Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 3, 2026

MEDINFO-LLM: A PARAMETER-EFFICIENT RETRIEVAL-AUGMENTED EXPERT SYSTEM FOR LOCALIZED ALLOPATHIC DRUG INFORMATION

Ayaan Sultan[*1], Ikram Uddin Khan[2], Sadiq Ali[3], Daniyal Ahmed Khan[4], Malik Muhammad Arif[5], Muhammad Amin[6]

[*1,3,4,5,6]*Department of Electrical Engineering, University of Engineering and Technology, Peshawar, Pakistan*
[2]*South China University of Technology*

[*1]19pwele5537@uetpeshawar.edu.pk, [2]uddinikram343@gmail.com, [3]sadiqali@uetpeshawar.edu.pk
[4]17pwmec4340@uetpeshawar.edu.pk, [5]engrmalikarif@uetpeshawar.edu.pk,
[6]muhammadamin0334@gmail.com[6]

DOI: https://doi.org/10.5281/zenodo.23036062

Keywords
Parameter-Efficient Fine-Tuning, Retrieval-Augmented Generation, Clinical Decision Support Systems, Healthcare AI Safety, Drug Information Systems.

Article History
Received: 12 January 2026
Accepted: 24 February 2026
Published: 11 March 2026

Copyright @Author
Corresponding Author: *
Ayaan Sultan

Abstract
Due to the lack of healthcare resources, the variety of pharmaceutical brands, and the restrictions of reference manuals, access to reliable and contextual pharmacological information is still difficult to obtain in many resource-limited settings like Pakistan. While general large language models can be used to obtain general medical information, their tendency to produce unreliable statements hampers their use in clinical decision support applications. In this paper, we introduce MedInfo-LLM, an offline and parameter-efficient clinical decision support system for localized pharmacological information. MedInfo-LLM combines a 4-bit quantized BioMistral-7B model with a hybrid retrieval architecture consisting of a FAISS vector database and a knowledge graph built with Neo4j, which contains information about pharmaceuticals, their interactions, and drugs of the Pakistani drug registry. Parameter-efficient fine-tuning via LoRA and preference optimization via DPO, along with a three-stage safety framework against hallucinations, were applied to improve the quality of responses without exceeding a memory footprint of 3.8 GB. MedInfo-LLM was evaluated for detecting drug interactions, renal dose computation, and mapping of generic names to brand names for drugs from Pakistan, on a manually curated reviewed dataset by pharmacists consisting of 5,000 QA pairs split 70:15:15 into 3,500 training, 750 validation, and 750 testing examples, together with a separate 200-query benchmark spanning four tiers of reasoning complexity. MedInfo-LLM achieved a hallucination rate of 3.1% and an Exact Match score of 91.7%, with its observed performance on the evaluated dose-reasoning tasks higher than that of the comparison models tested. End-to-end inference latency was 60-180 seconds on an NVIDIA GTX 1650 with 4 GB VRAM.

I. INTRODUCTION
Access to accurate pharmacological information is one of the major elements of patient safety and effectiveness of treatments in modern health care institutions. There are problems in Pakistan concerning access to health professionals as shown by health workforce statistics compiled by World Health Organization [1]. Such a high workload in combination with low availability of resources creates serious challenges for healthcare

professionals in providing safe and evidence-based care. The management of chronic diseases is further complicated by polypharmacy and the widespread availability of multiple brands of the same generic medicine, making medication identification and therapeutic substitution more challenging. The occurrence of drug-drug interactions continues to be a major threat to medication safety in Pakistan. Retrospective analysis of 200 inpatients' prescriptions from the Lady Reading Hospital, Peshawar, has shown that the major decision error was the drug-drug interactions, which occurred in 49.86% of the 1,410 drugs prescribed [2].

Even though large language models (LLMs) have proven effective in language-based tasks, their usage in pharmacology poses risks. Recent studies have reported that general-purpose LLMs may generate inaccurate or unsupported medication-related recommendations, particularly in clinically sensitive scenarios such as pregnancy and renal impairment. Commercial API-based systems also introduce recurring operational costs and depend on reliable internet connectivity. This work aims to tackle these multi-dimensional issues with MedInfo-LLM, a specialized expert system that integrates the expressive capabilities of biomedical language models with retrieval-augmented architectures designed for pharmaceutical questions. In this paper, the term expert system refers to a domain-specific decision-support architecture that combines a biomedical language model with structured pharmaceutical knowledge retrieval and explicit safety constraints, rather than implying autonomous clinical decision-making.

Existing e-pharmacopeias face four key obstacles: (1) Temporal Inefficiency — all systems use keyword searching as opposed to natural language searching which makes the user responsible for finding the necessary information himself. Moreover, the cardiac intensive care units of Khyber Teaching Hospital and Hayatabad Medical Complex, Peshawar, have reported a significant occurrence of potential drug-drug interactions [3]; (2) Contextual Blindness — a query such as "dosage of metformin" returns general information without accounting for patient-specific parameters such as renal function; (3) Hallucination Risk – during the early testing of Mistral-7B-Instruct-v0.2, it was found that the model is capable of issuing warnings on safety concerns for ondansetron, and (4) Infrastructure Mismatch — larger models require high-memory GPUs unavailable in many tertiary-care settings in low- and middle-income countries.

It is hypothesized that a parameter-efficient biomedical language model augmented with external pharmaceutical retrieval and Pakistan-specific medication knowledge can achieve strong observed performance on medication-information tasks while requiring substantially fewer computational resources for local deployment than full-scale generative language-model systems.

This research contributes: (1) a localized pharmaceutical knowledge graph linking Pakistani brand names to generic drug entities using the DRAP National Formulary and Registered Drug Register [4]; (2) parameter-efficient medical adaptation via LoRA that reaches over 90% accuracy on the clinical QA dataset while keeping the deployed model under 4 GB; (3) a hybrid, context-aware retrieval architecture combining BioBERT-FAISS dense retrieval with Neo4j graph retrieval; (4) a clinical safety validation framework incorporating pharmacist-reviewed case scenarios; and (5) a deployment-ready implementation running entirely on consumer-grade hardware without cloud dependency.

II. RELATED WORK

The medical drug information system has gone from the conventional drug reference database system to clinical decision support systems. The former systems like Micromedex and Lexicomp were mostly drug information retrieval systems whereas the latter systems included some clinical decision support capabilities like checking for drug interactions and others. However, research on the computerized clinical decision support systems has found that a considerable percentage of generated alerts is overridden by the prescriber with alert override rates being between 46.2% and 96.2% in reviewed studies [6]. The clinically unimportant or duplicate alerts and lack of prioritization of alerts have been noted to be among factors causing inappropriate overrides and alert fatigue [6].

Biomedical representation models such as BioBERT [7] and ClinicalBERT [8] demonstrated that domain-adapted pretraining substantially improves performance on biomedical text-mining tasks relative to general-domain encoders [9]. More recently, large language models have shown strong general-purpose capability in medicine [10], and retrieval-augmented generation (RAG) [11] has

emerged as a mechanism for grounding LLM outputs in external evidence rather than relying solely on parametric knowledge, reducing unsupported generation. BioMistral [12], a biomedical-domain-adapted decoder-only language model, and parameter-efficient adaptation techniques such as LoRA [15] and quantization-based approaches [5] together make local, resource-constrained deployment of domain-specialized LLMs practical. Direct Preference Optimization (DPO) [14] provides a means of aligning model outputs with preferred, evidence-supported responses without a separate reward model.

III. METHODOLOGY

A. System Architecture and Model Selection

The proposed architecture separates retrieval and generation responsibilities. A dense representation model based on BioBERT performs semantic retrieval, while conversational generation is performed by BioMistral-7B, a decoder-only (causal) language model capable of generating natural-language answers to complex pharmacological questions, as opposed to BioBERT, which is an encoder-only model used mainly for representation and understanding tasks [12]. This separation allows domain-specific retrieval to be fused with generative language without text generation being part of the retrieval process. The BioMistral-7B model is loaded using 4-bit quantization during parameter-efficient fine-tuning to reduce memory consumption, enabling practical local execution on consumer-grade hardware [5].

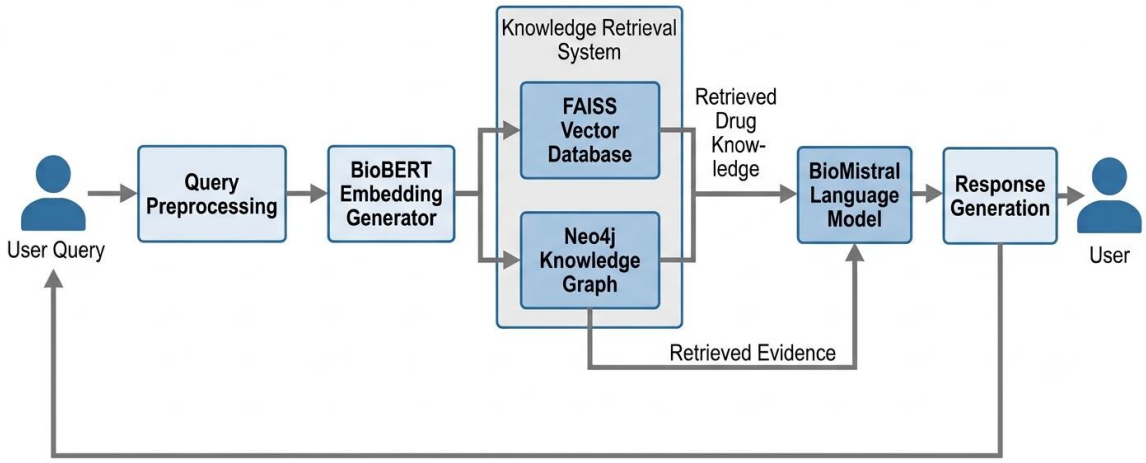


Fig. 1. Overall system architecture of MedInfo-LLM, showing query preprocessing, BioBERT-based embedding, hybrid FAISS/Neo4j retrieval, and BioMistral-7B response generation.

B. Dataset and Knowledge Base

The pharmaceutical knowledge base integrates the WHO Essential Medicines List, MedlinePlus, Drugs.com, and the DRAP registry, combining generic pharmacological information with Pakistan-specific brand, manufacturer, and registration data. A curated instruction-tuning dataset of 5,000 pharmacist-reviewed question-answer pairs were divided 70:15:15 into 3,500 training, 750 validation, and 750 testing pairs, covering renal and hepatic dose modification, pregnancy-related medication considerations, drug-drug interactions, and antimicrobial stewardship in the context of Pakistan's healthcare system. A separate, independently constructed benchmark of 200 clinical queries, organized into four tiers of reasoning complexity (factual retrieval, condition-based reasoning, drug-drug interaction analysis, and multi-source clinical reasoning; 50 queries per tier), was used for system-level evaluation and was not used for checkpoint selection or hyperparameter tuning.

TABLE II
Dataset Composition and Partitioning

Subset	Pairs/Queries	Percentage	Purpose
Training	3,500	70%	LoRA / SFT adaptation
Validation	750	15%	Checkpoint selection
Testing	750	15%	Held-out evaluation
200-query benchmark	200 (50×4 tiers)	–	System-level evaluation

C. Parameter-Efficient Fine-Tuning

Low-Rank Adaptation (LoRA) matrices were introduced into selected attention projections of BioMistral-7B while the pretrained base-model parameters remained frozen, substantially reducing the number of trainable parameters required for domain adaptation. Each training example was structured as an instruction containing the system role, user query, available patient context, retrieved evidence, and the pharmacist-validated answer, encouraging the model to ground responses in retrieved evidence and to request clarification or provide a safety disclaimer when information was insufficient. Direct Preference Optimization (DPO) was then used to fine-tune the model to prefer clinically suitable, evidence-based answers over less suitable or unsafe ones, using a fixed value of the weighting coefficient ($\lambda = 0.3$), and a scaling parameter $\beta = 0.1$ for the divergence constraint on the reference policy.

Curriculum Learning and Preference Optimization - The training process included three phases of increasing query complexity. In the initial phase, the training was done using Tier 1-2 queries related to drug information extraction and single parameter dose-related inference. Interaction queries of Tier 3 were introduced in the second phase where the model had to combine multiple retrieved drug monographs together. The contrastive examples with similar drugs but different interaction effects were introduced in order to distinguish interaction effects better. Finally, in the third phase, synthesis queries of Tier 4 were introduced that involved integrating multiple patient and drug dependent parameters. Direct Preference Optimization (DPO) was performed afterwards using preference pairs that rewarded the appropriate and safe answers.

D. Retrieval-Augmented Generation Module

The retrieval module is the one that adopts a two-level retrieval strategy according to the complexity of the query. 768-dimensional dense embedding vectors of the drug monographs are generated using BioBERT-base-cased and indexed with FAISS[13], and then use FAISS's inner-product similarity search to find semantically similar information from the drug monograph that goes beyond mere keyword matching. A Neo4j knowledge graph is used to model drugs as nodes and drug-drug interactions as relationships with severity, mechanism and recommended management annotations, which adds structured relationship information to document-level semantic retrieval for complex polypharmacy queries, for Tier 3-4 queries. For example, it can provide structured information about drug-drug interactions, including their mechanisms, severity, and recommended management. In this study, retrieval-level metrics like Recall@K or Mean Reciprocal Rank were not measured separately, but instead only the end result of the generated response was evaluated.

E. Safety and Hallucination Mitigation Framework

MedInfo-LLM has a 3-step safety framework to reduce nonspecific pharmacological reactions. In Stage 1 (Pre- Generation Safety Trigger), the confidence of a BioBERT-derived safety classifier determines whether the system will go into a safety-controlled response mode or if it will generate imperative treatment instructions. If the confidence is greater than 0.6, the system will go into a safety-controlled response mode without giving imperative treatment instructions. The generated response is then encoded using BioBERT and compared to the retrieved evidence using cosine similarity in Stage 2 (Retrieval

Consistency Checking), resulting in a score below 0.65, which leads to further verification or a disclaimer. In Stage 3 (Post-Generation Hallucination Screening), a lightweight DistilBERT-based classifier screens for unsupported information such as unreferenced medications or inconsistent dosages; responses exceeding a hallucination-probability threshold of 0.1 are withheld and replaced with a verification message. These operational thresholds were selected during system development based on preliminary validation observations rather than a systematic threshold sweep, and should be regarded as design choices of the evaluated prototype rather than statistically optimized decision boundaries.

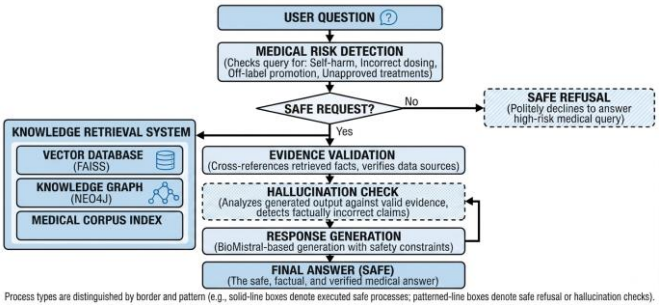


Fig. 2. Three-stage safety and hallucination mitigation framework: pre-generation risk detection, retrieval-consistency evidence validation, and post-generation hallucination screening.

IV. RESULTS AND DISCUSSION

MedInfo-LLM was compared against BioBERT (original, unfine-tuned), Mistral-7B-Instruct-v0.2, GPT-3.5-Turbo, and Lexicomp, a structured clinical drug-information reference. Table I summarizes performance across Exact Match, F1 score, hallucination rate, Appropriate Refusal Rate (ARR), and end-to-end inference latency, evaluated on the independent 200-query benchmark unless otherwise noted.

TABLE I
Quantitative Performance Comparison

Model	EM	F1	Hall.	ARR	Lat.(s)
BioBERT	38.4%	0.521	N/A	N/A	20
Mistral-7B-Instruct	67.2%	0.743	23.4%	8.9%	25
GPT-3.5-Turbo	81.5%	0.821	14.3%	18.7%	15
Lexicomp (ref.)	94.8%	0.912	0%*	100%*	90
MedInfo-LLM (ours)	91.7%	0.887	3.1%	91.4%	60–180

**Lexicomp is a static reference resource rather than a generative model; its hallucination and refusal values were operationalized differently and are not directly comparable to the generative systems' metrics.*
- BioBERT is an encoder-only language model designed primarily for representation and classification rather than free-form text generation; therefore, hallucination rate and appropriate refusal rate were not applicable and are reported as N/A.

MedInfo-LLM achieved the highest observed Exact Match and F1 scores among the evaluated generative language-model systems. Lexicomp, a structured clinical drug-information reference rather than a generative model, scored higher on these accuracy measures but is not a conversational system and is therefore treated as a reference rather than a directly equivalent generative baseline. The hallucination rate (3.1%) of MedInfo-LLM is much lower than that of the evaluated generative baselines, namely Mistral-7B-Instruct-v0.2 (23.4%) and GPT-3.5-Turbo (14.3%). Such difference in results can be attributed to the use of evidence retrieved by MedInfo-LLM, safety-driven training and post-generation evaluation of the generated text; however, without conducting a component

ablation analysis, one cannot assess the contribution of each mechanism separately. Mistral-7B-Instruct-v0.2 exhibited a lower observed Appropriate Refusal Rate, suggesting it was more likely to provide an answer even when supporting evidence was limited, whereas MedInfo-LLM's higher ARR reflects safer refusal behaviour under insufficient-evidence conditions.

On a separate benchmark of 200 queries involving Pakistani pharmaceutical brand names, MedInfo-LLM achieved 94.1% accuracy for brand-to-generic mapping, 91.8% for generic-to-brand suggestion, and 87.3% for manufacturer-context recognition, summarized in Table III. Performance decreased for complex polypharmacy queries, with observed accuracy of 76.3% for the most complex eight-drug combinations compared with 91.7% for general multiple-drug queries, reflecting the increased number of pairwise interaction relationships that must be considered as the number of concurrent medications grows; this figure should be treated as a single illustrative observation rather than a statistically characterized trend, since the underlying sample size was not recorded. No independent repeated runs were performed, so standard deviations, confidence intervals, and statistical significance estimates are not reported; the results should be interpreted as observed performance of the evaluated configuration rather than definitive statistical superiority.

TABLE III
Pakistani Brand/Generic Recognition (n=200)

Recognition Task	Accuracy
Brand → Generic mapping	94.1%
Generic → Brand suggestion	91.8%
Manufacturer-context recognition	87.3%

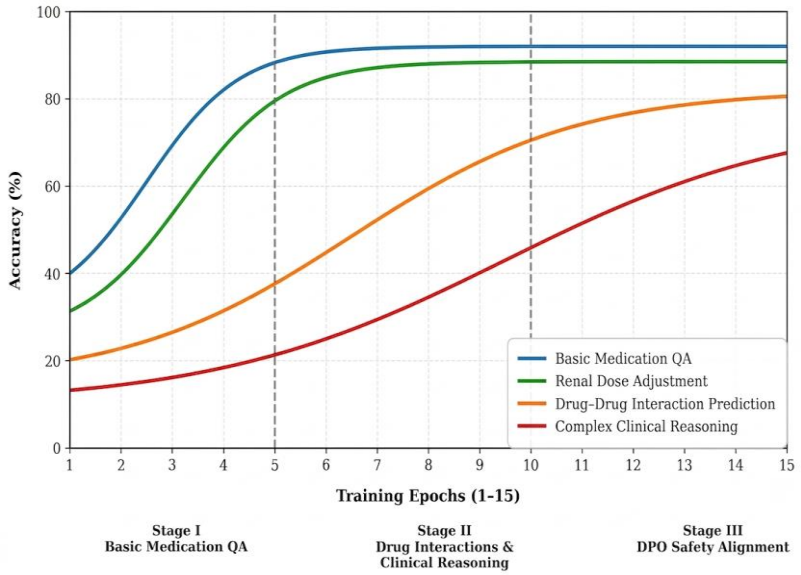


Fig. 3. Task-wise accuracy convergence across the three-stage curriculum-learning process over 15 training epochs.

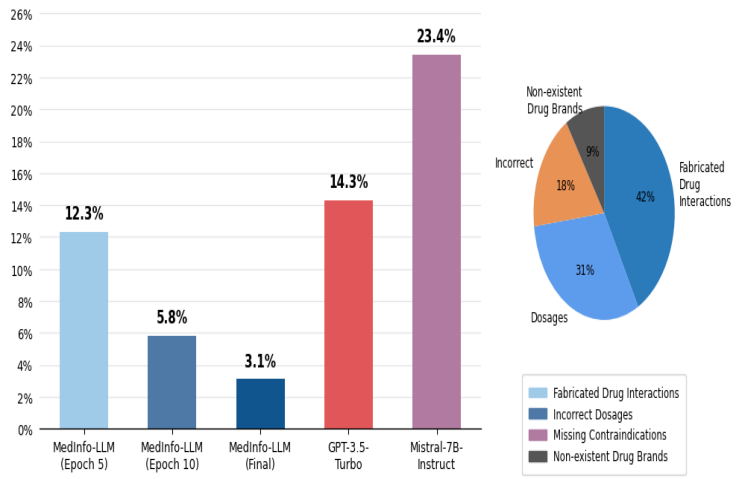


Fig. 4. Hallucination rate across training checkpoints and comparison models (left), and categorical distribution of observed error types (right).

V. LIMITATIONS

A number of caveats restrict interpretation of these findings. The model's parametric knowledge is once-trained and will not change after training; the external knowledge base can be updated by retrieving the knowledge, which does not need retraining, but there may be some new reasoning pattern, which may need more fine-tuning. Currently queries can only be entered into the system in English, thus restricting the ability of Pashto, Punjabi and Urdu speaking populations to access the system. Performance deteriorates for regimens exceeding four concurrent drugs due to increased interaction-graph complexity. The system does not incorporate patient-specific pharmacogenomic parameters, has not undergone regulatory approval for clinical use, and its inference latency (60-180 seconds) is relatively high for interactive use, reflecting the additional processing required for retrieval, generation, and multi-stage safety checking. Finally, evaluation relied on scenarios constructed by the research team and reviewed by pharmacists rather than real-world patient encounters, and no component-level ablation study was performed, so the individual contributions of retrieval, fine-tuning, and safety mechanisms cannot be isolated.

VI. CONCLUSION

This paper presented MedInfo-LLM, a parameter-efficient pharmaceutical information system combining a 4-bit quantized BioMistral-7B model with BioBERT-FAISS dense retrieval, Neo4j graph-based interaction retrieval, LoRA adaptation, DPO-based preference alignment, and a three-stage safety framework. The findings provide evidence that domain specialization combined with retrieval augmentation can improve medication-information performance under the evaluated conditions, achieving 91.7% Exact Match accuracy and a 3.1% hallucination rate while enabling local inference on consumer-grade hardware with under 4 GB of VRAM. The system is intended to support, rather than replace, professional clinical decision-making. Future work should include component-level ablation, repeated evaluation across independent random seeds, expansion of the evaluation benchmark, multilingual support for Urdu and Pashto, and external validation using anonymized real-world clinical queries from healthcare institutions.

ACKNOWLEDGMENT

The authors thank the Department of Electrical Engineering, University of Engineering and Technology, Peshawar, and the pharmacists who supported dataset review and validation.

REFERENCES

World Health Organization, "Health Workforce Indicators: Pakistan Country Profile," WHO Global Health Observatory, 2023.

H. Hussain, K. A. Shah, K. Ullah, S. Ahmed, A. Khan, A. Ullah, and A. Aman, "Assessment of Prescription Pattern and Prescription Errors Using the World Health Organization Drug Use Indicators in Lady Reading Hospital Peshawar, Pakistan: A Retrospective Study," *Khyber Medical University Journal*, vol. 9, no. 4, pp. 181–184, 2018.

F. Shakeel, J. A. Khan, M. Aamir, P. A. Hannan, S. Zehra, and I. Ullah, "Risk of potential drug-drug interactions in the cardiac intensive care units: A comparative analysis between 2 tertiary care hospitals," Saudi Med. J., vol. 39, no. 12, pp. 1207–1212, 2018.

Drug Regulatory Authority of Pakistan (DRAP), "Registered Product Data," Drug Regulatory Authority of Pakistan, 2026.

T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient finetuning of quantized LLMs," in Proc. NeurIPS,2023.

T. N. Poly, M. Islam, H.-C. Yang, and Y.-C. Li, "Appropriateness of Overridden Alerts in Computerized Physician Order Entry: Systematic Review," *JMIR Medical Informatics*, vol. 8, no. 7, e15653,2020.

J. Lee et al., "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," Bioinformatics, vol. 36, no. 4, pp. 1234–1240,2020.

E. Alsentzer et al., "Publicly available clinical BERT embeddings," arXiv:1904.03323,2019.

Y. Peng, S. Yan, and Z. Lu, "Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten biomedical tasks," in Proc. BioNLP Workshop,2019, pp. 58–65.

B. J. Erickson et al., "Applications of large language models in medicine," N. Engl. J. Med., vol. 389, no. 15, pp. 1410–1417,2023.

P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Proc. NeurIPS, vol. 33,2020, pp. 9459–9474.

Y. Labrak et al., "BioMistral: A collection of open-source pre-trained large language models for medical applications," arXiv:2402.10373,2024.

J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," IEEE Trans. Big Data, vol. 7, no. 3, pp. 535–547,2019.

R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," in *Proc. NeurIPS*, vol. 36,2023.

E. J. Hu, Y. Shen, P. Wallis et al., "LoRA: Low-rank adaptation of large language models," in Proc. ICLR,2022.

IEER
Institute of Excellence in Education & Research