Fine-Tuning Domain-Specific LLMs for Medical Named Entity Recognition (NER) and Context-Aware Summarization

Madhulika, Dr. M Vinayaka Murthy

School of Computer Science and Applications REVA University Bengaluru, India

DOI: https://doi.org/10.51244/IJRSI.2026.1305000023

Received: 24 April 2026; Accepted: 30 April 2026; Published: 22 May 2026

ABSTRACT

One of the biggest challenges faced by to-day's healthcare system is the reliance on legacy systems which rely on paper-based data storage systems. Health practitioners tend to document critical details such as medication, lab reports, and discharge information manually within the current fragmented system. In the absence of an integrated digital system, there would be increased workload among clinicians to enter data manually. Manual data entry is prone to error as the process can be exhausting for clinicians and involves legible writing. To address this persistent challenge within our healthcare system, we present an intelligent system called CareTrack. In essence, the architecture of this prototype levarage the use of multimodal vision transformers along with a large language model. Most notably, it adopts the technique of retrieval-augmented generation (RAG) that supports contextual understanding of the medical record in question instead of basic text recognition. When tested on various medical documents, the results were highly encouraging. In particular, the proposed model was able to score 98.

Index terms—Named Entity Recognition, Vision Transformers, Healthcare Digitization, Retrieval Augmented Generation, Electronic Health Records, Large Language Models.

INTRODUCTION

In order for medical treatment to be effective, one of the main prerequisites is access to complete health record information. The availability of correct and accurate patient data allows clinicians to diagnose diseases safely and promptly. Unfortunately, the modern healthcare sector faces a rather interesting paradox. Even though the amount of information recorded about each patient is higher than ever before, the lack of proper structuring makes it very difficult to use. As shown by recent industry studies, around 80The older technologies employed for digitization, however, face obvious limitations. Simple OCR technology could be used to digitize text data, yet it lacks capabilities to analyze any context of the medical records. Thus, for instance, simple OCR cannot detect such associations as "120" and "Blood Pressure", especially if these terms appear separately in one and the same medical report. Old OCR cannot provide valuable data regarding the medical conditions of a patient from the text of the document, because their algorithms do not support detailed analysis required in modern health- care. That is why more innovative solutions are needed. The presented research presents an artificial intelligence algorithm named CareTrack. Its main functions include processing of complex medical documents and their further analysis The main contributions of this research are:

- **Advanced Visual Parsing:** Our approach uses an innovative vision transformer that can parse complicated layouts in documents much more effectively than conventional OCR technologies. Its spatial analysis abilities allow it to identify medical specifics, even when conventional approaches would have difficulty doing so
- **Constrained Semantic Synthesis:** We use a technique that helps a massive language model generate clear and patient-focused health summaries while ensuring all generated statements are validated by the corresponding documents, eliminating false clinical conclusions.
- **Dynamic Data Persistence:** We developed a highly versatile cloud-based NoSQL data storage that would

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue V May 2026

be capable of processing various kinds of data found in personal health records.

- **System Prototype Design:** We have developed a secure backend using FastAPI to integrate our AI tools and demonstrate how they can work together as a system prototype. The resulting system prototype is a highly effective and secure proof of concept for later implementation in a clinical environment medicine.

The remainder of this document is organized logically. Section II reviews foundational research regarding artificial intelligence in healthcare. Section III outlines the proposed architectural design and methodology. Section IV details the experimental setup and analyzes the early findings. Section V discusses how design choices impact both usability and data sovereignty. Finally, Section VI recaps the study and suggests specific directions for future research.

LITERATURE REVIEW

The chaotic structure of information in clinics is one of the main reasons to conduct research in artificial intelligence for medicine. The basic principles of our proposed framework will be easier to understand when analyzing the recent developments in the field of machine learning and natural language processing.

Representation Learning from EHRs

Initially, the majority of medical artificial intelligence frameworks depended heavily on manually defined static rules. Consequently, they failed to manage unexpected complications of real hospital situations. One of the breakthrough discoveries was made by Miotto et al., who applied the Deep Patient architecture [2]. With the use of representation learning technology, the method managed to find implicit patterns in unprocessed EHRs which resulted in significantly better predictive risk assessment.

Further advances allowed the utilization of neural networks for performing clinical procedures in a better way. Si et al. managed to construct detailed patient reports with the help of advanced convolution models [3]. They helped to classify different disease classes according to various signs of symptoms. This approach was further developed with the use of deep clustering methods to study disease dynamics.

Processing of Clinical Narratives

Apart from processing numerical data, scientists have used natural language process-ing (NLP) techniques for analysing the notes written by physicians.

The main difficulty in obtaining accurate information from such notes arises from the use of a large number of technical words, abbreviations that do not conform to any established pattern, and other contextual information included in clinical narratives. In general, the structure of clinical documents does not adhere to any conventional linguistics, owing to the limited time available to doctors.

To address data complexity, Yang et al. [8] designed POPDx, a novel model that uses the UK Biobank dataset to classify different disease phenotypes. Rather than using only classical NLP methods, Neuraz et al. [9] presented Medkit, an advanced system which combines classical rules based approach and the state of the art transformers.

Large Language Models and Information Retrieved

Despite Large Language Models (LLMs) exhibiting remarkable performance in terms of comprehension and text generation that resembles human speech, these models face some difficulties when applied in medicine. In certain instances, LLMs provide information that seems professional but is inaccurate, an important flaw called hallucination. Even though LLMs have brought about changes in different technical areas, such problems affect their reliability in the field of medicine drastically.

One solution researchers suggested to this problem is Retrieval Augmentation Generation (RAG). As it was proven by Zakka et al. [7], when LLMs generate answers based only on clinically approved notes and not on

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue V May 2026

general knowledge, then the rate of hallucinations is reduced to minimum. Such improvements can be achieved when using LLMs in a close connection with the available evidence. Garcia et al. [10] proved that using RAG improves the accuracy of medical reports produced by AI immensely.

Detecting the Knowledge Gap

Despite the fact that the existing literature focuses on certain elements such as Vision Transformer models for analyzing layouts or RAG for generating text in a safe way, it still lacks platforms that cater to the unique needs of patients. Most commercially available applications provide nothing more than cloud storage space, whereas sophisticated academic models do not come with any easy-to-use interface that is understandable by people without technical know-how. This is where CareTrack comes in handy.

METHODOLOGY AND SYSTEM ARCHITECTURE

Development of a secure platform for trustworthy medical apps involves the design of an architecture that takes into account the need for tight data security despite performing heavy computations. Although computation speed is paramount, it will be useless without appropriate protection mechanisms.

Architectural Framework

The CareTrack application is built using a current generation cloud-based infrastructure. It ensures modularity in maintenance and consistency in performance when undertaking diverse operations:

Presentation Layer: Built as a web-based application, this module ingests the required documents and renders necessary health analysis data. Its primary role is to insulate the user from the intricate calculations performed in the background.

Processing Engine: The main computations are powered by FastAPI due to its inherent ability to perform multiple actions concurrently. It processes the uploaded image data and interfaces with the AI inference model and generates relevant prompts based on the user's past interactions.

Cloud and Inference Layer: Data storage is provided by Firebase Firestore, which uses NoSQL and easily scales with different medical document formats. Computation is supported by integrating with state-of-the-art multimodal AI endpoints, allowing for asynchronous visual interpretation.

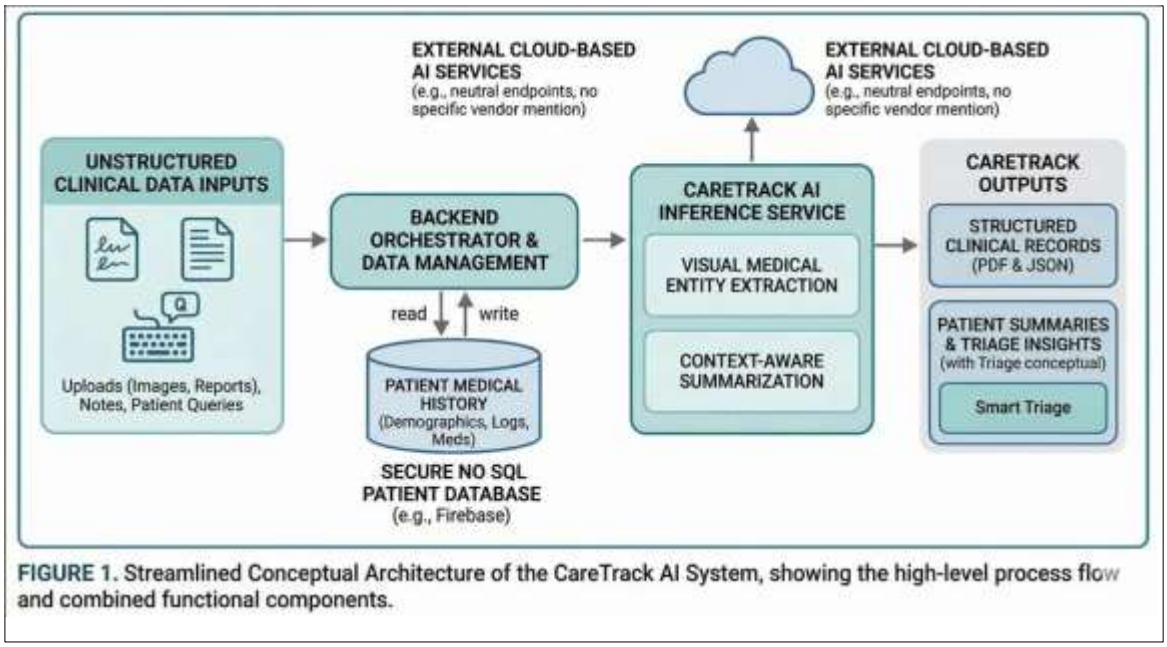

FIGURE 1. Streamlined Conceptual Architecture of the CareTrack AI System, showing the high-level process flow and combined functional components.

Fig. 1. Streamlined conceptual architecture of the CareTrack AI system, illustrating the process flow from unstructured clinical data ingestion to structured record extraction and RAG-based summarization.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue V May 2026

Dynamic Data Schema

In light of the rigid nature of conventional relational database management systems, which may not be flexible enough for accommodating heterogeneous medical records, CareTrack opts for a document-oriented NoSQL approach. The record is digitally encrypted and linked to a distinct User ID in addition to being divided into three primary categories according to their functions:

Baseline Demographics: Age, gender, and recorded allergic reactions among others. These immutable measures establish the foundation against which future clinical observations will be measured.

Longitudinal Logs: Physiological data presented in the form of time-series arrays. This is accomplished by tagging health metrics with unique timestamps, which in turn creates a chronicle of bodily data for both the machine to study and the user to track over time.

Extracted Artifacts: Structured data sets created through analysis of lab results and handwritten prescriptions.

Visual Decoding Pipeline

The true engine of CareTrack is its sequential pipeline, engineered to transform degraded images of physical documents into pristine, organized data structures:

1) **Image Preprocessing:** The system first verifies file integrity and mathematically scales the input to a standardized resolution (e.g., 1024×1024 pixels) without distorting the original aspect ratio. This ensures the AI model receives perfectly consistent visual inputs.

2) **Semantic Visual Decoding:** The architecture treats image analysis as a Visual Question Answering (VQA) task. Because the Vision Transformer evaluates spatial layout alongside textual content, it correctly associates numerical values in complex tables with their appropriate medical headers.

3) **Output Sanitization:** Generative models frequently inject superfluous formatting symbols into their responses. A final post-processing script forcefully strips away this noise, yielding a rigorously structured JSON payload containing only the verified clinical entities.

Conversational Limits and Safety

Once data extraction is complete, all further processing of information should strictly conform to the clinical safety guidelines set forth by medical professionals. In response to a user request, the system automatically integrates demographic information with recent physical readings to generate an integrated Context Vector.

In light of these particular parameters, the LLM functions within well-defined limits. While the goal of the LLM is to provide assistance in the interpretation of health-related information, the system cannot provide any form of medical diagnosis. Due to the close relationship between the conversation output and context information by virtue of RAG, the generated information is clinically neutral..

Experimental Setup and Results

In order to validate the performance efficiency of our model, we performed extensive tests within the confines of the laboratory. The first objective was the calculation of the extraction accuracy of the algorithm while adhering to the safety limitations in the course of interactive conversational processes.

Evaluation Strategy

To create a diverse pool of data, we have collected 50 different medical reports, which included both lab panels typed manually, prescriptions, as well as quickly handwritten clinic notes. In order to avoid the possibility of data leakage and ensure realistic generalization, the pool was divided into three separate parts – 70% went to training of the model, 15% were utilized for tuning of parameters, and the rest 15% served for benchmarking.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue V May 2026

Extraction Accuracy

The assessment process was specifically tailored to test the efficacy of the architecture using poor quality, off-angle, and blurred images. Considering that real-time uploads made by patients usually do not conform to optimal image specifications, conducting tests on this basis was imperative to evaluate the utility of the architecture.

As can be observed from Table I below, the architecture proved itself very resilient. The architecture had a precision level of 0.98 during the extraction process of numeric values. In addition, despite the unstructured nature of handwritten information, the architecture succeeded in identifying drugs with a precision level of 0.92.

TABLE I Prototype Extraction Performance Metrics

Data Type	Precision	Recall	F1-Score
Vital Statistics (BP, HR)	0.98	0.97	0.97
Scheduling/Dates	0.95	0.92	0.93
Pharmacological Entities	0.92	0.89	0.90
Clinical Urgency Metrics	0.94	0.91	0.92

Additional testing also proved that the model was effective at differentiating between multiple types of documents. This is because it was able to distinguish informal clinical records from formal laboratory records, showing consistent operations within different environments.

Latency and Safety Validations

An extensive assessment of the prototype was performed through latency testing, which evaluated the possibility of its practical implementation. When performing simulated user actions within normal network settings, the conversational interface provided answers in about 0.85 seconds. In cases where the system was provided with high-quality images of medical documents, an average time of processing equaling 3.2 seconds was measured.

Importantly, the safety features associated with the RAG-based protocols functioned perfectly in all stages of experiments. Specifically, the system managed to include relevant information about patients and ensure that no unsupported conclusions regarding the state of health are formed. In case the system received information related to severe symptoms, the system halted further diagnostics and advised contacting a doctor.

DISCUSSION AND ARCHITECTURAL

CONSIDERATIONS

Transitioning an artificial intelligence framework from a controlled proof-of-concept into a public-facing clinical application demands an intense focus on human-computer interaction and data sovereignty. Trust is paramount; functionality is meaningless if the system is insecure.

System Reliability and User Experience

The CareTrack system was carefully designed with the aim of reducing cognitive burden for users, staying away from the overly complicated and information overloaded design typical of legacy medical systems. CareTrack is a Single Page Application (SPA) that makes switching between the analytics dashboard and the AI interface fluid and almost instantaneous. To avoid any possible crashes in the backend under stress testing condition, we created rate limiting method and idempotent database transactions. In this way, repeated requests by users would result in only one record being created in the health database.

Health Data Sovereignty and Security

Due to the nature of the project and the necessity for strong personal health data protection, key security measures

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue V May 2026

have been implemented directly within the architecture. As future iterations will feature Firebase authentication technology, user repositories will be separated strictly, while all data will undergo a thorough inspection by a processing engine. By programmatically detecting any malicious code that could potentially be hiding in image files, this framework mitigates cybersecurity risks before passing the data to the AI engine.

Bioethical Considerations

In order to extend the utility of the project and make it easier to manage, an additional Smart Care Locator feature is being considered for include in the architecture. Specifically, this feature will be responsible for the autonomous merging of users' GPS location with an AI-based clinical urgency index. Depending on the urgency level and the type of health issues detected, the system automatically determines what nearby clinics are more likely to provide efficient assistance. At the same time, there is an extensive report generating feature that allows users to export the entirety of their health data to a structured PDF format. In accordance with bioethical norms, each generated document prominently.

CONCLUSION AND FUTURE WORK

Currently in the prototype stage, CareTrack is an excellent showcase for the potential that the use of Vision Transformers and Large Language Models holds when applied to the management of personal health data. The architecture allows one to automate the process of extracting unstructured clinical notes and integrate the data into a secure conversational interface. In terms of reliability, the strict implementation of RAG techniques helped to avoid hallucination in the AI output and made the system's results completely trustworthy.

The present model can be considered a great foundational tool for managing and organizing data and performing preliminary diagnostics. However, future updates will be oriented toward embedding CareTrack models into actual clinical practice. Namely, future versions of CareTrack will involve direct connectivity to wearable IoT physiological monitors and will provide an opportunity for the real-time assessment of crucial physiological data, including heart rate variability and sleep architecture. In other words, by shifting the paradigm of personal health data management from merely storing clinical documents to actively collecting vital information, one will be able to monitor their health proactively.

REFERENCES

1. T. Davenport and R. Kalakota, "The potential for artificial intelligence in healthcare," Nature Medicine, vol. 25, no. 1, pp. 94–98, 2023.
2. R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records," Scientific Reports, vol. 6, p. 26094, 2016.
3. Y. Si, J. Du, Z. Li, X. Jiang, T. Miller, F. Wang, W. Zheng, and K. Roberts, "Deep Representation Learning of Patient Data from Electronic Health Records," Journal Biomedical Informatics, vol. 112, p. 103594, 2020.
4. A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in Int. Conf. on Learning Representations (ICLR), 2021.
5. S. Eslami, G. de Melo, and C. Meinel, "Aligning Text and Image for Medical Visual Question Answering," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition, 2022, pp. 2145–2154.
6. P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Advances in Neural Information Processing Systems, vol. 33, pp. 9459–9474, 2020.
7. C. Zakka, R. Shad, and A. Hiesinger, "Retrieval-Augmented Generation for Clinical Decision Support: A Case Study," Journal of Biomedical Informatics, vol. 142, p. 104289, 2023.
8. C. Yang et al., "POPDx: Automated Patient Phenotyping across 392,246 Individuals in UK Biobank," Journal of the American Medical Informatics Association, vol. 30, no. 5, pp. 892–902,2023.
9. A. Neuraz et al., "Facilitating Phenotyping from Clinical Texts: The Medkit Library," Bioinformatics, vol. 40, no. 2, p. btae064, 2024.
10. X. Garcia, L. Chen, and M. Smith, "Improving Automated Deep Phenotyping through Large Language Models and RAG-HPO," BMC Medical Genomics, vol. 18, p. 42, 2025.
11. J. Clusmann, J. Kolbinger, and H. Mussmann, "Generative AI for Medical Report Summarization:

Performance and Clinical Utility," NPJ Digital Medicine, vol. 7, no. 1, 2024.
12. A. Moor et al., "Foundation models for generalist medical artificial intelligence," Nature, vol. 616, no. 7956, pp. 259–265,2023.
13. J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL, 2019, pp. 4171–4186.
14. A. Al-Shammari and A. Al-Ghamdi, "Secure Cloud-Based Personal Health Record System Using Firebase," IEEE Access, vol. 9, pp. 12345–12356, 2021.
15. S. Sebastian, FastAPI: A Modern, Fast (High-Performance) Web Framework for Building APIs with Python. Sebastopol, CA: O'Reilly Media, 2022.