BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Leveraging large language models to mimic domain expert labeling in unstructured text-based electronic healthcare records in non-english languages

Izzet Turkalp Akbasli[2,1*], Ahmet Ziya Birbilen[1*] and Ozlem Teksam[1]

Abstract

Background The integration of big data and artificial intelligence (AI) in healthcare, particularly through the analysis of electronic health records (EHR), presents significant opportunities for improving diagnostic accuracy and patient outcomes. However, the challenge of processing and accurately labeling vast amounts of unstructured data remains a critical bottleneck, necessitating efficient and reliable solutions. This study investigates the ability of domain specific, fine-tuned large language models (LLMs) to classify unstructured EHR texts with typographical errors through named entity recognition tasks, aiming to improve the efficiency and reliability of supervised learning AI models in healthcare.

Methods Turkish clinical notes from pediatric emergency room admissions at Hacettepe University İhsan Doğramacı Children's Hospital from 2018 to 2023 were analyzed. The data were preprocessed with open source Python libraries and categorized using a pretrained GPT-3 model, "text-davinci-003," before and after fine-tuning with domain-specific data on respiratory tract infections (RTI). The model's predictions were compared against ground truth labels established by pediatric specialists.

Results Out of 24,229 patient records classified as poorly labeled, 18,879 were identified without typographical errors and confirmed for RTI through filtering methods. The fine-tuned model achieved a 99.88% accuracy, significantly outperforming the pretrained model's 78.54% accuracy in identifying RTI cases among the remaining records. The fine-tuned model demonstrated superior performance metrics across all evaluated aspects compared to the pretrained model.

Conclusions Fine-tuned LLMs can categorize unstructured EHR data with high accuracy, closely approximating the performance of domain experts. This approach significantly reduces the time and costs associated with manual data labeling, demonstrating the potential to streamline the processing of large-scale healthcare data for AI applications.

Keywords Artificial intelligence, Large language models, Electronic healthcare records, Respiratory tract infections

*Correspondence:
Izzet Turkalp Akbasli
iakbasli@hacettepe.edu.tr
Ahmet Ziya Birbilen
ahmet.birbilen@hacettepe.edu.tr

[1]Present address: Division of Pediatric Emergency, Department of Pediatrics, Faculty of Medicine, Hacettepe University, Ankara, Turkey
[2]Life Support Center, Digital Health and Artificial Intelligence on Critical Care, Hacettepe University, Ankara, Turkey

Introduction

The healthcare industry is undergoing a transformative era, fueled by the rapid advancement of technology and the ever-increasing volume of data. The advent of Big Data (BD), characterized by its vastness, velocity, and variety, has unlocked unprecedented opportunities for healthcare providers and researchers [1–3]. In recent years, artificial intelligence (AI) applications have demonstrated significant improvements in safety, quality, and diagnostic accuracy across various clinical settings. Leveraging BD from Electronic Health Records (EHR), these AI-based techniques offer the potential to revolutionize medicine by enhancing outcomes and providing numerous benefits [2, 4–11]. However, these large-scale data are often unstructured, requiring extensive processing and labeling, which poses the most significant bottleneck [12]. In a precise field like medicine, errors in the labeling and preprocessing process can lead to poor outcomes in terms of the reliability of AI models and the impact of model results [11, 13–17]. Therefore, domain experts are often employed for labeling tasks in the present day, a process that is both time-consuming and costly [6, 18].

Moreover, when considering non-English datasets such as those in Turkish, the challenges intensify. The Turkish language presents distinct challenges for automated text analysis due to its agglutinative nature, morphological richness, and extensive inflection [19]. These linguistic characteristics can significantly complicate the extraction and classification of information from medical records, where precision is crucial [20, 21]. Addressing these challenges, our study leverages a fine-tuned large language model to effectively interpret and categorize unstructured EHR in Turkish. This adaptation not only enhances the accuracy of data processing but also contributes valuable insights into the application of AI in underrepresented linguistic contexts within the healthcare domain.

Challenges in generating datasets and data extractions from clinical notes for artificial intelligence models from BD sources containing unstructured EHR texts are primarily attributed to the complex process required for structuring and standardizing these texts for effective supervised learning AI models utilization. Unstructured EHRs are characterized by a wide array of data formats, including free-text clinical notes, laboratory findings, and imaging narratives. Each of these formats exhibits unique terminological and syntactical features, ambiguous jargon, and non-standard phrasal structures [17, 22–26]. To mitigate such complexity, the encoding of patients' diseases in EHRs using universally accepted disease classification coding systems like the International Classification of Disease (ICD) facilitates the clustering of patients, providing convenience. However, these codes can sometimes be misencoded due to intentional or unintentional information transfer by the patient or clinician [27], and these inaccuracies can significantly impact the performance of supervised learning AI models, including machine learning (ML), deep learning (DL), time series analysis, and Natural Language Processing (NLP) [26, 28, 29].

Recently, NLP methods for EHR-based computational phenotyping have seen extensive development, in information technology, the knowledge graph can transform complex unstructured data into structured form [30, 31]. Serving as a pivotal task in the construction of knowledge graphs, Named Entity Recognition (NER) enables the automatic extraction of predefined entities from extensive volumes of intricate texts, thereby facilitating the structuring of information. Through NER methods, the extraction of information from large-scale unstructured text-based datasets is substantially simplified [32–37]. However, human-induced typo errors, such as homophone, typographical, grammatical, and spacing errors, can still be present in manually entered data, with reported error rates ranging from 5 to 17% [38, 39]. These errors significantly impact the performance of NER methods [26, 40, 41]. In 2017, Google's introduction of transformer architecture marked a significant breakthrough in artificial intelligence, paving the way for the creation of advanced large language models (LLM). Trained on vast amounts of internet data using self-supervised learning techniques, these LLMs showcased an unprecedented ability to comprehend and produce text closely resembling human writing [6, 42]. Furthermore, in NER tasks, transformer-based language models have demonstrated the highest performance [26]. Unlike other LLMs, OpenAI's GPT model is used more frequently than others due to its availability through the ChatGPT interface and an API [43]. It has been demonstrated that ChatGPT can accurately predict diagnoses for patients based on clinical notes, achieving results comparable to those of human practitioners in the domain of clinical information extraction from such notes [6, 44–47].

In this research, the ability to precisely classify target labels containing typographical errors through NER tasks was explored, aiming to alleviate the detrimental effects of missing data on the efficacy of supervised learning AI models. This investigation was conducted utilizing domain specific, fine-tuned LLMs, highlighting their potential to enhance model accuracy and reliability.

Method

Data structure

In the present study, the primary data source comprised clinical notes from Pediatric Emergency Department (PED) admissions, which were extracted from the EHR system of Hacettepe University İhsan Doğramacı Children's Hospital. The structure of the data is centered

around the initial assessments conducted by pediatric residents at the PER triage point. These assessments include a variety of patient information, such as presenting complaints, evaluations based on the pediatric assessment triangle [45], body temperature, heart rate, respiratory rate, and SpO2% levels. This information forms the basis of the triage process, wherein patients are categorized for further examination. Notably, during this initial triage phase, patients are not assigned specific diagnostic codes due to the preliminary nature of the assessments. Instead, patient complaints are categorized into specific, institution-prepared complaint categories such as abdominal pain, headache, and fever, in a structured format. When the patient's presenting complaint does not align with these predefined structured categories, the physician recording the triage selects the "Others ()" category, and this input is taken as unstructured Turkish text. Consequently, this results in poorly labeled data, which poses challenges for researchers in subsequent retrospective studies.

Data collection

Data collection for this study was conducted by encompassing all patient visits to the PED at Hacettepe University İhsan Doğramacı Children's Hospital during the period from 2018 to 2023. Records were obtained from the hospital's Electronic Health Record (EHR) system, through which a systematic approach was employed to compile relevant clinical notes and assessment data.

Data preprocessing

For preprocessing tasks, open-source Python libraries such as Pandas, NLTK, and Re (regex) were utilized. Initially, poorly labeled unstructured texts categorized as "Others (Complaints)" from structured categorical diagnostic descriptions were selected, ensuring that the anonymized dataset contained only the complaints without any personal patient information. These categories were then normalized by removing the "Others ()" part to leave only 'Complaints,' and subsequently, all characters were converted to lowercase as 'complaints' using a regular expression task designed with the NLTK and Re libraries. After this normalization, the filtered data were iteratively classified syntactically using common NLP methods to determine if they contained various combinations of well-known respiratory tract infections (RTI) findings, such as fever, cough, and shortness of breath. Using a simple NER task, findings were extracted from the lowercase poorly labeled texts and subsequently categorized with rule-based methods using a series of dictionaries. Subsequently, words that could not be processed in the iteration were identified as either typographical errors or combinations of extremely rare findings. A dataset containing these poorly labeled data, which included typographical errors and those that could not be classified with simple NLP methods, was prepared for further queries using GPT.

Prompt engineering and fine-tuning of a GPT-3 model and prediction

In this study, we utilized the "text-davinci-003" model, a GPT-3 language model accessible through OpenAI's API, established as of May 2023. This low-code approach allowed us to provide preprocessed, poorly labeled data directly to the model without extensive coding requirements. Using a predefined prompt, "Based on the symptoms and findings presented, does this align with the characteristics of an RTI? If the evidence strongly suggests an RTI, please respond with 'True'. If the findings do not support an RTI diagnosis, respond with 'False'," we iteratively collected model responses. These were recorded in a Boolean list to capture the model's diagnostic alignment with the RTI characteristics. Following this initial application, the "text-davinci-003" model was fine-tuned using a specific corpus describing RTI symptoms in Turkish, enhancing its diagnostic accuracy. The fine-tuned model was then reapplied to the dataset with the same prompt to evaluate improvements in prediction accuracy.

Ground truth establishment

For the ground truth labels, four pediatric specialists were asked to determine whether the presenting complaint data, which were distributed equally and randomly among them, indicated findings of an RTI.

Model evaluation and data analysis

In the evaluation of the model outcomes, assessments were conducted using classification metrics from the Scikit Learn library, including accuracy, ROC-AUC, precision, recall, F1 score, and MCC metrics. For this project, Python version 3.9 and OpenAI's Python library version 0.26.5 were used.

Results

Between 2018 and 2023, 321,672 patients presented to the Pediatric Emergency Room (PER). In this study, 31.9% ($n = 102,732$) of the patients were determined to have RTI complaints through standard filtering methods. Subsequently, 7.53% ($n = 24,229$) of the patients were recorded in the EHR system as "Others ()", with 77.91% ($n = 18,879$) of these patients accurately identified with RTI findings through filtering methods, showing no typographical errors. Moreover, standard filtering methods revelaed that 20.2% ($n = 3,828$) of these patients had RTI. The presenting complaint targets of the remaining 22% ($n = 5,350$) were assessed as poorly labeled. These 5,350 patients received ground truth labels from four pediatric

specialists within two business days. From these labels, 16.9% ($n = 909$) were identified as RTI cases. In Table 1, the most frequent occurrences of presenting complaints containing RTI findings across the data clusters are displayed. Following the correction of errors within the unstructured poorly labeled data and typographical error-containing data cluster, the most frequently presenting complaints were, in order, control revisits, falling, diarrhea, patients sent for hospitalization from the outpatient clinic, patients with suspected COVID among upper respiratory tract infections, epistaxis, constipation, patients receiving injections, nasal discharge, and cough.

The labeling process, which was conducted by four pediatric specialists, each of whom dedicated two business days, was completed within a total of eight business days, resulting in a labeling rate of 27 labels per hour. The pretrained LLM completed the same task using a zero-shot approach in approximately six hours, with a labeling rate of 891 labels per hour. The fine-tuning process of the pretrained model, utilizing a document containing 4,724 tokens pertaining to RTI findings in Turkish, lasted approximately three hours. Similarly, employing a zero-shot approach, the fine-tuning process completed the entire labeling task in approximately six hours, akin to the performance of the pretrained model. The performances of both models were evaluated against the established ground truth labels. The pretrained model identified 714 (78.54%) patients with RTIs, and the fine-tuned model identified 908 (99.88%) patients with RTIs. The data processing stages are also demonstrated in Fig. 1, and the detailed performance metrics are available in Table 2.

Table 1 Distribution of presenting complaints by data clusters

Presenting Complaints	Structured text data	Unstructured text data ("Others ()")	Typographical errors
Total RTI Patient	98,904	3828	909
Fever	76,408	1131	246
Cough	53,866	890	143
Fatigue	19,926	33	19
Sore throat	10,897	302	47
Ear pain	9568	162	21
Respiratory Distress	4630	31	16
Non-cardiac chest pain	4077	78	8
URTI	3513	70	-
Crackles	718	249	2
Wheezing	246	5	34
COVID	9	289	423
Other RTI complaints[a]	80	1449	117
Total words in the text[b]	717,153	64,529	12,117
Total categorized label[c]	183,938	4689	1076

This table summarizes the distribution of presenting complaints from patients admitted to the PED. The complaints are categorized into structured text data, unstructured text data ("Others ()"), and typographical errors. The data includes RTI-related complaints like fever, cough, and sore throat, as well as non-respiratory issues such as ear pain and non-cardiac chest pain. The table also presents the total word count and categorized labels extracted through standard filtering methods. The structured text data contains the highest number of RTI complaints, while the unstructured category reflects poorly labeled cases, many of which were identified as RTIs after further analysis. **a**: Other RTI labels in English are: Flu, Cold, Nasal congestion, Wheezing, Rhonchi, Asthma, Croup, Bronchiolitis, Pneumonia, Febrile convulsion, Lymphadenitis, Tonsillitis, Influenza, Laryngitis, Sputum. **b**: Total words in the text: The total counts of words within the text data, segmented by data clusters. **c**: Total categorized label: The number of categorical variables that can be extracted from the content of text data through standard filtering methods

Discussion

Due to typographical errors, the categorization of unstructured text-based EHR clinical notes that cannot be classified through standard filtering and NER tasks is a costly and time-consuming process when dealing with large-scale data. As the data scale increases, it becomes imperative to automate the processes of data manipulation that require domain knowledge for more efficient supervised learning AI models. In the context of the pediatric emergency room visits where this study was conducted, nearly one-third of the patients had RTI, representing the largest patient cohort. Therefore, it is valuable to demonstrate that RTIs as presenting complaints can be recognized by LLMs. In this study, a solution to this bottleneck is presented, demonstrating that LLMs fine-tuned on a specific subject can be categorized with an accuracy approaching that of domain experts, in contrast to the general-use LLM models.

Our findings are particularly significant given the complexity of the Turkish language, which has been underrepresented in NLP research, especially in medical applications. The successful application of LLMs to Turkish EHR texts not only demonstrates the model's robustness but also its adaptability to diverse linguistic contexts. This capability is crucial for extending AI applications to non-English datasets, which are often less studied but equally in need of advanced analytical tools.

Evaluating the accuracy of LLM in medical data classification

In this study, ground truth labels were determined by pediatric specialists, and the primary focus was not a direct comparison between humans and LLMs, but rather an investigation into how closely LLMs could approximate domain expert human encoders. Accordingly, the performance of a general-use GPT-3 submodel, "text-davinci-003," resulted in 78% accuracy, while its version fine-tuned specifically for RTI findings demonstrated a significantly higher accuracy of 99.88%, surpassing that of the general model and closely matching the performance of domain experts. This efficiency and the low-code integration of the API not only expedited the research process but also minimized the potential for errors typically associated with manual coding, thus

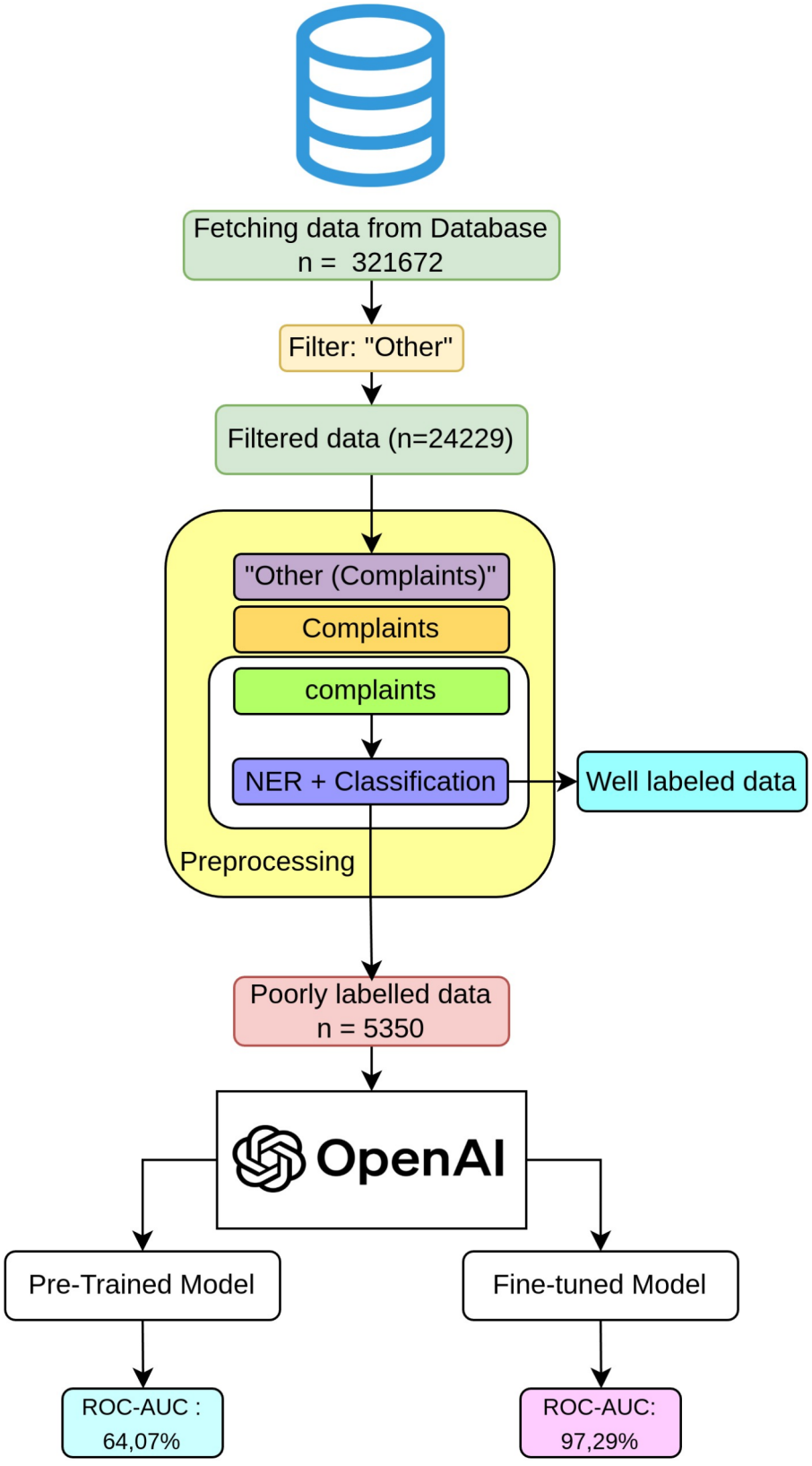

Fig. 1 Data Processing and Model Performance for RTI Identification: This figure shows the filtering of URTI symptoms from the dataset after processing all cases. It focuses on the analysis of 5,350 poorly labeled cases, comparing the ROC-AUC performance of the pretrained and fine-tuned GPT-3 models. The fine-tuned model demonstrates significant improvement in identifying RTI cases

Table 2 Comparison of performance between pretrained models and fine-tuned models

Performance Metrics	Pretrained Model (%)	Fine-tuned Model (%)
Accuracy	78.54	99.88
ROC-AUC	64.07	97.29
Precision	50.96	98.88
Recall	51.46	97.78
F1 Score	41.79	97.22
MCC	47.05	97.24

This table compares the performance metrics of the pretrained GPT-3 model and the fine-tuned model in identifying RTI cases from unstructured clinical notes. Key performance indicators, including accuracy, ROC-AUC, precision, recall, F1 score, and MCC, are provided for both models. The fine-tuned model significantly outperforms the pretrained model across all metrics, demonstrating improved diagnostic accuracy and precision. The fine-tuned model achieved an accuracy of 99.88%, with a MCC of 97.24, reflecting its enhanced capability in identifying RTI cases based on the fine-tuned Turkish RTI-specific dataset. ROC-AUC: Receiver Operating Characteristic - Area Under the Curve, MCC: Matthews Correlation Coefficient

enhancing the reliability and speed of medical data analysis. The ease of API utilization democratizes the use of sophisticated AI in clinical research, expanding the potential for broader adoption across various medical disciplines.

This finding aligns with existing literature, where comparisons between humans and LLMs, including a meta-analysis by Takita et al., revealed that the pooled accuracy of all models was 57%. Specifically, for the GPT-3 model utilized in this study, the average accuracy was reported to be 60% (range 51–69%). Additionally, model performance across specialties showed the highest efficacy in pediatric studies (93%) [48]. The above-average performance of the general-use model in our study could be attributed to this relatively high efficacy of LLMs in pediatric contexts.

In a related vein, Rosoł et al.'s study comparing humans and LLMs in medical exam questions found that the GPT-4 model, even without fine-tuning, outperformed the GPT-3 model [49]. Furthermore, the MedPaLM2 model, which is fine-tuned for the medical domain, demonstrated a high accuracy of 86.5% in Singhal et al.'s study, matching the performance of the GPT-4 model used in the study by Nori et al., which also showed an accuracy of 86.1% in USMLE exam questions [50, 51]. These findings highlight the substantial performance improvements brought about by fine-tuning, as reaffirmed in our study and supported by meta-analyses by Takita et al., where the pooled accuracy of the PaLM2 model was 43%. This underscores the significant enhancement effect of fine-tuning on model performance [48, 50, 52–56].

Moreover, another promising method for obtaining domain-specific responses through LLMs is the retrieval augmented generation (RAG) method [57], which enables a pretrained LLM to generate task-specific answers by sourcing information from specific external resources. This approach may offer an alternative solution for NER tasks [58–61]. Naik and colleagues, for instance, developed a language model that performs binary classification of clinical outcomes from EHR clinical notes using RAG methods, which have been shown to enhance answer generation performance [62, 63]. Balaguer et al.'s study comparing LLMs utilized with RAG and fine-tuning found that while the fine-tuned model produced correct answers 47% of the time, the use of RAG alone increased this to 72%, and to 74% when both were used in conjunction. The utilization of RAG methods in unstructured text-based EHR data holds significant potential for NER tasks, as demonstrated in various studies, and could provide a cost-effective alternative to solely fine-tuning models [64].

Time efficiency and cost comparison

Compared with humans, LLMs are capable of labeling both more rapidly and in a continuous, uninterrupted manner. Wang et al. demonstrated that labeling with GPT-3 is not only faster but also less expensive. Their comparison involved the GPT-3 model and human labellers on the Google Cloud Platform, where billing is based on the number of tokens. According to their findings, utilizing GPT resulted in a cost reduction of 50–96%, translating into an approximate cost of $453 for this study [65]. The work of the human encoders in this research was voluntary, with no compensation requested, and the study itself was not focused on cost analysis. However, the comparison is considered striking. Approximately $13 was spent on the labeling process in this study, including the use of a fine-tuning model that can be subsequently utilized with GPT-3. Consequently, achieving an accuracy of 98%, this method, which operates 33 times faster and can be 34 times less expensive, allows expert clinicians to allocate their time more effectively to other tasks.

Agarwal et al. work highlights the potential of using weak supervision to deploy smaller, task-specific models, thereby emphasizing the importance of models that are cost-effective and capable of generating more issue-specific responses [17]. Recently, various open-source, fine-tunable tiny language models have become available. Tasks such as those in our study can be trained on these models, significantly reducing costs through local usage while also addressing ethical concerns by enhancing local security.

Ethical consideration and data security

There are major concerns about the impact of LLMs on patient data. The large datasets used in the training of LLMs may contain sensitive patient information, which is

thought to increase the risk of data breaches or unauthorized access [66]. In this study, a language model accessed via API provided by OpenAI was used, and OpenAI's data usage policy guarantees that data used through APIs cannot be accessed or used by anyone, including model developers. This security is ensured through special protocols such as SAML SSO, SOC2, AES-256, and TLS. Some countries have their own data policies, and for this study, Turkey's personal data protection law was considered. No personal information of the patient was present in any text sent to the model via API, thus in these processes, data security is more dependent on developer compliance. Additionally, as mentioned by Agrawal et al. and Jimenez et al., security can be enhanced by using smaller, task-specific language models that can run on local systems, avoiding the use of APIs [67].

Conclusion

In conclusion, this study demonstrates that fine-tuned LLMs can effectively categorize unstructured EHR data with high accuracy, mirroring the performance of domain experts. By utilizing a fine-tuned GPT-3 model, the classification of pediatric emergency room data on respiratory tract infections achieved a remarkable accuracy of 99.88%. Notably, this performance was achieved even with data in a non-English language, highlighting the model's versatility and effectiveness. This approach significantly enhances the efficiency and cost-effectiveness of data labeling, reducing reliance on manual processes. Moreover, the successful adaptation to diverse linguistic contexts suggests a scalable model for global health systems, potentially addressing language barriers in medical data analytics. The findings underscore the potential of LLMs to streamline large-scale healthcare data processing, paving the way for more efficient and reliable AI applications in clinical settings.

Acknowledgements
Not applicable.

Author contributions
ITA conceptualized the study, developed the methodology, handled the software, validated the results, performed the formal analysis, and contributed to data curation, writing the original draft, and visualization. AZB contributed to conceptualization, validation, resources, data curation, and drafting the original manuscript. OT was responsible for the investigation, provided resources, reviewed and edited the manuscript, supervised the project, managed project administration, and acquired funding. All authors have read and approved the final manuscript.

Funding
This study did not receive any funding.

Data availability
All data produced in the present study are available upon reasonable request to the authors.

Declarations

Ethics approval and consent to participate
The Hacettepe University Clinical Research Ethics Committee approved our study's design and procedures under protocol number GO-23/508, ensuring adherence to the ethical standards in clinical research. The data sourced from Hacettepe University İhsan Doğramacı Children's Hospital, which underwent a de-identification process through the redaction of protected health information, received approval for utilization in a quality improvement project by the hospital. In this context, the Hacettepe University Research Ethics Board granted a waiver for the necessity of its approval and the procurement of informed consent for this study. Furthermore, all procedures complied with the relevant guidelines and standards outlined in the Declaration of Helsinki.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 4 March 2024 / Accepted: 14 January 2025
Published online: 31 March 2025

References

1. Saggi MK, Jain S. A survey towards an integration of big data analytics to big insights for value-creation. Inf Process Manag. 2018;54(5):758–90.
2. Pastorino R, De Vito C, Migliara G, Glocker K, Binenbaum I, Ricciardi W, et al. Benefits and challenges of Big Data in healthcare: an overview of the European initiatives. Eur J Public Health. 2019;29(Suppl 3):23–7.
3. Mishra S, Tripathy HK, Mishra BK, Sahoo S. Usage and Analysis of Big Data in E-Health Domain. In: Research Anthology on Big Data Analytics, Architectures, and Applications. IGI Global; 2022 [cited 2024 Feb 8]. pp. 417–30. Available from: https://www.igi-global.com/chapter/usage-and-analysis-of-big-data-in-e-health-domain/www.igi-global.com/chapter/usage-and-analysis-of-big-data-in-e-health-domain/290994
4. Yin J, Ngiam KY, Teo HH. Role of artificial intelligence applications in real-life clinical practice: systematic review. J Med Internet Res. 2021;23(4):e25759.
5. Bates DW, Levine D, Syrowatka A, Kuznetsova M, Craig KJT, Rui A, et al. The potential of artificial intelligence to improve patient safety: a scoping review. NPJ Digit Med. 2021;4(1):54.
6. Levine DM, Tuwani R, Kompa B, Varma A, Finlayson SG, Mehrotra A et al. The Diagnostic and Triage Accuracy of the GPT–3 Artificial Intelligence Model. medRxiv. 2023;2023.01.30.23285067.
7. Meskó B, Görög M. A short guide for medical professionals in the era of artificial intelligence. Npj Digit Med. 2020;3(1):1–8.
8. Agrawal R, Prabakaran S. Big data in digital healthcare: lessons learnt and recommendations for general practice. Heredity. 2020;124(4):525–34.
9. Matheny ME, Whicher D, Israni ST. Artificial intelligence in health care: a report from the National Academy of Medicine. JAMA. 2020;323(6):509–10.
10. Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. JAMA. 2018;319(13):1317–8.
11. Ahmed Z, Mohamed K, Zeeshan S, Dong X. Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. Database. 2020;2020:baaa010.
12. Zhou H, Albrecht MA, Roberts PA, Porter P, Della PR. Using machine learning to predict paediatric 30-day unplanned hospital readmissions: a case-control retrospective analysis of medical records, including written discharge documentation. Aust Health Rev Publ Aust Hosp Assoc. 2021;45(3):328–37.
13. Wang F, Preininger A. AI in Health: state of the art, challenges, and future directions. Yearb Med Inf. 2019;28(1):16–26.
14. Beam AL, Manrai AK, Ghassemi M. Challenges to the Reproducibility of Machine Learning Models in Health Care. JAMA. 2020;323(4):305–6.
15. Zhang P, Wang F, Hu J, Sorrentino R. Label propagation prediction of drug-drug interactions based on Clinical Side effects. Sci Rep. 2015;5:12339.
16. Curchoe CL, Flores-Saiffe Farias A, Mendizabal-Ruiz G, Chavez-Badiola A. Evaluating predictive models in reproductive medicine. Fertil Steril. 2020;114(5):921–6.
17. Agrawal M, Hegselmann S, Lang H, Kim Y, Sontag D. Large language models are few-shot clinical information extractors. In: Goldberg Y, Kozareva Z, Zhang Y, editors. Proceedings of the 2022 Conference on Empirical Methods in

Natural Language Processing. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics; 2022 [cited 2024 Feb 8]. pp. 1998–2022. Available from: https://aclanthology.org/2022.emnlp-main.130
18. Goel A, Gueta A, Gilon O, Liu C, Erell S, Nguyen LH et al. LLMs Accelerate Annotation for Medical Information Extraction. In: Proceedings of the 3rd Machine Learning for Health Symposium. PMLR; 2023 [cited 2024 Feb 8]. pp. 82–100. Available from: https://proceedings.mlr.press/v225/goel23a.html
19. Ünlütabak B, Bal O. Theory of mind performance of large language models: a comparative analysis of Turkish and English. Comput Speech Lang. 2025;89:101698.
20. Penedo G, Malartic Q, Hesslow D, Cojocaru R, Cappelli A, Alobeidli H et al. The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only. arXiv; 2023 [cited 2024 Sep 10]. Available from: http://arxiv.org/abs/2306.01116
21. Ullman T. Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks. arXiv; 2023 [cited 2024 Sep 10]. Available from: http://arxiv.org/abs/2302.08399
22. Nguyen-Dinh LV, Rossi M, Blanke U, Tröster G. Combining crowd-generated media and personal data: semi-supervised learning for context recognition. In: Proceedings of the 1st ACM international workshop on Personal data meets distributed multimedia. New York, NY, USA: Association for Computing Machinery; 2013 [cited 2024 Feb 7]. pp. 35–8. (PDM '13). Available from: https://doi.org/10.1145/2509352.2509396
23. Lake BM, Salakhutdinov R, Tenenbaum JB. Human-level concept learning through probabilistic program induction. Science. 2015;350(6266):1332–8.
24. Mozafari B, Sarkar P, Franklin M, Jordan M, Madden S. Scaling up crowd-sourcing to very large datasets: a case for active learning. Proc VLDB Endow. 2014 Ekim;8(2):125–36.
25. Qing L, Linhong W, Xuehai D. A novel neural network-based method for Medical text classification. Future Internet. 2019;11(12):255.
26. Lee EB, Heo GE, Choi CM, Song M. MLM-based typographical error correction of unstructured medical texts for named entity recognition. BMC Bioinformatics. 2022;23(1):486.
27. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring diagnoses: ICD Code Accuracy. Health Serv Res. 2005;40(5p2):1620–39.
28. Kim J, Kim T, Choi JH, Choo J. End-to-end Multi-task Learning of Missing Value Imputation and Forecasting in Time-Series Data. In: 2020 25th International Conference on Pattern Recognition (ICPR). 2021 [cited 2024 Feb 8]. pp. 8849–56. Available from: https://ieeexplore.ieee.org/document/9412112
29. Muller M, Wolf CT, Andres J, Desmond M, Joshi NN, Ashktorab Z et al. Designing Ground Truth and the Social Life of Labels. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery; 2021 [cited 2024 Feb 7]. pp. 1–16. (CHI '21). Available from: https://doi.org/10.1145/3411764.3445402
30. Murali L, Gopakumar G, Viswanathan DM, Nedungadi P. Towards electronic health record-based medical knowledge graph construction, completion, and applications: a literature study. J Biomed Inf. 2023;143:104403.
31. Sim Jah, Huang X, Horan MR, Stewart CM, Robison LL, Hudson MM, et al. Natural language processing with machine learning methods to analyze unstructured patient-reported outcomes derived from electronic health records: a systematic review. Artif Intell Med. 2023;146:102701.
32. Li I, Pan J, Goldwasser J, Verma N, Wong WP, Nuzumlalı MY, et al. Neural Natural Language Processing for unstructured data in electronic health records: a review. Comput Sci Rev. 2022;46:100511.
33. Wang Y, Afzal N, Fu S, Wang L, Shen F, Rastegar-Mojarad M, et al. MedSTS: a resource for clinical semantic textual similarity. Lang Resour Eval. 2020;54(1):57–72.
34. Zeng Z, Deng Y, Li X, Naumann T, Luo Y. Natural Language Processing for EHR-Based computational phenotyping. IEEE/ACM Trans Comput Biol Bioinform. 2019;16(1):139–53.
35. Kundeti SR, Vijayananda J, Mujjiga S, Kalyan M. Clinical named entity recognition: Challenges and opportunities. In: 2016 IEEE International Conference on Big Data (Big Data). 2016 [cited 2024 Feb 11]. pp. 1937–45. Available from: https://ieeexplore.ieee.org/abstract/document/7840814
36. Fraile Navarro D, Ijaz K, Rezazadegan D, Rahimi-Ardabili H, Dras M, Coiera E, et al. Clinical named entity recognition and relation extraction using natural language processing of medical free text: a systematic review. Int J Med Inf. 2023;177:105122.
37. Ahmad PN, Shah AM, Lee K. A review on Electronic Health Record text-mining for Biomedical Name Entity Recognition in Healthcare Domain. Healthcare. 2023;11(9):1268.
38. Hersh WR, Campbell EM, Malveau SE. Assessing the feasibility of large-scale natural language processing in a corpus of ordinary medical records: a lexical analysis. Proc Conf Am Med Inform Assoc AMIA Fall Symp. 1997;580–4.
39. Zhou L, Mahoney LM, Shakurova A, Goss F, Chang FY, Bates DW et al. How Many Medication Orders are Entered through Free-text in EHRs? - A Study on Hypoglycemic Agents. AMIA Annu Symp Proc. 2012;2012:1079–88.
40. Hamdi A, Pontes EL, Sidere N, Coustaty M, Doucet A. In-depth analysis of the impact of OCR errors on named entity recognition and linking. Nat Lang Eng. 2023;29(2):425–48.
41. Fetahu B, Chen Z, Kar S, Rokhlenko O, Malmasi S. arXiv.org. 2023 [cited 2024 Feb 11]. MultiCoNER v2: a Large Multilingual dataset for Fine-grained and Noisy Named Entity Recognition. Available from: https://arxiv.org/abs/2310.13213v1
42. Yang R, Tan TF, Lu W, Thirunavukarasu AJ, Ting DSW, Liu N. Large language models in health care: development, applications, and challenges. Health Care Sci. 2023;2(4):255–63.
43. Coello CEA, Alimam MN, Kouatly R. Effectiveness of ChatGPT in Coding: a comparative analysis of Popular large Language models. Digital. 2024;4(1):114–25.
44. Knebel D, Priglinger S, Scherer N, Siedlecki J, Schworm B. Assessment of ChatGPT in the preclinical management of ophthalmological emergencies–an analysis of ten fictional case vignettes. medRxiv; 2023 [cited 2024 Feb 8]. p. 2023.04.16.23288645. Available from: https://www.medrxiv.org/content/https://doi.org/10.1101/2023.04.16.23288645v1
45. Nastasi AJ, Courtright KR, Halpern SD, Weissman GE. Does ChatGPT Provide Appropriate and Equitable Medical Advice? A Vignette-Based, Clinical Evaluation Across Care Contexts. medRxiv; 2023 [cited 2024 Feb 8]. p. 2023.02.25.23286451. Available from: https://www.medrxiv.org/content/https://doi.org/10.1101/2023.02.25.23286451v1
46. Rao A, Pang M, Kim J, Kamineni M, Lie W, Prasad AK et al. Assessing the Utility of ChatGPT Throughout the Entire Clinical Workflow. medRxiv; 2023 [cited 2024 Feb 8]. p. 2023.02.21.23285886. Available from: https://www.medrxiv.org/content/https://doi.org/10.1101/2023.02.21.23285886v1
47. Fraser H, Crossland D, Bacher I, Ranney M, Madsen T, Hilliard R. Comparison of Diagnostic and Triage Accuracy of Ada Health and WebMD Symptom checkers, ChatGPT, and Physicians for patients in an Emergency Department: Clinical Data Analysis Study. JMIR MHealth UHealth. 2023;11(1):e49995.
48. Takita H, Walston SL, Tatekawa H, Saito K, Tsujimoto Y, Miki Y et al. Diagnostic Performance of Generative AI and Physicians: A Systematic Review and Meta-Analysis. medRxiv; 2024 [cited 2024 Feb 11]. p. 2024.01.20.24301563. Available from: https://www.medrxiv.org/content/https://doi.org/10.1101/2024.01.20.24301563v1
49. Rosoł M, Gąsior JS, Łaba J, Korzeniewski K, Młyńczak M. Evaluation of the performance of GPT–3.5 and GPT–4 on the Medical Final Examination. medRxiv; 2023 [cited 2024 Feb 10]. p. 2023.06.04.23290939. Available from: https://www.medrxiv.org/content/https://doi.org/10.1101/2023.06.04.23290939v2
50. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of gpt–4 on medical challenge problems. ArXiv Prepr ArXiv230313375. 2023.
51. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L et al. Towards Expert-Level Medical Question Answering with Large Language Models. arXiv; 2023 [cited 2024 Feb 12]. Available from: http://arxiv.org/abs/2305.09617
52. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv; 2023 [cited 2024 Feb 12]. Available from: http://arxiv.org/abs/2307.09288
53. Tian S, Jin Q, Yeganova L, Lai PT, Zhu Q, Chen X, et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. Brief Bioinform. 2024;25(1):bbad493.
54. Johnson D, Goodman R, Patrinely J, Stone C, Zimmerman E, Donald R et al. Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model. Res Sq. 2023.
55. Latif E, Zhai X. Fine-tuning chatgpt for automatic scoring. Comput Educ Artif Intell. 2024;100210.
56. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172–80.
57. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. Adv Neural Inf Process Syst. 2020;33:9459–74.
58. Guu K, Lee K, Tung Z, Pasupat P, Chang M. Retrieval Augmented Language Model Pre-Training. In: International Conference on Machine Learning. PMLR; 2020 [cited 2024 Feb 12]. pp. 3929–38. Available from: https://proceedings.mlr.press/v119/guu20a.html

59. Cuconasu F, Trappolini G, Siciliano F, Filice S, Campagnano C, Maarek Y et al. The Power of Noise: Redefining Retrieval for RAG Systems. arXiv; 2024 [cited 2024 Feb 12]. Available from: http://arxiv.org/abs/2401.14887
60. Zhang L, Jijo K, Setty S, Chung E, Javid F, Vidra N et al. Enhancing Large Language Model Performance To Answer Questions and Extract Information More Accurately. arXiv; 2024 [cited 2024 Feb 12]. Available from: http://arxiv.org/abs/2402.01722
61. Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. Brief Bioinform. 2022;23(6):bbac409.
62. Naik A, Parasa S, Feldman S, Wang LL, Hope T. Literature-Augmented Clinical Outcome Prediction. arXiv; 2022 [cited 2024 Feb 12]. Available from: http://arxiv.org/abs/2111.08374
63. Zakka C, Shad R, Chaurasia A, Dalal AR, Kim JL, Moor M, et al. Almanac — Retrieval-Augmented Language models for Clinical Medicine. NEJM AI. 2024;1(2):AIoa2300068.
64. Balaguer A, Benara V, Cunha RL, de Filho F, de Hendry R, Holstein T. D, RAG vs Fine-tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture. arXiv; 2024 [cited 2024 Feb 12]. Available from: http://arxiv.org/abs/2401.08406
65. Wang S, Liu Y, Xu Y, Zhu C, Zeng M. Want To Reduce Labeling Cost? GPT–3 Can Help. arXiv; 2021 [cited 2024 Feb 11]. Available from: http://arxiv.org/abs/2108.13487
66. Gutiérrez BJ, McNeal N, Washington C, Chen Y, Li L, Sun H et al. Thinking about GPT–3 In-Context Learning for Biomedical IE? Think Again. arXiv; 2022 [cited 2024 Jun 9]. Available from: http://arxiv.org/abs/2203.08410
67. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med. 2023;6:120.

Publisher's note