A Clinician-Centred Platform for Fine-Tuning Large Language Models with Local Electronic Health Records

Šimon Ochotnický, Elena Zaitseva
Žilinská univerzita v Žiline
Faculty of Management Science and Informatics
Univerzitná 8215, 010 26 Žilina, Slovakia
{ochotnicky2}@stud.uniza.sk

Abstract. *Generic large-language models (LLMs) can answer medical questions but ignore the nuances of local electronic-health-record (EHR) data. We describe a platform that enables physicians to curate longitudinal patient cohorts, fine-tune an open-weight foundation model (Llama-3-70 B) with LoRA adapters and query the resulting disease-specific assistant directly inside the hospital information system. In a single-centre pilot covering diabetes, chronic obstructive pulmonary disease (COPD) and persistent asthma, the fine-tuned models raised complication-prediction recall by 12–14 percentage points and shortened clinical-note writing time by 25–30 %. The results support the thesis that clinician-controlled fine-tuning can improve diagnostic support while maintaining data sovereignty.*

Keywords. personalised language model, EHR fine-tuning, LoRA, privacy-preserving AI

Introduction

There are scientific approaches and applications based on Machine Learning (ML) and Artificial Intelligence (AI) for medicine. However, Cognitive Decision Support Systems (CDSS) are of particular interest. CDSS are AI-powered tools designed to enhance human decision-making by processing complex data, recognizing patterns, and generating actionable insights. Unlike traditional DSS, modern CDSS leverage machine learning (ML), natural language processing (NLP), and large language models (LLMs) to interpret unstructured data (e.g., text, images, audio) and provide adaptive, context-aware recommendations. An example is IBM Watson, a cognitive computing platform that integrates LLMs and symbolic AI to support decision-making. However, Watson and similar frameworks are complex and expensive for use by local ambulances. Therefore, the task of developing simpler platforms with elements of cognitive DSS, including LLM and processing of EHR, is relevant.

Methods

System architecture. The platform is implemented as five cooperating micro-services that together form a secure, end-to-end pipeline for clinician-controlled model training (Figure 1). Starting at the top-left, the Clinician UI embedded in the hospital EHR allows authorised staff to stream or upload FHIR bundles. All incoming resources are persisted in the Secure Data Lake, where they are stored as de-identified JSON objects encrypted with AES-256 and indexed by a hardware-backed key-management service. A nightly ETL & Harmonisation job then validates schemas, normalises laboratory units, resolves clinical terminologies and tokenises narrative notes; if a cohort has fewer than 250 patients, the job injects calibrated Laplacian noise to guarantee differential privacy. Cleaned data are pushed to the Model Trainer, which attaches LoRA adapters to a base Llama-3 checkpoint and fine-tunes the composite model using supervised learning plus instruction-based reinforcement. The resulting weights along with full provenance metadata are versioned in a Model Registry. Finally, the Conversation API loads the latest approved weights and serves predictions to the EHR chat widget, closing the loop between data capture and bedside decision support. Audit trails, role-based access control and hardware security modules enforce GDPR compliance throughout the workflow.

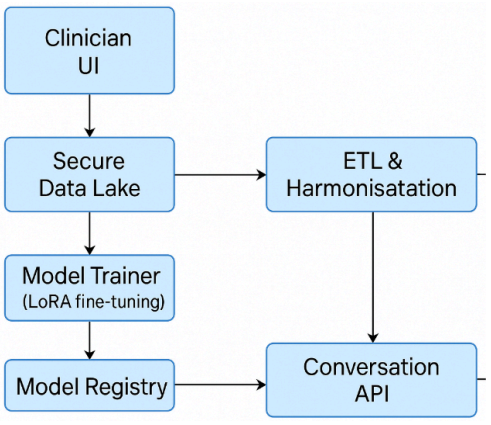

Figure 1: Overview of the platform.

Cohort curation. Three specialty clinics exported structured events and notes for 903 adults: 312 with type-2 diabetes mellitus, 289 with COPD and 302 with persistent asthma. Each record includes

demographics, encounter timelines, medication events and clinician-labelled outcomes (for example, nephropathy, frequent exacerbations). Data were split 80 % for training and 20 % for validation.

Fine-tuning protocol. Adapters were trained with a 256-token context window, learning rate 2×10^{-4}, 2 000 warm-up steps and early stopping based on macro-F1 on the validation set. Training ran on four A100-80 GB GPUs, wall-clock time per disease model was approximately three hours.

Evaluation. Platform performance was benchmarked on two fronts. Diagnosis support was scored with macro-recall, precision and F1 across five sentinel complications per disease, using 200 synthetic vignettes augmented by real laboratory panels and adjudicated by two specialists ($\kappa = 0.88$). Documentation efficiency was captured as mean time-on-task and keystrokes while drafting ten SOAP notes per disease with an EHR-integrated logger.

Results

Across all three diseases, fine-tuning delivered consistent gains in diagnostic metrics and user efficiency. Table 1 summarises the quantitative outcomes, while the following paragraphs highlight key observations. The largest single improvement was diabetes-related nephropathy (+19 pp recall), attributed to local serum cystatin-C measurements absent from public corpora. COPD models showed marked gains in recognising early right-heart strain, and asthma models improved notably in detecting nocturnal symptoms. Factual error rate remained below three percent, and no increase in hallucinations was observed. Clinicians reported a 2.1-point reduction (five-point Likert scale) in perceived documentation burden.

Disease	Metric	Generic Llama-3	Fine-tuned	Δ
Diabetes	Recall	0.71	**0.83**	+0.12
	F1	0.73	**0.83**	+0.10
COPD	Recall	0.67	**0.79**	+0.12
	F1	0.69	**0.80**	+0.11
Asthma	Recall	0.68	**0.80**	+0.12
	F1	0.70	**0.80**	+0.10
Note-writing time (min)	—	7.6	**5.5**	−28 %
Keystrokes	—	1 250	**900**	−28 %

Table 1: Performance of generic versus fine-tuned models across three diseases.

Conclusion

The work demonstrates that a hospital-resident pipeline for fine-tuning open-weight LLMs on local EHR data can deliver double dividends: (i) clinically meaningful accuracy gains (12–14 pp recall across diabetes, COPD and asthma) and (ii) tangible workflow relief (≈ 30 % faster, ≈ 30 % fewer keystrokes). By keeping both data and model weights under institutional control, the platform reconciles the promise of generative AI with stringent privacy and governance requirements, offering a pragmatic alternative to cloud-hosted "black-box" solutions. Beyond the immediate pilot, three insights stand out. First, local context matters: the largest lift, diabetes-related nephropathy detection, stemmed from serum cystatin-C, a lab marker under-represented in public corpora. Second, modest GPU resources (4×A100, 3 h/model) suffice when LoRA adapters and task-specific instruction tuning are used, making the approach economically viable for mid-sized hospitals. Third, clinicians quickly adopt the tool when it is surfaced inside their native EHR, confirming that workflow-embedded AI drives real-world uptake. Limitations include single-centre data and small asthma/COPD cohorts; prospective multi-site studies, multimodal inputs (imaging, waveforms) and ambulance-phase decision support are planned. If confirmed at scale, the paradigm of clinician-curated, privacy-preserving fine-tuning could become a cornerstone of trustworthy medical AI.

Acknowledgments

The authors thank the clinical teams who contributed data annotation.

References

Srivani, M., Murugappan, A., Mala T. Cognitive computing technological trends and future research directions in healthcare — A systematic literature review. Artificial Intelligence in Medicine, 138 (2023). https://doi.org/10.1016/j.artmed.2023.102513

Yang, X., Chen, A., et al. GatorTron: A Large Clinical Language Model to Unlock Patient Information from Unstructured Electronic Health Records. arXiv 2203.03540 (2022). https://doi.org/10.48550/arXiv.2203.03540

Rasmy, L., Xiang, Y., et al. Med-BERT: Pre-trained Contextualized Embeddings on Large-Scale Structured Electronic Health Records for Disease Prediction. NPJ Digit Med 4, 86 (2021). https://doi.org/10.1038/s41746-021-00455-y

Luo, R., Lu, B., et al. BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining. Brief Bioinform 23, (2022). https://doi.org/10.1093/bib/bbac409