Survey Report

Large Language Models-LLMs in Healthcare

Introduction

A structured literature review was conducted using Google Scholar search engine and Research Rabbit.

Keywords included:

- LLMs in Healthcare and Medical Q&A
- LLMs in Healthcare,
- Comparative evaluation of LLMs/ Comparative analysis of LLMs in medicine
- Limitations of LLMs in healthcare
- Systematic review of LLMs in medicine
- Boolean search: *("large language models" OR "LLMs" OR "transformer models") AND ("healthcare" OR "medicine") AND ("comparative analysis" OR "evaluation" OR "benchmarking")*

The below table 1 shows the medical/healthcare domain specific LLMs research papers survey. The models cover a range of biomedical NLP tasks including named entity recognition (NER), relation extraction (RE), question answering (QA), and clinical prediction (e.g., hospital readmission, diagnosis). They differ in scale, methodology, and evaluation strategies, yet collectively contribute to improving domain-specific performance in healthcare language modeling.

Model	**Multimodal**	**Objective/Problem Statement**	**Methodology**	**Dataset(s)**	**Results**	**Gap/Limitation/Research Gap**
BioBERT [1]	No	How BERT can be improved for biomedical corpora.	It has same architecture has BERT (bidirectional encoder only) and uses WordPiece tokenization. It applies standard full fine-tuning with task-specific single output layers: token-level classification for NER (BIO2 tagging), [CLS]-based classification for relation extraction, and start/end token span prediction for question answering.	It is trained on text corpora: Books corpus, English Wikipedia, PubMed Abstracts, PMC Full-text articles.	It outperformed the BERT mode in different tasks: biomedical NER (0.62% F1 score improvement), biomedical RE (2.80% F1 score improvement) and biomedical QA (12.24% MRR improvement).	• Not multimodal • Due to resource limitations, only the BERTBASE architecture was used. • BioBERT reuses BERT's general-domain vocabulary, which may limit its ability to represent biomedical-specific terms effectively.
ClinicalBER [2]	No	Clinical notes contain more information (medications, and lab values) but these have been underused as compared to	ClincalBERT uses base model; BERT (bidirectional encoder only). BERT is fine tuned for predicting the hospital readmission. Admissions were split into five folds in order to ensure independence b/w pre-training & fine tuning; 4 folds for pre-	Clinical notes from MIMIC-III.	ClinicalBERT is compared with 03 base models: Bag of words, Bi-LSTM, BERT. `Each model is evaluated using 03 metrics:	• Not multimodal • Poor performance on conversational medical QA or complex reasoning tasks. • It is trained on fixed dataset therefore

		structure data, since the clinical notes are sparse and high dimensional. This research paper use clinical notes to find thirty days hospital readmission at different times of admission such as discharge and early stage.	training and fine tuning whereas 1 fold for testing during fine tuning. Fine tuning details are: 3 epochs Batch size 56 Learning arte 2×10^{-5} For the binary output: A 3-layer feedforward with architecture: 768->2048->76-> 1, using sigmoid activation function.		1. Area under the receiver operating characteristic curve (AUROC) 2. Area under the precision-recall curve (AUPRC) 3. Recall at precision of 80% (RP80). ClinicalBERT outperforms baseline models with up to 3 to 4% higher AUROC and 2 to 3% higher AUPRC, indicating better prediction accuracy and precision. It also attains the highest RP80, indicating more effective detection of patients at high risk of early hospital readmission.	cannot adopt newly emerging medical knowledge.
Med-BERT [3]	No	To evaluate whether pre-trained contextualized embeddings on large-scale structured EHR data can benefit downstream disease prediction tasks, especially in low-data scenarios.	• Adapted the BERT model for structured EHR data using: • Code embeddings (diagnosis codes) • Visit embeddings (to differentiate visits) • Serialization embeddings (code order in visit) • Pre-training tasks: • Masked Language Model (MLM) • Prolonged Length of Stay (LOS) prediction. • Fine-tuned on two disease prediction tasks across three cohorts (heart failure & pancreatic cancer)	Pre-training: 1. Cerner Health Facts® (Large-scale structured EHR dataset) having 28M+ Cohort size. Evaluation: 03 cohorts; 1. DHF-Cerner (diseases related heart failure) having 672K+ Cohort size. 2. PaCa-Cerner(pancreatic cancer) having 29K+ cohort size. 3. PaCa-Truven(pancreatic cancer) having 42K+ cohort size.	- Improved AUC by 2.02–7.12% across GRU, Bi-GRU, and RETAIN models - Achieved strong performance with very small training sets (300–500 samples) - Med-BERT alone (FFL only) outperformed non-pretrained baselines - Showed generalization on unseen data (PaCa-Truven)	• Not multimodal • Only diagnosis information in ICD-9/10 format was used. • It did not incorporate the actual time intervals between patient visits; instead, it relied on relative visit positions, which may result in a loss of important temporal information. • Did not thoroughly explore the ordering of concepts within each visit, relying instead on potentially unreliable code priority rankings. • Future work will involve developing improved pre-training tasks, expanding evaluation beyond disease prediction, and incorporating additional data types such as time, medications,

						procedures, and lab tests.
BioGPT [4]	No	To develop a domain-specific language model for biomedical text generation and understanding, improving performance on biomedical NLP tasks.	BioGPT is based on GPT-2 medium (24 layers, 1024 hidden, 16 heads, 347M params) and pretrained from scratch using causal language modeling on 15M PubMed abstracts. Fine-tuned per task: token classification for NER, relation classification using entity markers, and multiple-choice classification for QA tasks.	Pretraining: 15M PubMed abstracts (4.5B words); Evaluation: PubMedQA, BioASQ, MedMCQA, BLURB benchmark tasks (NER, RE, QA, etc.)	They have achieved 44.98%, 38.42% and 40.76% F1 score on BC5CDR, KD-DTI and DDI end-to-end RE(elation extraction) tasks respectively, and 78.2% accuracy on PubMedQA, therefore BioGPT outperformed previous models (including BioBERT and PubMedBERT) on biomedical QA, NER, and relation extraction tasks.	• Not multimodal • BioGPT was pretrained only on PubMed abstracts and not full-text biomedical articles. • Although model is fine-tuned on diverse downstream tasks but pretrained on abstracts only not the fill articles. • BioGPT's 1024-token context limit, inherited from GPT-2, restricts its ability to process long biomedical documents or clinical records effectively.
BioMedLM [5]	No	To introduce BioMedLM, a 2.7B parameter GPT-style biomedical language model trained on PubMed abstracts and full articles, aimed at exploring the potential of domain-specialized, smaller models as competitive and transparent alternatives to large, opaque biomedical LLMs.	BioMedLM is a 2.7B parameter GPT-2 style autoregressive Transformer trained from scratch on PubMed abstracts and full-text articles using causal language modeling. The model uses domain-specific tokenization (SPT). It was evaluated without task-specific fine-tuning on medical QA datasets (e.g., MedQA, PubMedQA) to assess zero-shot and few-shot capabilities.	Pretraining: Pile-Med (800 million tokens), consisting of PubMed Central, PubMed abstracts, and other biomedical sources. Evaluation: MedQA (USMLE), PubMedQA, and MedMCQA for medical QA tasks.	BioMedLM demonstrates strong performance across biomedical QA tasks, outperforming larger models like Galactica on MedQA (54.1 vs. 44.4) and BioASQ (95.7 vs. 94.3), and closely matching Flan-PaLM on MedMCQA (57.38 vs. 57.6), despite being 200× smaller. It also provides generally accurate long-form answers, though some hallucinations—especially numerical inaccuracies—were observed in generated responses.	Not multimodal BioMedLM struggles with hallucination in generated answers, especially around numerical facts; it lacks real-world clinical evaluation (e.g., EHRs), and future work includes adopting multi-phase fine-tuning strategies (as used in BioGPT) to enhance task-specific performance.
(LinkBERT) BioLinkBERT [6]	No	Existing LMs like BERT are trained on isolated documents and thus cannot effectively learn or reason over knowledge that spans across documents. BioLinkBERT addresses this by incorporating	Pretraining on a citation graph constructed from PubMed abstracts. BioLinkBERT uses two self-supervised objectives: • Masked Language Modeling (MLM) – predict masked tokens using context, potentially including linked documents. • Document Relation Prediction (DRP) – classify the relation between two text segments as one of	• PubMed abstracts (21GB; same corpus as PubmedBERT) • Citation links extracted using Pubmed Parser.	• Outperforms PubmedBERT on all tasks in the BLURB benchmark, improving average score by +2% (base) and +3% (large) absolute. • Achieves 44.6% accuracy on MedQA-USMLE,	• Not multimodal • Only abstracts used (not full text). • Focused only on citation links; other types of inter-document relationships (e.g., shared authorship, co-mentions) were not explored. • While DRP improves performance, the

		citation links between PubMed articles to internalize multi-document knowledge during pretraining.	{contiguous, random, linked}.		outperforming PubmedBERT (38.1%) and other baselines. • On MMLU-professional medicine, gets 50.7%, outperforming GPT-3 (175B) and UnifiedQA (11B), despite being only 340M parameters.	model still depends on the quality and coverage of citation data—meaning incomplete citation graphs could limit learning.
Med-PaLM [7]	No	Med-PaLM aims to improve & evaluate LLMs for medical use by introducing a new benchmark (MultiMedQA) and a human-based evaluation method to check answers for accuracy, reasoning, bias, and potential harm—overcoming the limits of earlier automated tests.	Med-PaLM is built on the 540B-parameter PaLM architecture (decoder-only transformer) and fine-tuned using instruction prompt tuning on curated medical QA datasets spanning consumer, professional, and research domains (Section 3.2). To align the model with clinical values, reinforcement learning from human feedback (RLHF) is applied using a reward model trained on expert preference ratings (Section 3.3). Evaluation is conducted using a human framework with clinicians rating outputs on factuality, comprehension, reasoning, harm, and bias (Section 4).	Used publicly available and internal medical QA datasets, including MedQA (USMLE), PubMedQA, MedMCQA, and HealthSearchQA. They also created MultiMedQA by combining existing datasets and curated additional expert-written questions.	Med-PaLM achieved 67.6% accuracy on MedQA (USMLE), outperforming previous state-of-the-art by 17%, and was the first model to reach the expert-level performance threshold. It also showed improvements in helpfulness, factuality, and safety across multiple medical QA datasets.	Not multimodal Model's occasional hallucinations, lack of full transparency in decision-making, and performance inconsistencies across domains. Future work includes scaling to multimodal inputs (Med-PaLM M) and improving model alignment, robustness, and evaluation with real-world clinicians and patients.
Med-PaLM 2 [8]	No		Med-PaLM 2 is based on PaLM 2, a 540B parameter LLM, fine-tuned using instruction-tuned and expert-curated datasets from MultiMedQA. The alignment process includes reinforcement learning from human feedback (RLHF) and further safety tuning via preference modeling and expert review pipelines to optimize helpfulness and factuality.	MedQA (USMLE), PubMedQA, MedMCQA, MMLU (medical subset), HealthSearchQA, LiveQA, and MultiMedQA expert-curated questions.	Med-PaLM 2 surpassed Med-PaLM on MedQA (achieving 86.5% vs. 67.6%), and outperformed ChatGPT and GPT-4 on multiple medical benchmarks. It also achieved higher ratings from clinicians on helpfulness, accuracy, and minimal harm compared to previous models.	Not multimodal Despite gains, Med-PaLM 2 still hallucinates and lacks verifiable citations or full reasoning transparency. Future work includes improving long-context reasoning, incorporating multimodal inputs (e.g., imaging), and testing in real clinical workflows with healthcare providers and patients.
Clinical Camel [9]	No	To develop an open, expert-level medical LLM that overcomes proprietary model limitations by enabling transparent,	• Base model: LLaMA-2 (13B and 70B variants) • Training method: QLoRA for efficient fine-tuning • Dialogue-Based Knowledge Encoding (DBKE): Converts dense medical texts (e.g., clinical review articles) into	• ShareGPT (multi-turn dialogue data) • 20,000 PubMed clinical articles (pre-2021), transformed into ~100,000	• Five-shot evaluations: Clinical Camel-70B outperforms GPT-3.5 across all benchmarks	• Not multimodal • Not yet systematically evaluated across clinical settings • DBKE method not compared with other

		efficient training and supporting rigorous clinical evaluation.	synthetic, multi-turn dialogues using GPT-4 as a teacher model. • Human input masked during training. Trained on single H100 GPU for 1 epoch.	dialogues via DBKE • MedQA (USMLE): 4000 questions transformed into dialogues with retrieved justifications from source texts using GPT-4	• USMLE Sample Exam: 64.3% (vs. GPT-3.5 at 58.5%) • PubMedQA: 77.9% (vs. GPT-3.5 at 60.2%) • MedQA (USMLE): 60.7% (vs. GPT-3.5 at 53.6%) • MedMCQA: 54.2% (vs. GPT-3.5 at 51.0%) Surpasses GPT-4 on PubMedQA in five-shot setting.	augmentation techniques • Updating medical knowledge remains a challenge • Potential for inaccurate outputs; requires extensive validation before use in practice
PMC-LLaMA [10]	No	To develop an instruction-tuned LLaMA-based model trained specifically on PMC biomedical research articles, enabling low-resource biomedical LLM development through efficient fine-tuning (e.g., using LoRA).	The authors fine-tuned LLaMA-1 models (7B and 13B) using instruction-tuning on biomedical question answering datasets. They also built a medical-aligned instruction dataset using prompts derived from publicly available sources like MedQA, PubMedQA, and LiveQA. The training process was completed using LoRA for parameter-efficient fine-tuning.	Pretraining data: PubMed Central Open Access (PMC-OA) subset. Fine-tuning datasets: MedQA (USMLE), PubMedQA, LiveQA, MedMCQA.	PMC-LLaMA significantly outperforms general-domain models like LLaMA-13B and GPT-J on biomedical QA benchmarks, achieving notable accuracy gains on tasks such as MedQA and PubMedQA.	The model may still hallucinate or misinterpret ambiguous clinical questions. There is limited interpretability, and future work includes improving long-form reasoning, adding multimodal capabilities, and conducting real-world clinical validation.
MediPhi (SLM Framework) [11]	No	To overcome the lack of clinical data and the high cost and latency of LLMs by proposing a modular framework (SLM) optimized for clinical NLP tasks	Modular framework with: (1) pre-instruction tuning (PIT) on domain-specific corpora, (2) model merging of expert SLMs using techniques like SLERP and BreadCrumbs, (3) supervised fine-tuning and Direct Preference Optimization (DPO) using synthetic instructions from the MediFlow dataset	Five groups: PubMed (PMC, abstracts), Clinical (NoteChat, MTSamples), MedCode (ICD9/10 & ATC), Guidelines, MedWiki. Synthetic data (MediFlow: 2.5M instructions on 14 clinical tasks), CLUE+ benchmark (12 tasks, e.g. RRS QA, SDoH, MeDiSumCode)	MediPhi-Instruct (3.8B) achieved a CLUE+ score of 43.4%, showing a 6.9% absolute improvement over the base model, with 64.3% gain on SDoH, 49.5% on radiology QA, and 14% better performance than GPT-4-0125 on ICD-10 coding (Table 4, Section 4.3, Pages 5, 26–27).	Relies heavily on synthetic data generation; performance may vary across real-world domains. Future work includes incorporating multimodal data (e.g., radiology), enhancing transparency and trust in outputs, and deploying SLMs in real clinical workflows.

Table 1 Comparative overview of biomedical language models across key dimensions

Table 2 shows the comparison of biomedical LLM's architecture, scale and resource utilizations. While reviewing biomedical LLMs, I have observed significant variation in the level of technical detail provided in the research papers. For many models, parameter counts and training infrastructure were not explicitly reported. In such cases, I have inferred values based on the underlying base models (e.g., BERT-base ≈ 110M parameters). I indicate where estimates or assumptions were made and highlight the lack of transparency in CO_2 usage reporting, suggesting a need for standardized reporting practices.

Model	**Architecture**	**Parameters**	**Training Resources**	**Electricity/CO_2**
BioBERT	BERT-base	110M same as those for pre-training BERT	BioBERT is the first domain-specific BERT based model pretrained on biomedical corpora for 23 days on **eight NVIDIA V100 GPUs**	Not specified
ClinicalBER	BERT-base	110M (same as BERT)	Not specified	Not specified
Med-BERT	Custom BERT variant	~17M	Not specified	Not specified
BioGPT BioGPT-Large	GPT-2 medium GPT-2 XL (large)	347M 1.5B	Not specified	Not specified
BioMedLM	GPT-style (GPT-2)	2.7B	BioMedLM was trained on MosaicML Cloud using 128 A100-40GB GPUs for 6.25 days, processing 300B tokens with a sequence length of 1024 and batch size of 1024. The training leveraged PyTorch FSDP and the Composer library for efficient distributed training	Not specified
(LinkBERT)Bio LinkBERT	BERT-large	340M	Not specified	Not specified
Med-PaLM	PaLM (decoder-only)	Up to 540B	Google Cloud TPU v4	Not specified
Med-PaLM 2	PaLM 2 (decoder-only)	Not specified	Google Cloud TPU v4	Not specified
Clinical Camel	Not publicly documented	Not specified	Not specified	Not specified
PMC-LLaMA	LLaMA-based	13B	Not specified	Not specified
MediPhi (SLM Framework)	LLaMA-based	Not specified	Not specified	Not specified

Table 2 Overview of Biomedical Language Models – Architecture, Scale, and Resources.

Conclusion

This survey highlights the rapid progress of LLMs in healthcare, showing that domain-specific models like BioBERT, ClinicalBERT, and Med-PaLM outperform general-purpose LLMs on biomedical tasks. While these models offer strong results in QA, NER, and clinical prediction, challenges remain—including lack of multimodal input, hallucinations, and limited real-world validation. Future work should focus on improving transparency, clinical integration, and standardized evaluation.

References

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

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

[3] D. Rasmy, Y. Xiang, Z. Xie, and H. Zhi, “Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction,” NPJ Digital Medicine, vol. 4, no. 1, pp. 1–13, 2021.

[4] Y. Luo, Q. Sun, Y. Wang, H. Wang, X. Qin, and Y. Yang, “BioGPT: Generative pre-trained transformer for biomedical text generation and mining,” Briefings in Bioinformatics, vol. 24, no. 2, 2023.

[5] Y. Boiko, A. Dohan, J. Zhang, et al., “BioMedLM: A domain-specific language model for biomedical research,” Stanford CRFM, 2023. [Online]. Available: https://crfm.stanford.edu/2023/biomedlm.html

[6] Z. Yao, Y. Liu, and Z. Lin, “BioLinkBERT: Pretraining biomedical language models with document links,” arXiv preprint arXiv:2204.12110, 2022.

[7] K. Singhal, A. Azizi, T. Tu, et al., “Large language models encode clinical knowledge,” Nature, vol. 620, pp. 172–180, 2023.

[8] K. Singhal, A. Azizi, M. Nori, et al., “Med-PaLM 2: Expert-level medical question answering with multimodal self-consistency,” arXiv preprint arXiv:2305.09617, 2023.

[9] M. Wang, S. Zheng, R. Xu, et al., “Clinical Camel: Open-source medical large language model with expert-level performance,” arXiv preprint arXiv:2309.11185, 2023.

[10] M. Wang, J. Zhang, R. Xu, et al., “PMC-LLaMA: A biomedical LLM adapted from LLaMA using PMC articles,” arXiv preprint arXiv:2310.09699, 2023.

[11] J.-P. Corbeil, A. Dada, J.-M. Attendu, et al., “A modular approach for clinical SLMs driven by synthetic data with pre-instruction tuning, model merging, and clinical-tasks alignment,” arXiv preprint arXiv:2505.10717, 2025.