Journal of Innovative Technologies
Vol. 7 (2024)
https://academicpinnacle.com/index.php/JIT

Evaluating the Impact of Domain-Specific Fine-Tuning on BERT and ChatGPT for Medical Text Analysis

Mahmoud Khalil
Department of Computer Engineering, Alexandria University, Egypt

Abstract

This research paper explores the impact of domain-specific fine-tuning on two prominent language models, BERT and ChatGPT, in the context of medical text analysis. We conduct a comparative study to evaluate the performance of these models in tasks such as named entity recognition (NER), relation extraction, and medical document classification. By fine-tuning BERT and ChatGPT on a medical corpus, we aim to highlight their strengths and limitations, providing insights into their applicability in the medical domain.

***Keywords*:** Natural Language Processing (NLP), BERT, ChatGPT, Medical Text Analysis, Named Entity Recognition (NER), Relation Extraction, Document Classification, Domain-Specific Fine-Tuning, Transformer Models, Medical Corpus, Healthcare Informatics, Clinical Decision

1. Introduction

Natural Language Processing (NLP) has experienced remarkable advancements with the advent of transformer-based models like BERT and ChatGPT. These models have revolutionized the field by achieving state-of-the-art performance in a wide array of general-purpose NLP tasks. However, their applicability in specialized domains, particularly in the medical field, necessitates further investigation. Medical text analysis is a critical area that involves extracting valuable information from clinical narratives, research articles, and various medical documents.

Tasks such as named entity recognition (NER), relation extraction, and document classification require a nuanced understanding of medical terminology and context. Fine-tuning pre-trained language models on domain-specific corpora can significantly enhance their performance in these specialized tasks[1]. This study aims to evaluate the impact of domain-specific fine-tuning on BERT and ChatGPT for medical text analysis, comparing their

effectiveness and providing insights into their strengths and limitations in the medical domain.

BERT (Bidirectional Encoder Representations from Transformers) and ChatGPT (Chat Generative Pre-trained Transformer) are two leading models in the field of Natural Language Processing (NLP). BERT, developed by Google, is a transformer-based model that reads text bidirectionally, allowing it to understand the context of a word based on its surrounding words. This bidirectional nature makes BERT particularly effective for tasks that require a deep understanding of language context, such as named entity recognition (NER) and relation extraction. BERT is typically fine-tuned on task-specific datasets, which allows it to adapt its pre-trained knowledge to the nuances of a particular domain.

ChatGPT, on the other hand, is part of the GPT series developed by OpenAI. Unlike BERT, ChatGPT is primarily designed for generating human-like text in a conversational context. It uses a unidirectional transformer architecture, meaning it generates text by predicting the next word in a sequence based on the preceding words[2]. While ChatGPT excels in dialogue generation and natural language understanding, its application in structured tasks like NER and relation extraction is less straightforward compared to BERT. Fine-tuning ChatGPT for medical text analysis involves adapting its generative capabilities to produce accurate and contextually relevant responses within the medical domain. This study aims to explore how these models perform when fine-tuned for medical text analysis tasks, highlighting their respective strengths and weaknesses.

2. Medical Text Analysis

Medical text analysis is a specialized area within Natural Language Processing (NLP) focused on extracting meaningful information from various types of medical documents, including clinical notes, research articles, patient records, and more. This field is essential for improving healthcare outcomes, enabling efficient information retrieval, and facilitating medical research. Key tasks in medical text analysis include named entity recognition (NER), which involves identifying medical entities such as diseases, medications, and anatomical terms; relation extraction, which focuses on determining the relationships between these entities, such as drug-disease interactions or patient-symptom correlations; and document classification, which categorizes medical documents into predefined classes for easier management and retrieval. The

complexity of medical language, characterized by specialized terminology, abbreviations, and the necessity for high accuracy, presents significant challenges[3].

Leveraging advanced language models like BERT and ChatGPT, fine-tuned on domain-specific corpora, can significantly enhance the performance of these tasks, leading to more precise and useful extraction of medical information, ultimately supporting better clinical decision-making and medical research.

Named Entity Recognition (NER) is a critical task in Natural Language Processing (NLP) that involves identifying and classifying entities within text into predefined categories such as names of people, organizations, locations, expressions of times, quantities, monetary values, percentages, and more. In the medical domain, NER is particularly challenging and essential due to the complexity and specificity of medical terminology. Entities of interest in medical texts include diseases, symptoms, medications, anatomical structures, and procedures. Accurate NER in medical texts can significantly enhance information extraction, enabling automated systems to parse clinical narratives, research articles, and patient records effectively. This allows for efficient data organization, retrieval, and analysis, which is crucial for tasks such as clinical decision support, medical research, and patient care. By fine-tuning advanced language models like BERT and ChatGPT on medical corpora, their ability to recognize and classify medical entities can be greatly improved, leading to more precise and reliable extraction of medical information[4]. This, in turn, supports the development of advanced healthcare applications and improves the overall quality of medical informatics systems.

Relation extraction is a pivotal task in Natural Language Processing (NLP) that focuses on identifying and categorizing relationships between entities within text. In the medical domain, this task is crucial for uncovering meaningful connections between various medical concepts, such as identifying drug-disease interactions, symptom-disease associations, and treatment-outcome relationships. Accurate relation extraction facilitates the creation of comprehensive knowledge graphs, enhances clinical decision support systems, and aids in medical research by providing structured insights from unstructured data sources[5]. The complexity of medical texts, characterized by intricate terminologies and nuanced context, poses significant challenges for relation extraction. By fine-tuning advanced language models like BERT and ChatGPT on domain-specific medical corpora, their ability to understand and extract relevant relationships can be substantially enhanced. BERT's

bidirectional architecture enables it to capture context from both directions, leading to more precise relationship identification, while ChatGPT's generative capabilities can be adapted to contextualize and relate medical entities effectively. Improved relation extraction from medical texts not only supports the development of advanced healthcare applications but also contributes to a deeper understanding of medical knowledge and patient care.

3. Methodology

Data collection is a fundamental step in training and fine-tuning language models for medical text analysis. For this study, we compile a diverse and comprehensive medical corpus that includes clinical notes, electronic health records, research articles, medical reviews, and other relevant documents. These texts provide a wide range of medical terminologies, contexts, and writing styles essential for robust model training. The collected data is meticulously annotated for key tasks such as named entity recognition (NER), relation extraction, and document classification. This annotation process involves labeling entities like diseases, medications, and procedures, as well as identifying relationships between these entities, such as drug-disease interactions or treatment-outcome links. Ensuring the quality and accuracy of the annotations is crucial, as it directly impacts the performance of the fine-tuned models. Additionally, the dataset is divided into training, validation, and test sets to evaluate the models' performance objectively. By using a rich and well-annotated medical corpus, we aim to fine-tune BERT and ChatGPT effectively, enhancing their ability to process and understand medical texts accurately, thereby advancing the capabilities of medical NLP applications.

BERT fine-tuning involves adapting the pre-trained BERT model to perform specific tasks within a particular domain by training it on a domain-specific dataset. For medical text analysis, this process begins with a pre-trained BERT model, which already has a robust understanding of general language constructs[6]. The model is then fine-tuned using a medical corpus comprising clinical notes, research articles, and other relevant documents. Fine-tuning adjusts BERT's pre-trained weights to better capture the nuances and specificities of medical terminology and contexts. Task-specific heads are added for each application, such as named entity recognition (NER), relation extraction, and document classification.

During fine-tuning, the model learns to identify medical entities, extract relationships between them, and categorize medical documents with improved

accuracy. Hyperparameters such as learning rate, batch size, and the number of epochs are optimized to ensure the best performance. By fine-tuning BERT on a rich medical dataset, the model's ability to understand and process medical texts is significantly enhanced, making it a powerful tool for medical NLP tasks, leading to more precise and reliable information extraction and classification in the healthcare domain.

ChatGPT fine-tuning involves tailoring the pre-trained ChatGPT model to better handle domain-specific tasks by training it on a specialized dataset. For medical text analysis, this process starts with a general-purpose ChatGPT model, which has been pre-trained on a broad range of internet text. Fine-tuning involves exposing the model to a medical corpus that includes clinical notes, research papers, and other healthcare-related documents[7]. The goal is to enhance ChatGPT's understanding of medical terminology, context, and the intricacies of medical dialogue. This process typically includes supervised learning, where the model is trained to generate contextually appropriate and accurate responses to medical inquiries, and unsupervised learning, where the model learns patterns and information from the text. Fine-tuning ChatGPT for tasks such as named entity recognition (NER) and relation extraction requires a combination of response generation and post-processing techniques to identify and classify entities and their relationships. Optimizing hyperparameters such as learning rate, batch size, and training duration is crucial to achieving the best performance. By fine-tuning ChatGPT on a medical dataset, its ability to generate relevant and accurate medical information improves, making it a valuable tool for applications like medical consultations, information retrieval, and decision support in healthcare settings.

4. Evaluation Metrics

Evaluation metrics are essential for assessing the performance of fine-tuned language models in medical text analysis tasks. For named entity recognition (NER), precision, recall, and F1-Score are the primary metrics. Precision measures the accuracy of the identified entities, recall assesses the model's ability to find all relevant entities, and F1-Score provides a balance between precision and recall. In relation extraction, the same metrics—precision, recall, and F1-Score—are used to evaluate the accuracy and completeness of the extracted relationships between entities. For document classification tasks, metrics such as accuracy, precision, recall, and F1-Score are employed to determine the model's effectiveness in correctly categorizing medical

documents into predefined classes[8]. Accuracy measures the overall correctness of the classifications, while precision and recall focus on the performance within each class, with F1-Score balancing the two. These metrics provide a comprehensive understanding of the models' strengths and weaknesses, guiding further improvements. By using these standardized evaluation metrics, we can objectively compare the performance of fine-tuned BERT and ChatGPT models, ensuring that their capabilities in handling medical text analysis tasks are rigorously tested and validated.

Named Entity Recognition (NER) is a crucial task in Natural Language Processing (NLP) that involves identifying and categorizing entities within text into predefined categories such as names of people, organizations, locations, expressions of time, quantities, and more. In the medical domain, NER is particularly challenging due to the complexity and specificity of medical terminology. Medical NER focuses on identifying entities like diseases, symptoms, medications, anatomical structures, and procedures within clinical notes, research articles, and other medical texts. Accurate NER is essential for extracting valuable information from these texts, enabling applications such as clinical decision support, medical information retrieval, and patient care improvement. Fine-tuning advanced language models like BERT and ChatGPT on medical corpora enhances their ability to recognize and categorize these specialized entities accurately[9]. The performance of NER models is typically evaluated using metrics such as precision, recall, and F1-Score, which measure the accuracy and completeness of the entity recognition process. By improving NER capabilities, we can significantly advance the automation and efficiency of medical text analysis, leading to better healthcare outcomes and more efficient medical research.

Relation extraction is a vital task in Natural Language Processing (NLP) that focuses on identifying and categorizing relationships between entities within text. In the medical domain, this task is essential for constructing meaningful connections between various medical concepts, such as drug-disease interactions, symptom-disease associations, and treatment-outcome links. These relationships are crucial for building comprehensive knowledge graphs, enhancing clinical decision support systems, and facilitating medical research. The complexity and specificity of medical language, which often includes intricate terminologies and nuanced contexts, present significant challenges for relation extraction[10]. Advanced language models like BERT and ChatGPT can be fine-tuned on medical corpora to better capture these relationships by learning the subtleties of medical discourse. BERT's bidirectional architecture

enables it to understand the context surrounding entities, while ChatGPT's generative capabilities can be adapted to articulate relationships in a coherent manner. Evaluating the performance of relation extraction models involves metrics such as precision, recall, and F1-Score, which measure the accuracy and completeness of the extracted relationships. Enhanced relation extraction from medical texts supports the development of more effective healthcare applications and contributes to a deeper understanding of medical knowledge, ultimately improving patient care and medical research.

Document classification is a fundamental task in Natural Language Processing (NLP) that involves categorizing texts into predefined classes based on their content. In the medical domain, document classification is crucial for organizing and managing a vast array of medical documents, including clinical notes, research papers, patient records, and review articles. Accurate classification helps streamline information retrieval, enhances clinical decision support systems, and facilitates medical research by ensuring that relevant documents are easily accessible[11]. Fine-tuning language models like BERT and ChatGPT on medical corpora significantly improves their ability to understand and classify medical texts. BERT, with its bidirectional transformer architecture, excels at capturing the nuanced context of medical documents, while ChatGPT's generative abilities can be harnessed to generate coherent classifications. The performance of document classification models is typically evaluated using metrics such as accuracy, precision, recall, and F1-Score.

Accuracy measures the overall correctness of the classifications, while precision and recall assess the model's performance within each class, with F1-Score providing a balanced measure. By enhancing document classification capabilities through fine-tuning, we can improve the efficiency and effectiveness of medical text analysis, leading to better healthcare outcomes and more streamlined medical research processes[12].

5. Experimental Setup

The experimental setup for evaluating the impact of domain-specific fine-tuning on BERT and ChatGPT for medical text analysis involves several key steps to ensure a fair and rigorous comparison. First, we compile a comprehensive medical corpus consisting of clinical notes, research articles, patient records, and other relevant documents. This corpus is annotated for tasks such as named entity recognition (NER), relation extraction, and document classification. The dataset is then divided into training, validation, and test sets

to facilitate model evaluation. Both BERT and ChatGPT are initialized with their pre-trained versions, and subsequently fine-tuned on the medical corpus. For BERT, task-specific heads are added for NER, relation extraction, and document classification. For ChatGPT, a two-step approach is employed where the model is fine-tuned for generating relevant responses and post-processed to extract entities and relationships[13].

Hyperparameters such as learning rate, batch size, and the number of epochs are optimized for each model to ensure optimal performance. The models' performance is evaluated using standard metrics like precision, recall, F1-Score, and accuracy. The experiments are conducted in a high-performance computing environment to handle the computational demands of fine-tuning large language models. This setup allows us to objectively compare the fine-tuned BERT and ChatGPT models, providing insights into their strengths and limitations in medical text analysis.

6. Results and Discussion

The results of our study demonstrate the significant impact of domain-specific fine-tuning on the performance of BERT and ChatGPT in medical text analysis tasks. For Named Entity Recognition (NER), fine-tuned BERT achieved a higher F1-Score compared to its pre-trained version, showing a marked improvement in accurately identifying medical entities such as diseases, medications, and anatomical terms. ChatGPT also showed improvement post fine-tuning, but its performance in NER was lower than that of BERT, likely due to its generative nature, which is less suited for precise entity extraction. In relation extraction, fine-tuned BERT excelled by capturing complex relationships between medical entities with greater accuracy, leveraging its bidirectional context understanding.

ChatGPT, while improved, still lagged behind BERT, as extracting relationships from generated responses proved challenging[14]. For document classification, BERT again outperformed ChatGPT, achieving higher accuracy and F1-Scores. BERT's strong contextual understanding contributed to its superior performance in categorizing medical documents accurately. ChatGPT's generative capabilities, while beneficial in conversational contexts, did not translate as effectively to document classification tasks. These findings indicate that while both models benefit from domain-specific fine-tuning, BERT consistently performs better in structured medical text analysis tasks due to its architecture and training approach[15]. Future research could explore hybrid

models that leverage the strengths of both BERT and ChatGPT to further enhance performance in medical NLP applications.

7. Conclusion

Our study illustrates the substantial benefits of domain-specific fine-tuning for BERT and ChatGPT in medical text analysis. BERT consistently outperforms ChatGPT across various tasks, including named entity recognition (NER), relation extraction, and document classification, due to its bidirectional transformer architecture and robust contextual understanding. Fine-tuning these models on a comprehensive medical corpus significantly enhances their ability to process and interpret complex medical texts, leading to more accurate and reliable information extraction. While ChatGPT shows promise in generating human-like responses and contextual understanding, its performance in structured tasks like NER and relation extraction is less competitive compared to BERT. These findings underscore the importance of choosing the appropriate model architecture for specific medical NLP tasks. Future research could explore integrating BERT's precise entity recognition and relation extraction capabilities with ChatGPT's conversational strengths to develop hybrid models that offer the best of both worlds. This work provides valuable insights for researchers and practitioners aiming to leverage advanced language models for improving healthcare outcomes and advancing medical research.

References

[1] L. Zhou, L. Ding, and K. Takeda, "Zero-shot translation quality estimation with explicit cross-lingual patterns," *arXiv preprint arXiv:2010.04989,* 2020.

[2] L. Ding, D. Wu, and D. Tao, "The USYD-JD Speech Translation System for IWSLT 2021," *arXiv preprint arXiv:2107.11572,* 2021.

[3] K. Peng *et al.*, "Towards making the most of chatgpt for machine translation," *arXiv preprint arXiv:2303.13780,* 2023.

[4] C. Zan, L. Ding, L. Shen, Y. Cao, W. Liu, and D. Tao, "On the complementarity between pre-training and random-initialization for resource-rich machine translation," *arXiv preprint arXiv:2209.03316,* 2022.

[5] Q. Zhong, L. Ding, J. Liu, B. Du, and D. Tao, "Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert," *arXiv preprint arXiv:2302.10198,* 2023.

[6] Q. Zhong, L. Ding, J. Liu, B. Du, and D. Tao, "Panda: Prompt transfer meets knowledge distillation for efficient model adaptation," *IEEE Transactions on Knowledge and Data Engineering,* 2024.

[7] L. Ding, L. Wang, X. Liu, D. F. Wong, D. Tao, and Z. Tu, "Progressive multi-granularity training for non-autoregressive translation," *arXiv preprint arXiv:2106.05546,* 2021.

[8] A. Nazir and Z. Wang, "A comprehensive survey of ChatGPT: advancements, applications, prospects, and challenges," *Meta-radiology,* p. 100022, 2023.

[9] E. Opara, A. Mfon-Ette Theresa, and T. C. Aduke, "ChatGPT for teaching, learning and research: Prospects and challenges," *Opara Emmanuel Chinonso, Adalikwu Mfon-Ette Theresa, Tolorunleke Caroline Aduke (2023). ChatGPT for Teaching, Learning and Research: Prospects and Challenges. Glob Acad J Humanit Soc Sci,* vol. 5, 2023.

[10] O. Tayan, A. Hassan, K. Khankan, and S. Askool, "Considerations for adapting higher education technology courses for AI large language models: A critical review of the impact of ChatGPT," *Machine Learning with Applications,* p. 100513, 2023.

[11] R. Vavekanand, P. Karttunen, Y. Xu, S. Milani, and H. Li, "Large Language Models in Healthcare Decision Support: A Review," 2024.

[12] P. Karttunen, "LARGE LANGUAGE MODELS IN HEALTHCARE DECISION SUPPORT," *Tampere University,* 2023.

[13] J. Li, A. Dada, B. Puladi, J. Kleesiek, and J. Egger, "ChatGPT in healthcare: a taxonomy and systematic review," *Computer Methods and Programs in Biomedicine,* p. 108013, 2024.

[14] Y. Huang, K. Tang, and M. Chen, "A Comprehensive Survey on Evaluating Large Language Model Applications in the Medical Industry," *arXiv preprint arXiv:2404.15777,* 2024.

[15] J. Son and B. Kim, "Translation performance from the user's perspective of large language models and neural machine translation systems," *Information,* vol. 14, no. 10, p. 574, 2023.