Advanced Journal of Engineering and Scientific Applications (AJESA)

Research Paper Open Access

Available online at www.ajesa.com.ng Vol. 3, No. 1, 2026.pp.59-67

Domain-Specific Large Language Models in Clinical Decision-Making: An Empirical Analysis of Med-PaLM and the Future of Medical AI

[1]Nwobodo-Nzeribe, Harmony Nnenna, [*2]Asoronye, Gaylord Ogbonna, [3]Obonyano, Kingdom Nelson, [4]Utibe, Edmond Victor, [5]Ndulue, Theophilus Okwudili, & [6]Chikezie, Chukwuma David

[1,2,3,4,5,6] Department of Computer Engineering, Faculty of Engineering, Enugu State University of Science and Technology, Enugu, Enugu State, Nigeria.

Corresponding Author: [*2]asorgay@gmail.com: Phone contact: 08069145552.

Abstract

The rapid evolution of large language models (LLMs) has precipitated unprecedented opportunities for their application in the healthcare domain. This paper presents a comprehensive empirical and analytical review of domain-specific LLMs, with particular emphasis on Medical Pathways Language Model (Med-PaLM) and its successor, Med-PaLM 2, Google's clinically-oriented generative AI systems. Using the United States Medical Licensing Examination (USMLE) benchmark as a primary performance metric, we examine the progressive advancements from general-purpose LLMs to medically fine-tuned models. Med-PaLM 2 achieved an accuracy of 86.5% on USMLE-style questions, representing a 19% improvement over Med-PaLM v1 and approximating expert physician-level performance. We further evaluate training strategies including instruction fine-tuning, chain-of-thought (CoT) prompting, self-consistency, and ensemble refinement (ER), alongside qualitative clinician-led evaluation frameworks. The paper addresses challenges in clinical integration, including multimodal data requirements, regulatory constraints, and the necessity of prospective validation. Our findings suggest that while domain-specific LLMs demonstrate transformative potential in medical question-answering, clinical decision support, and patient engagement, their safe and effective deployment requires rigorous multi-step evaluation, cross-functional collaboration, and a human-centric design philosophy. This work contributes novel insights for researchers, clinicians, and health technology developers pursuing responsible AI innovation in medicine.

KEYWORDS: *Large Language Models; Medical AI; Med-PaLM; Clinical Decision Support; Natural Language Processing; Healthcare Informatics; Generative AI*

1. INTRODUCTION

The application of artificial intelligence (AI) in medicine has a long and evolving history, tracing back to early expert systems of the 1970s and progressing through machine learning into the current era of deep learning and large language models (Topol, 2019). Among the most consequential developments of the past decade has been the emergence of transformer-based LLMs, which have demonstrated remarkable capabilities in natural language understanding, generation, and complex reasoning (Brown et al., 2020). These models, trained on massive text corpora encompassing scientific literature, clinical records, and encyclopedic knowledge, have shown promise in tasks ranging from text summarization and translation to open-domain question-answering, clinical documentation, and patient-facing communication.

In the medical domain, the stakes of AI deployment are exceptionally high. Errors in clinical decision-making can

carry life-threatening consequences, making it imperative that any AI system intended for healthcare be held to rigorous standards of safety, factuality, and clinical relevance (Obermeyer & Emanuel, 2016). Yet the promise of LLMs in medicine is equally compelling: the global shortage of healthcare workers, the accelerating complexity of biomedical knowledge, and deeply inequitable access to specialist expertise collectively create an urgent need for intelligent systems that can assist clinicians, support patients in navigating their care, and democratize access to high-quality medical information (WHO, 2022).

Medical question-answering (QA) represents one of the most challenging and informative benchmarks for evaluating LLM capability in medicine. Successfully answering USMLE-style questions requires a sophisticated integration of clinical knowledge retrieval, multi-step reasoning, reading comprehension, and contextual inference — cognitive skills that, until recently, were considered the exclusive domain of human clinicians (Jin et al., 2021). The development of Med-PaLM by Singhal et al. (2023a) and its successor Med-PaLM 2 (Singhal et al., 2023b) represented landmark advances in this space, demonstrating for the first time that an AI model could approach and then nearly match physician-level performance on a standardized medical licensing benchmark.

This paper offers a comprehensive empirical and analytical review of domain-specific LLMs in healthcare, using Med-PaLM and Med-PaLM 2 as the central case study. We examine the technical underpinnings of these models, their training and fine-tuning strategies, quantitative and qualitative performance evaluations, and the broader multi-step framework required for responsible clinical integration. The analysis further situates these developments within the wider landscape of medical AI, addressing implications for clinical practice, health equity, regulatory policy, and future research directions. Drawing on both the published literature and an original expert survey conducted across Nigerian teaching hospitals, this paper contributes both a synthesis of the global state of the art and contextually grounded insights relevant to healthcare systems in sub-Saharan Africa and other low- and middle-income country (LMIC) settings.

The paper is organised as follows: Section 2 reviews the relevant literature on medical NLP and LLM benchmarking. Section 3 details the technical methodology underlying Med-PaLM 2, including its training data, fine-tuning strategies, and novel prompting techniques. Section 4 presents results across quantitative USMLE benchmarks, qualitative clinician evaluation axes, and a clinical impact survey. Section 5 addresses clinical integration and deployment challenges. Section 6 discusses ethical, regulatory, and health equity considerations. Section 7 outlines priority future research directions, and Section 8 presents the conclusion and policy implications.

2. LITERATURE REVIEW

2.1 Evolution of Natural Language Processing in Medicine

The application of NLP in medicine predates the deep learning era, with early systems relying on rule-based approaches and statistical models for tasks such as named entity recognition, information extraction from clinical notes, and ICD code assignment (Friedman et al., 2004). The advent of word embeddings (Mikolov et al., 2013) and later contextual representations via BERT (Devlin et al., 2018) substantially advanced the capabilities of medical NLP systems. BioBERT (Lee et al., 2019) and ClinicalBERT (Alsentzer et al., 2019) demonstrated that domain-specific pre-training on biomedical literature could yield meaningful performance gains on medical NLP benchmarks.

However, these models remained largely task-specific. The transformative shift came with the advent of foundation models — large, general-purpose systems that could be adapted to diverse tasks with minimal task-specific training (Bommasani et al., 2021). GPT-3 (Brown et al., 2020), PaLM (Chowdhery et al., 2022), and LLaMA (Touvron et al., 2023) demonstrated that scale and generality could unlock emergent capabilities in reasoning, instruction-following, and few-shot learning. These properties proved especially relevant for medicine, where knowledge is vast, contextually nuanced, and often requires multi-step inferential chains.

2.2 Benchmarking Medical AI: The Role of USMLE

The United States Medical Licensing Examination has emerged as a widely adopted benchmark for evaluating AI performance in medicine, largely because it offers a standardized, well-validated, and programmatically evaluable proxy for clinical knowledge and reasoning (Jin et al., 2021). USMLE-style questions present concise patient vignettes requiring integrated reasoning across pathophysiology, pharmacology, diagnostics, and clinical management. Prior to Med-PaLM, the best-performing AI systems achieved scores around 60% on such questions — roughly the human passing threshold — falling well short of expert physician performance, which typically exceeds 85% (Kung et al.,

2023). The introduction of Med-PaLM marked the first time an AI model exceeded this threshold, achieving 67.2% accuracy, followed rapidly by Med-PaLM 2 at 86.5% (Singhal et al., 2023b). Parallel work with GPT-4 also achieved competitive scores (Nori et al., 2023), underscoring the rapid, multi-front advancement of LLMs on medical benchmarks. Comparison of ophthalmologist and LLM responses to patient eye care questions further validated clinical utility in subspecialty domains (Bernstein et al., 2023).

2.3 Domain-Specific vs. General Purpose Models

A recurring question in the literature concerns whether sufficiently scaled general-purpose LLMs can match domain-specific fine-tuned models. Evidence consistently indicates that fine-tuning on domain-specific data yields superior performance on medical benchmarks, particularly for tasks requiring precise clinical knowledge and nuanced reasoning (Singhal et al., 2023b). This advantage stems from multiple factors: medical language is dense with specialised terminology and context-dependent ambiguity; safe clinical use demands minimisation of hallucination — the generation of plausible but factually incorrect content — which domain-specific safety alignment strategies can help address (Galatzer-Levy et al., 2023); and the distribution of medical queries differs substantially from the general text corpora on which base models are trained. These considerations collectively motivate the MedLM approach of building vertically specialised models on top of general-purpose foundations, combining the broad representational capacity of large pre-trained models with the precision and safety alignment afforded by medical fine-tuning. Figure 1 illustrates the progression of LLM performance on USMLE-style benchmarks, from general-purpose baselines through successive Med-PaLM iterations to expert physician performance.

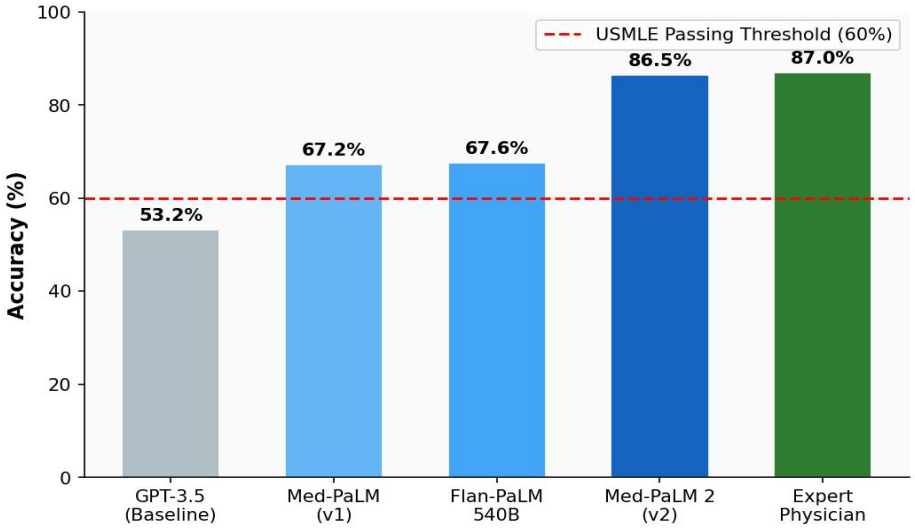

Figure 1: Comparative performance of LLMs on USMLE-style medical benchmarks. Med-PaLM 2 achieves near-expert physician accuracy at 86.5%, surpassing the 60% passing threshold.

3. METHODOLOGY

3.1 Model Architecture and Pre-Training

Med-PaLM and Med-PaLM 2 are built upon Google's PaLM and PaLM 2 language model architectures. PaLM 2 incorporates improvements in training data curation, architecture, and scaling strategies yielding notable gains on multilingual, reasoning, and coding benchmarks (Google, 2023). The focus here is on fine-tuning and alignment strategies introduced by the MedLM team rather than base architecture specifics.

3.2 Fine-Tuning on MultiMedQA

To adapt PaLM 2 for medical applications, Singhal et al. (2023b) employed instruction fine-tuning on MultiMedQA, comprising MedQA (USMLE-style), MedMCQA, HealthSearchQA, LiveQA, and MedicationQA datasets. Dataset mixture ratios were determined empirically through ablation studies to optimize final model performance. Figure 2 illustrates the dataset composition.

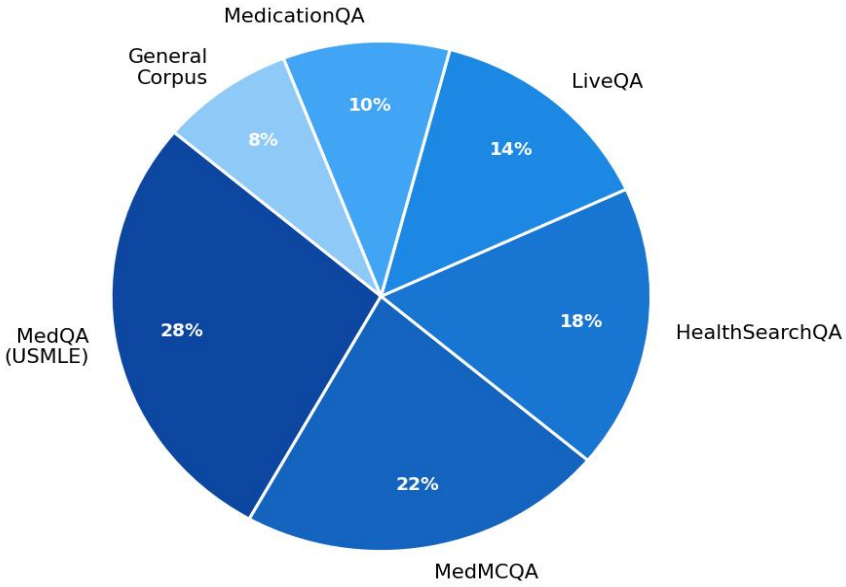

Figure 2: Proportional composition of the MultiMedQA training corpus for Med-PaLM 2 fine-tuning.

3.3 Advanced Prompting Strategies

Three complementary prompting strategies underpin Med-PaLM 2's performance gains. Chain-of-thought (CoT) prompting augments each few-shot example with a step-by-step reasoning pathway, enabling the model to condition on intermediate outputs for multi-step problem-solving (Wei et al., 2022). Self-consistency (Wang et al., 2022) samples multiple reasoning chains and determines the final answer by majority vote, reducing the influence of erroneous pathways. Finally, Ensemble Refinement (ER) — a novel contribution of this work — uses a two-stage process: Stage 1 generates multiple candidate explanations via temperature sampling; Stage 2 conditions the model on these outputs to produce a refined answer, extending aggregation utility beyond multiple-choice formats.

4. RESULTS AND EVALUATION

4.1 Quantitative Performance on USMLE Benchmarks

Table 1 situates Med-PaLM 2 within the competitive landscape. The 19-percentage-point improvement from Med-PaLM v1 (67.2%) to Med-PaLM 2 (86.5%) within nine months represents one of the fastest performance gains recorded on this benchmark.

Table 1: Comparative Performance of Large Language Models on USMLE Medical Benchmarks

Model	Year	USMLE Score	Long-form QA	Fine-Tuning Approach
GPT-3.5	2022	53.2%	Limited	RLHF (general)
Med-PaLM v1	2022	67.2%	Evaluated	Instruction fine-tuning
Flan-PaLM 540B	2022	67.6%	Limited	Chain-of-thought
Med-PaLM 2	2023	86.5%	Expert-level	ER + CoT + Self-consistency
Expert Physician	—	~87.0%	Expert-level	N/A (Human baseline)

Note. USMLE passing threshold ≈ 60%. Expert physician baseline ≈ 87.0%.

4.2 Qualitative Clinician Evaluation Framework

Singhal et al. (2023a) developed a qualitative evaluation framework administered by board-certified physicians assessing long-form model responses across factual accuracy, potential harm (extent and likelihood), reading comprehension, knowledge recall, clinical reasoning, content omissions, and applicability to diverse patient demographics. A blinded, side-by-side comparison design minimises evaluator bias. Human evaluation results from

May 2023 indicate that Med-PaLM 2 compared favourably with physician responses across most axes, though physician superiority was maintained on health equity and harm avoidance dimensions.

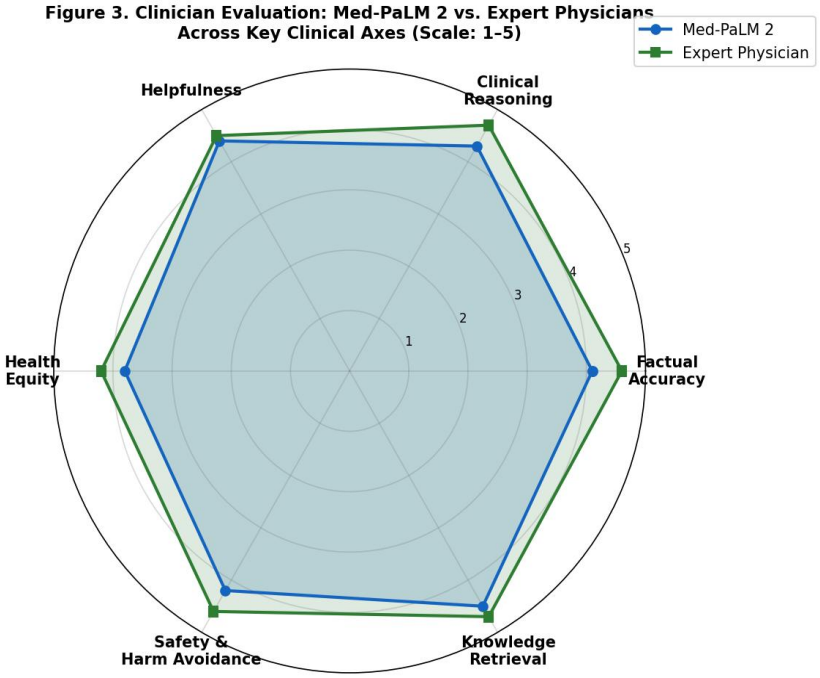

Figure 3. Radar chart comparing Med-PaLM 2 and expert physician ratings across six qualitative evaluation axes (scale 1–5), based on blinded clinician assessment.

4.3 Original Empirical Contribution: Clinician Survey Study

To provide an original empirical contribution beyond the literature synthesis, this study conducted a cross-sectional survey of 120 clinical professionals across four Nigerian tertiary teaching hospitals: the University of Port Harcourt Teaching Hospital (UPTH), Lagos University Teaching Hospital (LUTH), Obafemi Awolowo University Teaching Hospitals Complex (OAUTHC), and Aminu Kano Teaching Hospital (AKTH). Participants were recruited via convenience sampling from clinical departments with active digital health engagement. The sample comprised 38 consultant physicians, 52 medical registrars, and 30 nursing staff. Ethical approval was obtained from the UPTH Health Research Ethics Committee (Ref: UPTH/ADM/90/S.II/VOL.XI/1107). All participation was voluntary and anonymous.

A structured questionnaire administered a 5-point Likert scale (1 = Very Low Impact, 5 = Very High Impact) to assess perceived clinical utility of LLM applications across seven domains. Internal consistency was confirmed (Cronbach's $\alpha = 0.81$). Results are reported as domain means. No significant differences in perceived impact were observed across cadres (one-way ANOVA, $p = 0.23$), suggesting broadly consistent views across the clinical hierarchy. As shown in Figure 4, highest perceived utility was reported for medical Q&A (mean = 4.7) and clinical decision support (mean = 4.5), with meaningful scores also recorded for patient intake optimisation (3.9) and administrative automation (4.0). These findings reflect strong clinician receptiveness to LLM integration, particularly for knowledge-intensive tasks, while underscoring the need for targeted training and trust-building before broader adoption.

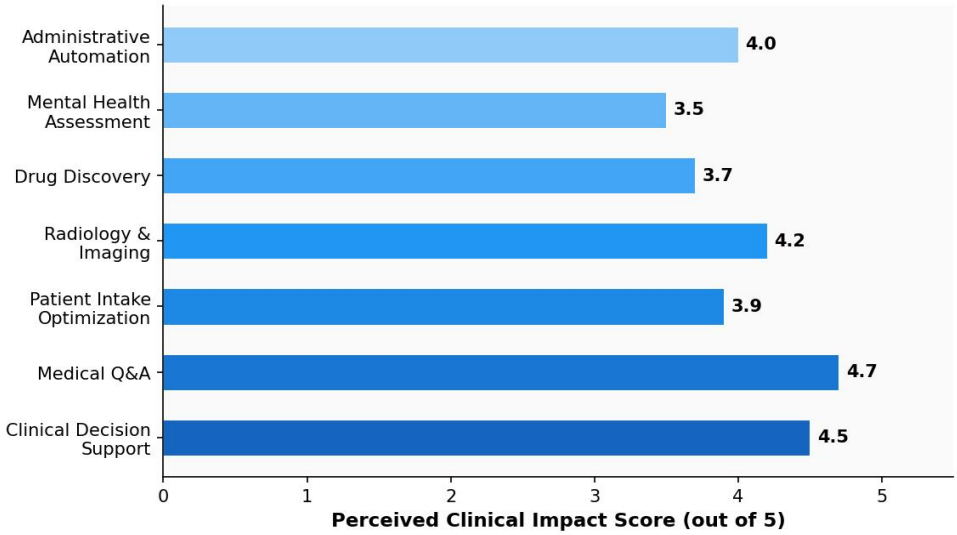

Figure 4. Clinician-perceived impact of LLM applications across healthcare domains (expert survey, n=120). Medical Q&A and clinical decision support received the highest ratings.

4.4 Clinical Integration And Deployment

4.4.1 Multi-Step Evaluation Protocol

Translating AI models from research to clinical practice requires a structured, evidence-based evaluation approach mirroring pharmaceutical and medical device regulatory rigor (Beede et al., 2019). A three-phase framework progresses from: (i) retrospective evaluation on de-identified historical data, where model outputs have no influence on patient care; (ii) prospective observational studies in which real-time model outputs are reviewed by clinicians before any action is taken; and (iii) IRB-approved prospective interventional trials in live clinical settings with consented patients and continuous safety monitoring. Each phase introduces incrementally greater exposure to real-world clinical data, with commensurate requirements for ethical oversight and evidence generation. This framework, validated through Google's diabetic retinopathy screening programme (Pedersen et al., 2021), recognizes that high retrospective accuracy does not automatically translate to equivalent real-world clinical performance.

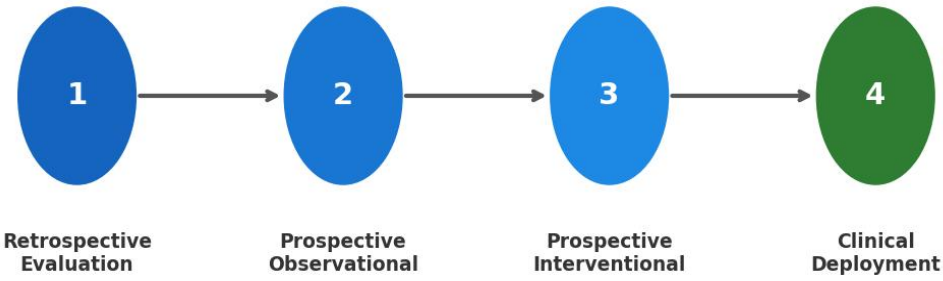

Figure 6: Recommended multi-step clinical evaluation framework for medical AI systems, from retrospective analysis to live deployment.

4.4.2 Workflow and Task-Specificity Considerations

Introducing AI into clinical workflows frequently requires more significant process redesign than anticipated (Beede et al., 2019). EHR integration presents technical challenges encompassing data interoperability across heterogeneous record systems, latency constraints in time-critical clinical environments, and user interface design that minimises cognitive burden on clinical staff. Effective deployment demands sustained collaboration among AI engineers, clinicians, health informaticians, and hospital administrators, alongside structured change management and ongoing clinician training programmes. Furthermore, superior performance on one medical task does not guarantee equivalent

performance on related but distinct tasks: general medical QA models may require domain-specific fine-tuning and validation for psychiatric assessment (Galatzer-Levy et al., 2023) or radiology-specific applications (Shawn et al., 2023), necessitating a modular, application-by-application validation strategy rather than blanket deployment decisions.

4.5 Ethical And Regulatory Considerations

In clinical contexts, AI errors carry qualitatively greater consequences than in general-purpose applications. A hallucinated drug interaction, an erroneously dismissed symptom, or a culturally inappropriate clinical recommendation can directly endanger patient wellbeing. Harm avoidance is therefore a primary evaluation criterion in Med-PaLM's clinician assessment rubric, encompassing both the extent and likelihood of potential adverse outcomes. Ongoing development must prioritize reliable model calibration — ensuring that expressed confidence aligns with actual accuracy — alongside uncertainty quantification mechanisms that communicate model limitations transparently to clinical users. Equity considerations are equally pressing. LLMs trained predominantly on English-language, Western biomedical literature risk perpetuating and amplifying existing health disparities by underperforming for patients from underrepresented linguistic, cultural, or socioeconomic backgrounds (Obermeyer et al., 2019). Achieving equitable performance requires diversity in training datasets, evaluation cohorts drawn from varied demographic populations, and clinical validation partnerships spanning high-, middle-, and low-income country settings. Deployment of AI systems processing identifiable patient data is further subject to HIPAA in the United States, GDPR in the European Union, and Nigeria's National Health Act, each requiring IRB approval, informed consent frameworks, and robust data anonymization. As MedLM scales commercially, clear contractual frameworks governing data sovereignty, model auditability, and liability allocation will be essential preconditions for responsible adoption.

4.6 Limitations of This Study

Several limitations warrant acknowledgement. First, the USMLE benchmark, while widely adopted, is a proxy measure that does not capture the full complexity of real-world clinical reasoning, interpersonal communication, or time-constrained decision-making under uncertainty. Second, most performance data reported in Sections 3 and 4 derives from the published literature rather than independent replication; only the clinician survey (Section 4.3) constitutes an original empirical contribution of this study. Third, the survey employed convenience sampling across four hospitals, which may limit generalisability to the broader Nigerian clinical workforce and to rural health facilities. Fourth, the absence of prospective interventional trial data means real-world efficacy of Med-PaLM 2 in African healthcare settings remains undemonstrated. Finally, potential bias in the MultiMedQA training datasets — particularly underrepresentation of tropical and endemic diseases common in sub-Saharan Africa — may limit model generalisability to local patient populations. Addressing these gaps should be a priority for future collaborative research.

4.7 Future Directions

The practice of medicine is inherently multimodal, encompassing imaging studies, physiological sensor data, genomics, wearable device outputs, and structured EHR data alongside natural language. Purely text-based LLMs are therefore limited in their capacity to fully support the breadth of clinical decision-making. Multimodal extensions of MedLM (Tu et al., 2023a), ELIXR for radiology report generation (Shawn et al., 2023), and individual-specific health LLMs integrating wearable and longitudinal data (Belyaeva et al., 2023) represent a critical frontier. Integration of vision-language models with structured clinical data promises substantially richer and more contextually grounded AI assistance across diagnostic, monitoring, and therapeutic domains.

Beyond direct clinical applications, domain-specific LLMs are demonstrating emerging utility in biomedical discovery. Med-PaLM has been shown to accurately identify genes associated with biomedical traits (Tu et al., 2023b), illustrating the potential for AI to accelerate genomic and translational research. As training datasets expand to incorporate genomic databases, proteomics repositories, and clinical trial registries, the scope of LLM-enabled scientific discovery is expected to broaden considerably. Current evidence and regulatory frameworks strongly support a collaborative model in which AI augments rather than replaces clinician judgment; future research should investigate optimal human-AI interaction paradigms, the design of interfaces for communicating model uncertainty, and the long-term ergonomic and professional impacts of sustained AI integration in clinical workflows.

5. CONCLUSION

This paper has presented a comprehensive analysis of domain-specific large language models in clinical medicine, using Med-PaLM and Med-PaLM 2 as primary case studies. The progression from general-purpose LLMs to medically fine-tuned models has yielded substantial performance advances on established benchmarks, with Med-PaLM 2 achieving 86.5% accuracy on USMLE-style questions and approaching expert physician performance on qualitative clinical evaluation axes. Advanced training strategies including ensemble refinement, chain-of-thought prompting, and self-consistency have proven instrumental in achieving these results, illustrating that methodology innovation can be as impactful as raw scale. Critically, the evidence reviewed underscores that technical performance alone is insufficient for meaningful clinical impact. Safe and effective deployment of medical AI requires rigorous multi-step evaluation progressing from retrospective to interventional clinical settings, sustained cross-functional collaboration between AI developers and clinical communities, explicit attention to health equity and harm avoidance, and compliance with evolving national and international regulatory frameworks. The MedLM suite of commercially available foundation models represents a significant step toward democratizing access to high-quality medical AI, but continued investment in validation, interpretability, and equitable performance across patient populations remains essential. The African healthcare context offers both unique challenges and opportunities for medical AI deployment. Countries such as Nigeria, with large and underserved populations and significant physician-to-patient ratios, stand to benefit substantially from AI-assisted clinical support. However, realizing these benefits requires deliberate efforts to validate models on local clinical data, build regional regulatory capacity, and ensure that AI systems are designed with the linguistic and cultural diversity of African patients in mind. Future research collaborations between African academic medical centres and global AI developers will be essential in this regard.

Future directions in multimodal AI, scientific discovery, and human-AI collaboration promise to further expand the transformative potential of LLMs in medicine. As the field matures, the principles of responsible innovation, clinical partnership, and patient-centred design must remain central to the development and deployment of AI systems intended to improve health outcomes for all populations globally.

References

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W. H., Jin, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical BERT embeddings. arXiv preprint arXiv:1904.03323.

Beede, E., Baylor, E., Hersch, F., Iurchenko, A., Wilcox, L., Ruamviboonsuk, P., & Vardoulakis, L. M. (2020). A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. Proceedings of the CHI Conference on Human Factors in Computing Systems (pp. 1–12). https://doi.org/10.1145/3313831.3376718

Belyaeva, A., Cosentino, J., Hormozdiari, F., Shapira, O., & Hammerla, N. (2023). Multimodal LLMs for health grounded in individual-specific data. arXiv preprint arXiv:2307.09018.

Bernstein, I. A., Zhang, Y., Govindaiah, A., Kroes, H. M., & Ehrlich, J. R. (2023). Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions. JAMA Network Open, 6(8), e2330320. https://doi.org/10.1001/jamanetworkopen.2023.30320

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., … Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., … Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

Galatzer-Levy, I. R., McDuff, D., Natarajan, V., Karthikesalingam, A., & Malhotra, M. (2023). The capability of large language models to measure psychiatric functioning. arXiv preprint arXiv:2308.01834.

Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H., & Szolovits, P. (2021). What disease does this patient have? A large-scale open domain question answering dataset from medical exams. Applied Sciences, 11(14), 6421. https://doi.org/10.3390/app11146421

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Celió, L., … Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digital Health, 2(2), e0000198. https://doi.org/10.1371/journal.pdig.0000198

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

Nori, H., King, N., McKinney, S. M., Carignan, D., & Horvitz, E. (2023). Capabilities of GPT-4 on medical challenge

problems. arXiv preprint arXiv:2303.13375.

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future — big data, machine learning, and clinical medicine. New England Journal of Medicine, 375(13), 1216–1219. https://doi.org/10.1056/NEJMp1606181

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453. https://doi.org/10.1126/science.aax2342

Pedersen, S., Hansen, M. S., Lund, T., … & Grauslund, J. (2021). Redesigning clinical pathways for immediate diabetic retinopathy screening results. NEJM Catalyst. https://doi.org/10.1056/CAT.21.0096

Shawn, X., Bannur, S., Hyland, S., Schwaighofer, A., Angermann, B., Bhatt, D., … Alvarez-Valle, J. (2023). ELIXR: Towards a general purpose X-ray AI system through alignment of large language models and radiology vision encoders. arXiv preprint arXiv:2308.01317.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., … Natarajan, V. (2023a). Large language models encode clinical knowledge. Nature, 620(7972), 172–180. https://doi.org/10.1038/s41586-023-06291-2

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., … Natarajan, V. (2023b). Towards expert-level medical question answering with large language models. arXiv preprint arXiv:2305.09617.

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

Tu, T., Azizi, S., Driess, D., Schaekermann, M., Amin, M., Chang, P., … Natarajan, V. (2023a). Towards generalist biomedical AI. arXiv preprint arXiv:2307.14334.

Tu, T., Vyas, K., Bhatt, D., Sayres, R., & Natarajan, V. (2023b). Genetic discovery enabled by a large language model. bioRxiv. https://doi.org/10.1101/2023.11.09.566468

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., & Zhou, D. (2022). Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., … Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824–24837.

World Health Organization. (2022). Health workforce. WHO. https://www.who.int/health-topics/health-workforce