, Kale DC, Elkan C, Wetzel R. Learning to diagnose with LSTM recurrent neural networks. arXiv preprint arXiv:1511.03677. 2015 Nov 11.
34. Afzal M, Alam F, Malik KM, Malik GM. Clinical context–aware biomedical text summarization using deep neural network: model development and validation. Journal of medical Internet research. 2020 Oct 23;22(10):e19810.
35. Pivovarov R, Elhadad N. Automated methods for the summarization of electronic health records. Journal of the American Medical Informatics Association. 2015 Sep 1;22(5):938-47.
36. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020 Feb 15;36(4):1234-40.
37. Huang K, Altosaar J, Ranganath R. Clinicalbert: Modeling clinical notes and predicting hospital readmission. arXiv preprint arXiv:1904.05342. 2019 Apr 10.
38. Beltagy I, Lo K, Cohan A. SciBERT: A pretrained language model for scientific text. arXiv preprint arXiv:1903.10676. 2019 Mar 26.

39. Cho HN, Jun TJ, Kim YH, Kang H, Ahn I, Gwon H, Kim Y, Seo J, Choi H, Kim M, Han J. Task-Specific Transformer-Based Language Models in Health Care: Scoping Review. JMIR Medical Informatics. 2024 Nov 18;12:e49724.
40. Zhang A, Xing L, Zou J, Wu JC. Shifting machine learning for healthcare from development to deployment and from models to data. Nature biomedical engineering. 2022 Dec;6(12):1330-45.
41. Gupta M, Agrawal P. Compression of deep learning models for text: A survey. ACM Transactions on Knowledge Discovery from Data (TKDD). 2022 Jan 8;16(4):1-55.
42. Ren L, Wang T, Jia Z, Li F, Han H. A lightweight and adaptive knowledge distillation framework for remaining useful life prediction. IEEE Transactions on Industrial Informatics. 2022 Nov 28;19(8):9060-70.
43. Abadeer M. Assessment of DistilBERT performance on named entity recognition task for the detection of protected health information and medical concepts. InProceedings of the 3rd clinical natural language processing workshop 2020 Nov (pp. 158-167).
44. Wasserblat M, Pereg O, Izsak P. Exploring the boundaries of low-resource BERT distillation. InProceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing 2020 Nov (pp. 35-40).
45. Jiao X, Yin Y, Shang L, Jiang X, Chen X, Li L, Wang F, Liu Q. Tinybert: Distilling bert for natural language understanding. arXiv preprint arXiv:1909.10351. 2019 Sep 23.
46. Sarkar S, Babar MF, Hassan MM, Hasan M, Karmaker Santu SK. Processing Natural Language on Embedded Devices: How Well Do Transformer Models Perform?. InProceedings of the 15th ACM/SPEC International Conference on Performance Engineering 2024 May 7 (pp. 211-222).
47. Guo Z, Wang P, Wang Y, Yu S. Improving small language models on PubMedQA via Generative Data Augmentation. arXiv preprint arXiv:2305.07804. 2023 May 12.
48. Lu Z, Li X, Cai D, Yi R, Liu F, Zhang X, Lane ND, Xu M. Small language models: Survey, measurements, and insights. arXiv preprint arXiv:2409.15790. 2024 Sep 24.
49. Tangsrivimol, J.A., Darzidehkalani, E., Virk, H.U.H., Wang, Z., Egger, J., Wang, M., Hacking, S., Glicksberg, B.S., Strauss, M. and Krittanawong, C. (2025). Benefits, limits, and risks of ChatGPT in medicine. *Frontiers in Artificial Intelligence*, *2025 Jan 30;8:1518049.*
50. Tenajas, R., & Miraut, D. (2025). The Hidden Risk of AI Hallucinations in Medical Practice. https://www.annfammed.org/content/hidden-risk-ai-hallucinations-medical-practice
51. Shakhadri SA, KR K, Aralimatti R. SHAKTI: A 2.5 Billion Parameter Small Language Model Optimized for Edge AI and Low-Resource Environments. In IFIP International Conference on Artificial Intelligence Applications and Innovations 2025 (pp. 434-447). Springer, Cham.
52. Zhao X, Lu J, Deng C, Zheng C, Wang J, Chowdhury T, Yun L, Cui H, Xuchao Z, Zhao T, Panalkar A. Beyond One-Model-Fits-All: A Survey of Domain Specialization for Large Language Models. arXiv preprint arXiv. 2023 May;2305.
53. Taylor N, Ghose U, Rohanian O, Nouriborji M, Kormilitzin A, Clifton DA, Nevado-Holgado A. Efficiency at scale: Investigating the performance of diminutive language models in clinical tasks. Artificial Intelligence in Medicine. 2024 Nov 1;157:103002.
54. Popov RO, Karpenko NV, Gerasimov VV. Overview of small language models in practice. InCEUR Workshop Proceedings 2025 (pp. 164-182).
55. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Advances in neural information processing systems. 2017;30.
56. Koroteev MV. BERT: a review of applications in natural language processing and understanding. arXiv preprint arXiv:2103.11943. 2021 Mar 22.
57. Levine DM, Tuwani R, Kompa B, Varma A, Finlayson SG, Mehrotra A, Beam A. The diagnostic and triage accuracy of the GPT-3 artificial intelligence model. MedRxiv. 2023 Feb 1.
58. Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531. 2015 Mar 9.
59. Bondarenko Y, Nagel M, Blankevoort T. Quantizable transformers: Removing outliers by helping attention heads do nothing. Advances in Neural Information Processing Systems. 2023 Dec 15;36:75067-96.

60. Ai C, Yang H, Ding Y, Tang J, Guo F. Low rank matrix factorization algorithm based on multi-graph regularization for detecting drug-disease association. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2023 May 9;20(5):3033-43.
61. Rohanian O, Nouriborji M, Kouchaki S, Clifton DA. On the effectiveness of compact biomedical transformers. Bioinformatics. 2023 Mar 1;39(3):btad103.
62. Wang C, Li M, He J, Wang Z, Darzi E, Chen Z, Ye J, Li T, Su Y, Ke J, Qu K. A survey for large language models in biomedicine. arXiv preprint arXiv:2409.00133. 2024 Aug 29.
63. Guluzade A, Heiba N, Boukhers Z, Hamiti F, Polash JH, Mohamad Y, Velasco CA. ELMTEX: Fine-Tuning Large Language Models for Structured Clinical Information Extraction. A Case Study on Clinical Reports. arXiv preprint arXiv:2502.05638. 2025 Feb 8.
64. Konečný J, McMahan HB, Yu FX, Richtárik P, Suresh AT, Bacon D. Federated learning: Strategies for improving communication efficiency. arXiv preprint arXiv:1610.05492. 2016 Oct 18.
65. Liu M, Ho S, Wang M, Gao L, Jin Y, Zhang H. Federated learning meets natural language processing: A survey. arXiv preprint arXiv:2107.12603. 2021 Jul 27.
66. Dere G. Biomedical applications with using embedded systems. InData acquisition-recent advances and applications in biomedical engineering 2021 Feb 24. IntechOpen.
67. Zhang J, Sun K, Jagadeesh A, Falakaflaki P, Kayayan E, Tao G, Haghighat Ghahfarokhi M, Gupta D, Gupta A, Gupta V, Guo Y. The potential and pitfalls of using a large language model such as ChatGPT, GPT-4, or LLaMA as a clinical assistant. Journal of the American Medical Informatics Association. 2024 Sep;31(9):1884-91.
68. Bedi S, Liu Y, Orr-Ewing L, Dash D, Koyejo S, Callahan A, Fries JA, Wornow M, Swaminathan A, Lehmann LS, Hong HJ. Testing and evaluation of health care applications of large language models: a systematic review. JAMA. 2024 Oct 15.
69. Shool S, Adimi S, Saboori Amleshi R, Bitaraf E, Golpira R, Tara M. A systematic review of large language model (LLM) evaluations in clinical medicine. BMC Medical Informatics and Decision Making. 2025 Dec;25(1):1-1.
70. Guo Z, Wang P, Wang Y, Yu S. Improving small language models on PubMedQA via Generative Data Augmentation. arXiv preprint arXiv:2305.07804. 2023 May 12.
71. Hasan, M. J., Rahman, F., & Mohammed, N. (2023). Distilling the Knowledge of Clinical Outcome Predictions in Large Language Models for Resource Constrained Healthcare Systems. Available at SSRN 4591013.
72. Wang X, Chen N, Chen J, Wang Y, Zhen G, Zhang C, Wu X, Hu Y, Gao A, Wan X, Li H. Apollo: A Lightweight Multilingual Medical LLM towards Democratizing Medical AI to 6B People. arXiv preprint arXiv:2403.03640. 2024 Mar 6.
73. Kim, H., Hwang, H., Lee, J., Park, S., Kim, D., Lee, T., Yoon, C., Sohn, J., Park, J., Reykhart, O. and Fetherston, T. Small language models learn enhanced reasoning skills from medical textbooks. NPJ digital medicine. 2025 May 2;8(1):240.
74. Buhnila I, Sinha A, Constant M. Retrieve, generate, evaluate: A case study for medical paraphrases generation with small language models. arXiv preprint arXiv:2407.16565. 2024 Jul 23.
75. Wang, X., Dang, T., Kostakos, V., & Jia, H. Efficient and personalized mobile health event prediction via small language models. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking* 2024 Dec 4 (pp. 2353-2358).
76. Qu Y, Dai Y, Yu S, Tanikella P, Schrank T, Hackman T, Li D, Wu D. A Novel Compact LLM Framework for Local, High-Privacy EHR Data Applications. arXiv preprint arXiv:2412.02868. 2024 Dec 3.
77. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L, Chen W. Lora: Low-rank adaptation of large language models. ICLR. 2022 Apr 25;1(2):3.
78. Björkdahl, L., Pauli, O., Östman, J., Ceccobello, C., Lundell, S., & Kjellberg, M. Towards Holistic Disease Risk Prediction using Small Language Models. arXiv preprint arXiv:2408.06943. 2024 Aug 13.
79. Farrington M. DeviceBERT: Applied Transfer Learning With Targeted Annotations and Vocabulary Enrichment to Identify Medical Device and Component Terminology in FDA Recall Summaries. arXiv preprint arXiv:2406.05307. 2024 Jun 8.
80. Gwon H, Seo J, Park S, Kim YH, Jun TJ. Medical language model specialized in extracting cardiac knowledge. Scientific Reports. 2024 Nov 23;14(1):29059.

81. Griewing S, Lechner F, Gremke N, Lukac S, Janni W, Wallwiener M, Wagner U, Hirsch M, Kuhn S. Proof-of-concept study of a small language model chatbot for breast cancer decision support–a transparent, source-controlled, explainable and data-secure approach. Journal of Cancer Research and Clinical Oncology. 2024 Oct 9;150(10):451.
82. Diwakar, Raj D. DistilBERT-based Text Classification for Automated Diagnosis of Mental Health Conditions. InMicrobial Data Intelligence and Computational Techniques for Sustainable Computing 2024 Mar 1 (pp. 93-106). Singapore: Springer Nature Singapore.
83. Kumar A, Sharma A, Sangwan SR. Transformer-Based Abstractive Summarization for Depression Detection Literature for Enhanced Medical Insights. Available at SSRN 5029494. 2025 Jan 9.
84. Kumar A, Sangwan SR, Sharma A. Improving Depression Detection through Biomedical Entity Linking: A Hybrid Approach Using Embedding Models and Full-Text Search., 2024
85. Zong C, Wan J, Tang S, Zhang L. EvidenceMap: Learning Evidence Analysis to Unleash the Power of Small Language Models for Biomedical Question Answering. arXiv e-prints. 2025 Jan:arXiv-2501.
86. Magnini M, Aguzzi G, Montagna S. Open-source small language models for personal medical assistant chatbots. Intelligence-based medicine. 2025;11:1-9.
87. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608. 2017 Feb 28.
88. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence. 2019 May;1(5):206-15.
89. Chen IY, Joshi S, Ghassemi M. Treating health disparities with artificial intelligence. Nature medicine. 2020 Jan;26(1):16-7.
90. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019 Oct 25;366(6464):447-53.
91. Kaissis GA, Makowski MR, Rückert D, Braren RF. Secure, privacy-preserving and federated machine learning in medical imaging. Nature Machine Intelligence. 2020 Jun;2(6):305-11.
92. McMahan B, Moore E, Ramage D, Hampson S, y Arcas BA. Communication-efficient learning of deep networks from decentralized data. InArtificial intelligence and statistics 2017 Apr 10 (pp. 1273-1282). PMLR.
93. Kumar A, Sharma A, Ranjan R, Han L. FTL-Emo: Federated Transfer Learning for Privacy Preserved Biomarker-Based Automatic Emotion Recognition. InInternational Conference on Data Analytics & Management 2023 Jun 23 (pp. 449-460). Singapore: Springer Nature Singapore.
94. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nature medicine. 2019 Jan;25(1):44-56.
95. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. New England Journal of Medicine. 2019 Apr 4;380(14):1347-58.
96. Fowowe OO, Anthony OC. Advancing Healthcare Frameworks in the US: Artificial Intelligence Applications Across Operations and Administration International Journal of Computer Applications Technology and Research Volume 14–Issue 02, 82 – 98, 2025.
97. London AJ. Artificial intelligence and black-box medical decisions: accuracy versus explainability. Hastings Center Report. 2019 Jan;49(1):15-21.
98. Kiseleva A, Kotzinos D, De Hert P. Transparency of AI in healthcare as a multilayered system of accountabilities: between legal requirements and technical limitations. Frontiers in artificial intelligence. 2022 May 30;5:879603.
99. Wang F, Casalino LP, Khullar D. Deep learning in medicine—promise, progress, and challenges. JAMA internal medicine. 2019 Mar 1;179(3):293-4.
100. Adler-Milstein J, Aggarwal N, Ahmed M, Castner J, Evans BJ, Gonzalez AA, James CA, Lin S, Mandl KD, Matheny ME, Sendak MP. Meeting the moment: addressing barriers and facilitating clinical adoption of artificial intelligence in medical diagnosis. NAM perspectives. 2022 Sep 29;2022:10-31478.
101. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Briefings in bioinformatics. 2018 Nov;19(6):1236-46.
102. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nature biomedical engineering. 2018 Oct;2(10):719-31.