Health. Innovation. Community: It Starts With Us
J. Bichel-Findlay (Ed.)

doi:10.3233/SHTI240886

Adapting Large Language Models for Automated Summarisation of Electronic Medical Records in Clinical Coding

Bokang BI [a,1], Leibo LIU [a] and Oscar PEREZ-CONCHA [a]
[a] *Centre for Big Data Research in Health, The University of New South Wales, Sydney, Australia*
ORCiD ID: Bokang Bi https://orcid.org/0009-0000-4301-6123, Leibo Liu https://orcid.org/0000-0002-7517-0050, Oscar Perez-Concha https://orcid.org/0000-0002-8823-7090

Abstract: Encapsulating a patient's clinical narrative into a condensed, informative summary is indispensable to clinical coding. The intricate nature of the clinical text makes the summarisation process challenging for clinical coders. Recent developments in large language models (LLMs) have shown promising performance in clinical text summarisation, particularly in radiology and echocardiographic reports, after adaptation to the clinical domain. To explore the summarisation potential of clinical domain adaptation of LLMs, a clinical text dataset, consisting of electronic medical records paired with "Brief Hospital Course" from the MIMIC-III database, was curated. Two open-source LLMs were then fine-tuned, one pre-trained on biomedical datasets and another on a general-content domain on the curated clinical dataset. The performance of the fine-tuned models against their base models were evaluated. The model pre-trained on biomedical data demonstrated superior performance after clinical domain adaptation. This finding highlights the potential benefits of adapting LLMs pre-trained on a related domain over a more generalised domain and suggests the possible role of clinically adapted LLMs as an assistive tool for clinical coders. Future work should explore adapting more advanced models to enhance model performance in higher-quality clinical datasets.

Keywords. summarisation, clinical coding, fine-tuning, large language model

1. Introduction

Clinical text summarisation for clinical coding involves shortening healthcare narratives into a condensed and informative version [1]. This is a challenging task as hospital electronic medical records (EMRs) document a patient's encounter in multiple, time-ordered linked files across a variety of categories, including progress notes, radiology reports, discharge summaries, and more. Numerous healthcare professionals write these documents, therefore making navigation and comprehension of EMRs challenging [2-6]. According to statistics, on average, only 10% of the text in a patient's EMR is relevant to the coding task, while the rest is redundant and potentially misleading, making manual coding time-consuming and prone to errors [7]. Therefore, an automated clinical text summarisation method that can generate a concise summary, highlighting valuable

[1] Corresponding Author: Bokang Bi, akatsukirin0205@gmail.com.

information for clinical coders from vast documents in EMRs for their coding task, is much needed.

In recent years, the development of large language models (LLMs) has demonstrated extraordinary capabilities in various natural language processing (NLP) tasks [8-10]. Previous studies on clinical text summarisation of chest X-ray reports [1, 2, 11] and echocardiography reports [12] have demonstrated domain adaptation of LLMs through fine-tuning. Recent works on generating the "Brief Hospital Course" (BHC) section in discharge summaries using Bidirectional Representations from Transformers (BERT) and Bidirectional and Auto-Regressive Transformers (BART) [4, 13] revealed that summarising a patient's hospital course from EMRs is challenging due to the lengthy content and intricate variability across contributors and document structures in EMRs, which has made it difficult for models to capture the complex clinical structure and terminologies required for effective clinical summarisation [1, 2, 4, 14].

This study investigated the potential for domain-specific adaptation of LLMs pre-trained general domain knowledge and subsequently pre-trained related biomedical domain knowledge. Adaptation was achieved by fine-tuning the curated dataset from MIMIC-III [15-17]. It was hypothesised that LLMs pre-trained on data from related domains would exhibit better performance than LLMs pre-trained on general domain data, as pre-training on related domains equips the models with a deeper understanding of specialised terminology and document structure in clinical text.

2. Methods

2.1 Dataset

This study collected de-identified free-text clinical notes from the MIMIC-III database [15-17]. This dataset consists of 2,083,180 distinct clinical notes from various categories, including healthcare provider reports and discharge summaries, from 53,423 admission events involving 38,597 patients in the Intensive Care Units of Beth Israel Deaconess Medical Center, USA, between 2001 and 2012 [15]. To enhance the fine-tuning process of the LLMs, extensive pre-processing and cleaning were undertaken to optimise the LLM's learning efficiency. A labelled dataset of 18,316 EMR and BHC pairs was constructed, randomly divided into 85%, 10% and 5% for model training, validation and test set, respectively.

From the discharge summaries in the MIMIC III dataset, regular expressions were applied to extract the BHC section as the true labels. The following sections were also extracted: Chief Complaint, Major Procedure, History of Present Illness, Physical Exam, Discharge Diagnosis and Discharge Disposition as these sections contain valuable information for the BHC summarisation [23].

Table 1. The concatenated multi-document EMR structure in the curated clinical dataset from MIMIC III, as input for model training.

Admission day: Reason of hospitalisation is {*Chief Complaints*}. History of present illness: {*History of the Present Illness*}.
Day 1: Extracted sections from EMR
Day 2: Extracted sections from EMR
.....
Discharge day: Patient physical examination {*Physical Exams*}. Patient is diagnosed {*Discharge diagnosis*}, received {*Major Procedure*} in hospital. Patient is discharged to {*Discharge Disposition*}.

The EMR in the dataset is constructed to capture the daily narrative of a patient's hospital stay, using regular expressions to extract representative sections from multiple clinical documents generated during the patient's hospital course. These extracted texts were concatenated in chronological order from the admission day to the discharge day. Special characters used for de-identification were replaced with artificial identifiers through pseudonymisation. For example, '*Ms. [**Known lastname**]'* is replaced with '*the patient*', and all the dates in the format of *'[**2119-01-16**]'* is replaced by '*Day X*' where 'X' represents the number of days since the admission date. The admission date itself is designated as *'Day 1'*. Additionally, the format of the EMRs was standardised by removing all the excessive existing line breaks and using new line breaks to separate each daily narrative.

2.2 Model Selection and Adaptation

In this study, domain-specific adaptation was explored via fine-tuning of BioMistral, currently (as of 14/03/2024) the best-performing open-source auto-regressive model at the 7B scale, which is pre-trained on bio-medical data [19], and Llama2 7B, which has demonstrated success in medical domain adaptation in several previous studies [11, 12, 18]. Large language models at the 7B scale were chosen, as they can be fine-tuned with constrained computational resources, and inference tasks are feasible on consumer-grade GPUs. Proprietary models such as GPT-turbo and Gemini were excluded from this study because adapting these models required uploading confidential health data to a central server [12], which violated privacy regulations [15].

Parameter-efficient fine-tuning (PEFT) has been proven to increase the performance of clinical text summarisation for adapted pre-trained large language models [3, 11, 12]. Quantised Low Rank Adaptation (QLoRA) was used to fine-tune the selected open-source LLM. QLoRA achieves comparable fine-tuning results to standard Low Rank Adaptation (LoRA) but significantly reduces the computational resources required through the 4-bit quantisation of the LLM [20]. Zero-shot prompting was used to combine the input EMR in the dataset with a specific prompt format and parsed it to selected models for BHC generation. This served as a baseline performance to evaluate the fine-tuned model performance.

In-context learning (ICL), another popular adaption method, is unsuitable for this study due to the maximum input token length limitation. Furthermore, previous studies have demonstrated that QLoRA results in better improvements in LLM performance compared to ICL [2, 12].

2.3 Evaluation

The model performance was evaluated using three metrics from the Recall-Oriented Understudy for Gisting Evaluation [21] (ROUGE): ROUGE-L, ROUGE-1 and ROUGE-2 for syntactic overlapping, and BERT-Score [22] for semantic similarity through BERT embedding.

3. Results

Both adapted models displayed better performance compared to their base models and demonstrated similar performance improvements after fine-tuning across all four metrics. In domain-adapted models, BioMistral demonstrated superior performance compared to the Llama2 model on all evaluated metrics after fine-tuning, particularly on the ROUGE metrics.

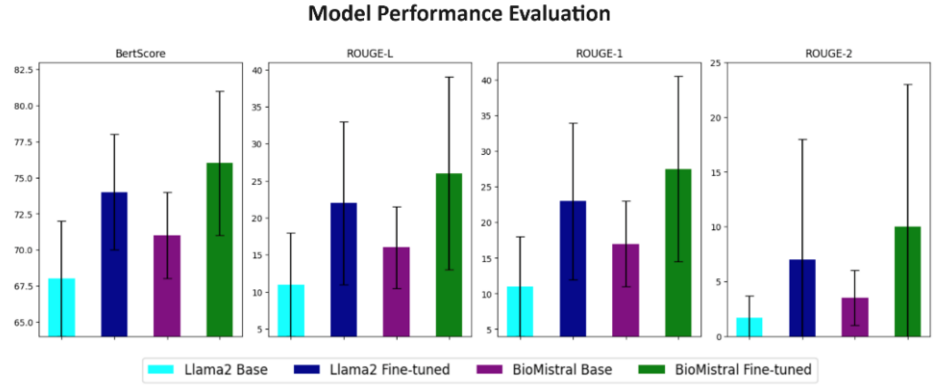

Figure 1. Different model performance across Bert-Score and Rouge metrics.

4. Discussion

The experiments revealed that the improvement in clinical text summarisation performance of domain-adapted LLMs depends on the performance of the base model. LLMs pre-trained on data closely related to the specific domain targeted for adaptation exhibited better performance after fine-tuning adaptation. This finding highlights the potential benefits of selecting LLMs pre-trained on biomedical or clinical data for further adaptation for downstream tasks in clinical applications. BioMistral, currently the best-performing open-source LLM pre-trained on biomedical data, demonstrated greater superiority in ROUGE metrics than the BERT-Score. ROUGE-1 measures the overlap of individual words, thereby assessing the coverage of specific medical jargon, such as treatment or diagnosis. The ROUGE-L measures how well sequential information is maintained via the longest common subsequence, reflecting the accuracy of chronological events in clinical narratives. This suggests that the BioMistral model is better at generating summaries that are more closely aligned with those written by physicians due to its pre-training on biomedical data. This in turn enables the BioMistral model to effectively capture the specialised terminology and clinical narrative structure in the fine-tuning dataset. Conversely, the Llama2 model, pre-trained on generalised data, is less capable of adapting to the medical jargon and underlying narrative structures of clinical EMRs during fine-tuning. This observation supports the hypothesis that pre-training in a specific domain significantly enhances a LLM's performance when subsequently fine-tuned to tasks in related domains.

The study demonstrated that domain adaptation through fine-tuning is an effective method for significantly enhancing the performance of pre-trained LLMs in clinical NLP tasks, both syntactically and semantically. A similar result is also illustrated in a benchmark study on generating BHC from discharge summaries using domain-adapted LLMs [23]. However, similarity metrics do not fully capture alignment with expert

preferences due to the intricate nature of medical text. Consequently, automated BHC summarisation should serve as an assistive tool for healthcare professionals, as their critical role in evaluation remains indispensable.

Limitations in the study are acknowledged. Due to time constraints, the adaptation of LLMs could not be explored with greater context length, which is ideal for processing the lengthy content of EMRs. The data pre-processing work takes inspiration from relevant studies, which addressed this issue through content extraction of clinical notes [4] and a day-to-day [13] approach to generating BHC. Additionally, both studies revealed a lack of high-quality datasets available for clinical text summarisation. Future work aims to 1) construct a more robust dataset that emphasises the relationship between EMRs and hospital course summaries. 2) extend the context window of LLMs to provide more information as input. The emerging state space architecture model, Mamba [24], shows promising performance in processing long sequences, such as EMRs and other medical documentation, and could be an alternative to transformer-based models in clinical text summarisation tasks; 3) include a clinician evaluation study to align model outputs with experts' preferences.

5. Conclusions

This study evaluated the clinical text summarisation potential of domain-adapted pre-trained, open-source LLMs through QLoRA fine-tuning on a curated MIMIC III dataset containing EMR and BHC pairs. Fine-tuning an LLM pre-trained in the biomedical domain for automated hospital course summarisation can serve as an assistive tool for clinical coders to reduce their workload.

Acknowledgements

The experiment data is accessed via PhysioNet, after completion of CITI Data or Specimens Only Research training course. Model training was performed on A100 GPUs from Google Cloud Computing Services.

References:

[1] Chuang YN, Tang R, Jiang X, Hu X. SPeC: A soft prompt-based calibration on performance variability of large language model in clinical notes summarization. arXiv. 2023. doi: 10.48550/arXiv.2303.13035.

[2] Van Veen D, Van Uden C, Blankemeier L, Delbrouck JB, Aali A, Bluethgen C, Pareek A, Polacin M, Reis EP, Seehofnerova A, Rohatgi N, Hosamani P, Collins W, Ahuja N, Langlotz CP, Hom J, Gatidis S, Pauly J, Chaudhari AS. Clinical text summarization: adapting large language models can outperform human experts. arXiv. 2023. doi: 10.48550/arXiv.2309.07430.

[3] Fleming SL, Lozano A, Haberkorn WJ, Jindal JA, Reis EP, Thapa R, Blankemeier L, Genkins JZ, Steinberg E, Nayak A, Patel BS, Chiang CC, Callahan A, Huo Z, Gatidis S, Adams SJ, Fayanju O, Shah SJ, Savage T, Goh E, Chaudhari AS, Aghaeepour N, Sharp C, Pfeffer MA, Liang P, Chen JH, Morse KE, Brunskill EP, Fries JA, Shah NH. MedAlign: A clinician-generated dataset for instruction following with electronic medical records. arXiv. 2023. doi: 10.48550/arXiv.2308.14089.

[4] Searle T, Ibrahim Z, Teo J, Dobson RJB. Discharge summary hospital course summarization of in patient electronic health record text with clinical concept guided deep pre-trained transformer models. J Biomed Inform. 2023 May;141. doi:10.1016/j.jbi.2023.104358.

[5] Liang J, Tsou CH, Poddar A. A novel system for extractive clinical note summarization using EHR data. In: Rumshisky A, Roberts K, Bethard S, Naumann T, editors. Proceedings of the 2nd Clinical Natural Language Processing Workshop. Association for Computational Linguistics; 2019. p. 46-54. doi:10.18653/v1/W19-1906.
[6] Kolhatkar G, Paranjape A, Gokhale O, Kadam D. Team Converge at ProbSum 2023: Abstractive text summarization of patient progress notes. In: The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks. Association for Computational Linguistics; 2023. p. 510-515. doi:10.18653/v1/2023.bionlp-1.50.
[7] Zhou T, Cao P, Chen Y, Liu K, Zhao J, Niu K, Chong W. Automatic ICD coding via interactive shared representation networks with self-distillation mechanism. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. 2021. p. 5948-5957. doi:10.18653/v1/2021.acl-long.463.
[8] Yang J, Jin H, Tang R, Han X, Feng Q, Jiang H, Yin B, Hu X. Harnessing the power of LLMs in practice: A survey on ChatGPT and beyond. arXiv. 2023. doi:10.48550/arXiv.2304.13712.
[9] Basyal L, Sanghvi M. Text summarization using large language models: a comparative study of MPT-7b-instruct, Falcon-7b-instruct, and OpenAI Chat-GPT models. arXiv. 2023. arXiv:2310.10449.
[10] Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, Lee P, Lee YT, Li Y, Lundberg S, Nori H, Palangi H, Ribeiro MT, Zhang Y. Sparks of artificial general intelligence: early experiments with GPT-4. arXiv. 2023. doi:10.48550/arXiv.2303.12712.
[11] Liu Z, Zhong A, Li Y, Yang L, Ju C, Wu Z, Ma C, Shu P, Chen C, Kim S, Dai H, Zhao L, Zhu D, Liu J, Liu W, Shen D, Li X, Li Q, Liu T. Radiology-GPT: a large language model for radiology. arXiv. 2023. doi:10.48550/arXiv.2306.08666.
[12] Chao CJ, Banerjee I, Arsanjani R, Ayoub C, Tseng A, Kane GC, Oh JK, Li FF, Adeli E, Langlotz C. EchoGPT: A large language model for echocardiography report summarization. medRxiv preprint. 2024. doi:10.1101/2024.01.18.24301503.
[13] Hartman V, Campion TR. A Day-to-Day Approach for automating the hospital course section of the discharge summary. AMIA Jt Summits Transl Sci Proc. 2022 May 23;2022:216-225. PMID: 35854728; PMCID: PMC9285173.
[14] Tang L, Sun Z, Idnay B, Nestor JG, Soroush A, Elias PA, Xu Z, Ding Y, Durrett G, Rousseau JF, Weng C, Peng Y. Evaluating large language models on medical evidence summarization. NPJ Digit Med. 2023 Apr;6(1):158. doi:10.1038/s41746-023-00896-7.
[15] Johnson A, Pollard T, Mark R. MIMIC-III Clinical Database (version 1.4). PhysioNet. 2016. doi:10.13026/C2XW26.
[16] Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG. MIMIC-III, a freely accessible critical care database. Sci Data. 2016 May;3:160035. doi:10.1038/sdata.2016.35.
[17] Goldberger A, Amaral L, Glass L, Hausdorff J, Ivanov PC, Mark R, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation. 2000 Jun;101(23):e215-e220.
[18] Wang H, Gao C, Dantona C, Hull B, Sun J. DRG-LLaMA: tuning LLaMA model to predict diagnosis-related group for hospitalized patients. NPJ Digit Med. 2024 Jan;7:Article 16. doi:10.1038/s41746-023-00989-3.
[19] Labrak Y, Bazoge A, Morin E, Gourraud PA, Rouvier M, Dufour R. BioMistral: a collection of open-source pretrained large language models for medical domains. arXiv. 2024. doi:10.48550/arXiv.2402.10373.
[20] Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L. QLoRA: efficient finetuning of quantized LLMs. arXiv. 2023. doi:10.48550/arXiv.2305.14314.
[21] Ganesan K. ROUGE 2.0: Updated and improved measures for evaluation of summarization tasks. arXiv. 2018. doi:10.48550/arXiv.1803.01937.
[22] Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: evaluating text generation with BERT. arXiv. 2020. doi:10.48550/arXiv.1904.09675.
[23] Aali A, Van Veen D, Arefeen YI, Hom J, Bluethgen C, Reis EP, Gatidis S, Clifford N, Daws J, Tehrani AS, Kim J, Chaudhari AS. A benchmark of domain-adapted large language models for generating brief hospital course summaries. arXiv. 2024. doi:10.48550/arXiv.2403.05720.
[24] Gu A, Dao T. Mamba: Linear-time sequence modeling with selective state spaces. arXiv. 2023. doi:10.48550/arXiv.2312.0075.