Parameter-Efficient Fine-Tuning of LLaMA for the Clinical Domain

Aryo Pradipta Gema[1] **Pasquale Minervini**[1] **Luke Daines**[2]
Tom Hope[3,4] **Beatrice Alex**[5,6]
[1]School of Informatics, University of Edinburgh [2]Usher Institute, University of Edinburgh
[3]Allen Institute of AI
[4]Hebrew University of Jerusalem
[5]Edinburgh Futures Institute, University of Edinburgh
[6]School of Literatures, Languages and Cultures, University of Edinburgh
{aryo.gema, p.minervini, luke.daines, b.alex}@ed.ac.uk
tomh@allenai.org

Abstract

Adapting pretrained language models to novel domains, such as clinical applications, traditionally involves retraining their entire set of parameters. Parameter-Efficient Fine-Tuning (PEFT) techniques for fine-tuning language models significantly reduce computational requirements by selectively fine-tuning small subsets of parameters. In this study, we propose a two-step PEFT framework and evaluate it in the clinical domain. Our approach combines a specialised PEFT adapter layer designed for clinical domain adaptation with another adapter specialised for downstream tasks. We evaluate the framework on multiple clinical outcome prediction datasets, comparing it to clinically trained language models. Our framework achieves a better AUROC score averaged across all clinical downstream tasks compared to clinical language models. In particular, we observe large improvements of 4-5% AUROC in large-scale multilabel classification tasks, such as diagnoses and procedures classification. To our knowledge, this study is the first to provide an extensive empirical analysis of the interplay between PEFT techniques and domain adaptation in an important real-world domain of clinical applications.[1]

1 Introduction

Large Language Models (LLMs) have consistently achieved state-of-the-art performance across various NLP tasks. However, while these models exhibit impressive generalisation abilities, they often struggle to perform in specialised domains such as clinical applications, primarily due to the absence of domain-specific knowledge. The complexity of medical terminology and the presence of incomplete sentences in clinical notes contribute to this challenge (Lehman and Johnson, 2023). Unfortunately, studies have indicated that even LLMs pretrained with datasets comprising biomedical publications still exhibit suboptimal performance when applied to downstream clinical applications, particularly when compared to LLMs pretrained with clinical notes (Alsentzer et al., 2019; Li et al., 2022; Yang et al., 2022). This observation suggests that there are intrinsic nuances specific to the clinical context that can only be effectively captured if LLMs undergo pretraining using clinical datasets.

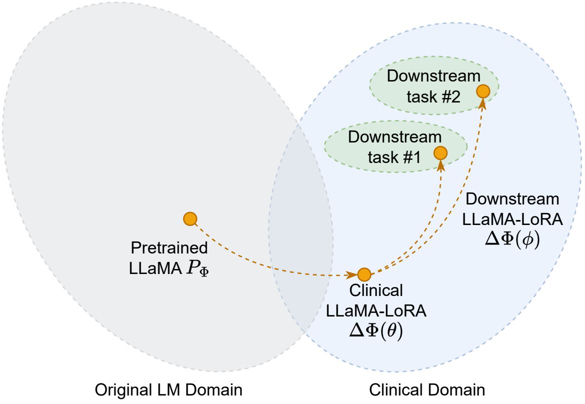
Downstream task #2
Downstream task #1
Downstream LLaMA-LoRA $\Delta\Phi(\phi)$
Pretrained LLaMA P_Φ
Clinical LLaMA-LoRA $\Delta\Phi(\theta)$
Original LM Domain
Clinical Domain

Figure 1: An illustration of the proposed two-step PEFT framework. Clinical LLaMA-LoRA fine-tunes the pretrained LLaMA to the clinical domain. Downstream LLaMA-LoRA further fine-tunes the domain-adapted model to downstream clinical tasks.

The current approach of adapting pretrained LLMs to the clinical domain typically involves fine-tuning the entire model parameters (Alsentzer et al., 2019; Peng et al., 2019; van Aken et al., 2021; Michalopoulos et al., 2021; Lehman and Johnson, 2023). However, due to the rapid increase in the size of LLMs, such a practice demands extensive computational resources, which may not be readily accessible to all researchers. Consequently, this challenge will further exacerbate the disparity between the resource-rich and resource-constrained research institutions (Ruder et al., 2022).

To address the substantial computational demands, studies have proposed various Parameter-

[1] The code is accessible via https://github.com/aryopg/clinical_peft.

Proceedings of the 6th Clinical Natural Language Processing Workshop, pages 91–104
June 21, 2024 ©2024 Association for Computational Linguistics

Efficient Fine-Tuning (PEFT) techniques. These techniques present a practical solution by fine-tuning a small subset of additional parameters while keeping the remaining pretrained parameters fixed. As a result, this strategy significantly alleviates the computational burden while achieving comparable performance to that of full fine-tuning.

In this study, we propose a two-step PEFT framework (see Figure 1). Firstly, we introduce Clinical LLaMA-LoRA, a Low-Rank Adaptation (LoRA, Hu et al., 2022) PEFT adapter built upon the open-source Large Language Model Meta AI (LLaMA) (Touvron et al., 2023). Then, we introduce Downstream LLaMA-LoRA, which is trained on top of the pretrained Clinical LLaMA-LoRA. Downstream LLaMA-LoRA is specifically designed for clinical downstream tasks. The fusion of the two adapters achieves better performance in clinical NLP downstream tasks compared to clinically trained LLMs while considerably reducing the computational requirements. This study presents the following contributions:

- We introduce Clinical LLaMA-LoRA, a PEFT-adapted version of the LLaMA model tailored specifically for the clinical domain.
- We provide comparisons of multiple PEFT techniques in terms of language modelling performance based on perplexity score, shedding light on the optimal PEFT techniques for the clinical domain-adaptive pretraining.
- We introduce Downstream LLaMA-LoRA, built on top of Clinical LLaMA-LoRA and tailored specifically for the clinical downstream tasks.
- We evaluate the proposed mixture of Clinical LLaMA-LoRA and Downstream LLaMA-LoRA on downstream clinical datasets and tasks. Our proposed framework showcases improvements in AUROC scores over the existing clinical LLMs.

2 Background

2.1 Biomedical Large Language Models

General-domain LLMs continue to face challenges when confronted with domain-specific tasks. The complexity associated with the requisite domain knowledge is recognised as a significant factor (Ling et al., 2023), particularly within the biomedical domain. Consequently, numerous studies have attempted to adapt LLMs specifically for the biomedical domain.

An early example of such adaptation is BioBERT (Lee et al., 2019), which was pretrained using biomedical research articles from PubMed and PubMed Central. This adaptation has shown improved performance across various biomedical NLP tasks. Recognising the significance of biomedical-specific vocabularies, Gu et al. (2022) proposed PubMedBERT, which is pretrained on biomedical data from scratch and initialised the model vocabulary with the biomedical corpus. The growing interest in biomedical NLP research has led to the adaptation of even larger models to the biomedical domain (Luo et al., 2022; Singhal et al., 2022; Wu et al., 2023; Singhal et al., 2023)

While these biomedical LLMs have demonstrated advancements in various biomedical NLP benchmarking tasks, studies have revealed that clinical LLMs still outperform their biomedical counterparts in numerous clinical downstream tasks (Alsentzer et al., 2019; Yang et al., 2022; Li et al., 2022; Lehman and Johnson, 2023). This suggests that domain-adaptive pretraining using clinical data is still the *de facto* protocol in adapting LLMs to the clinical domain.

2.2 Clinical Large Language Models

Clinical LLMs are often fine-tuned with clinical data from an LLM that is already pretrained with datasets that encompass broader topics. For instance, Bio+ClinicalBERT (Alsentzer et al., 2019) is domain-adaptively pretrained using clinical notes from the Medical Information Mart for Intensive Care (MIMIC)-III database (Johnson et al., 2016), starting from a pretrained BioBERT (Lee et al., 2019), which itself is pretrained on biomedical articles. BlueBERT (Peng et al., 2019) is domain-adaptively pretrained using PubMed abstracts and MIMIC-III clinical notes from a BERT model (Devlin et al., 2019), that is pretrained with general-domain texts. Similarly, Clinical-T5 (Lehman and Johnson, 2023) is domain-adaptively pretrained using the union of MIMIC-III and MIMIC-IV (Johnson et al., 2023) clinical notes from T5-base (Raffel et al., 2020), another general-domain LLM.

All these studies share a common approach, which is to fine-tune the entire model parameters. With massive LLMs, this method has become cost-prohibitive and inaccessible for many researchers.

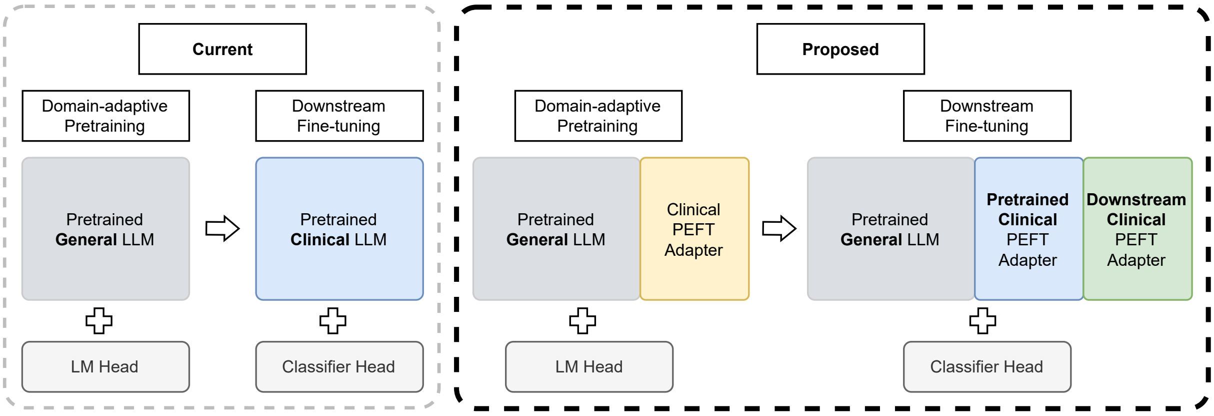

Figure 2: Frameworks of domain-adaptive and downstream fine-tuning to adapt a pretrained LLM from the general domain to the clinical domain. As opposed to a full fine-tuning process which can be prohibitively expensive (left), our approach leverages PEFT techniques to introduce a clinically-specialised adapter that is attached to a pretrained general LLM (right). Our proposed framework also introduces another clinical PEFT adapter trained on the downstream clinical tasks, such as clinical note classification.

2.3 Parameter-Efficient Fine-Tuning for Large Language Models

Suppose that we have a pretrained LLM $P_\Phi(y|x)$; fine-tuning it can be effectively defined as finding the most appropriate parameter changes $\Delta\Phi$ by optimising the fine-tuning objective. A conventional, full fine-tuning process means that the model needs to learn a $\Delta\Phi$ whose dimension is equal to the entire parameters of the pretrained LLM $|\Delta\Phi| = |\Phi_0|$, which is computationally expensive. PEFT techniques address this by tuning the *delta* $\Delta\Phi$, which corresponds to a very small fraction of additional trainable parameters during the fine-tuning process.

Adapter tuning (Houlsby et al., 2019) is an early PEFT method that involves adding small additional parameters called *adapters* to each layer of the pretrained model and strictly fine-tuning this small set of new parameters. LoRA (Hu et al., 2022) is another PEFT approach that trains low-rank matrices to represent the attention weights update of transformer-based models.

Another group of PEFT approaches leverages the concept of prompting. Prefix Tuning (Li and Liang, 2021) optimises a sequence of continuous task-specific vectors, called a *prefix*, which are trainable parameters that do not correspond to real tokens. P-Tuning (Liu et al., 2021b) uses a similar strategy as Prefix tuning with a focus on text understanding tasks, as opposed to generative tasks. Prompt tuning (Lester et al., 2021) simplifies Prefix tuning by introducing trainable tokens, called *soft prompts*, for each downstream task. Liu et al. (2021a) introduced P-tuning v2 which uses deep prompt tuning to address the lack of performance gain in the previous prompt tuning techniques.

By fine-tuning a small fraction of additional parameters, all PEFT approaches alleviate the issue of extensive computational resource requirements.

2.4 Multi-step Adaptation

Prior studies have explored the two-step adaptation framework, although they have fundamental differences from our proposed setup. For instance, Zhang et al. (2021) introduced a multi-domain unsupervised domain adaptation (UDA) with a two-step strategy, involving domain-fusion training with Masked Language Model loss on a mixed corpus, followed by task fine-tuning with a task-specific loss on the domain corpus. More recently, Malik et al. (2023) introduced UDApter which utilises PEFT adapters to do efficient UDA. However, unsupervised domain matching techniques such as UDApter rely on restrictive assumptions about the underlying data distributions that are often unsatisfied in real-world scenarios (Li et al., 2020). In our study, we experiment with the clinical domain as the target domain that is not available in the LLM's initial pretraining. Consequently, significant discrepancies exist between the distributions of the source and target domains. Leveraging the amount of available clinical notes, we adopt a self-supervised learning paradigm by continually pretraining the LLMs within the target domain rather than relying on the UDA paradigm.

Our approach shares theoretical similarities with the multi-step continual pretraining approach, proposed by Gururangan et al. (2020), which proposes domain- and task-adaptive pretraining. However, the main difference between our proposed approach and Gururangan et al. (2020) is in the discrepancy between the source and the target domains. Gururangan et al. (2020) experimented with adapting general-domain LLMs to domains encountered during their initial pretraining, such as news and biomedical domains. On the other hand, we experiment with the clinical domain which is entirely absent from the LLMs' initial pretraining due to legal constraints which restrict access to sensitive clinical notes. On top of that, adapting to the clinical domain poses a bigger challenge due to the complexity of medical terminology and the presence of incomplete sentences (Lehman et al., 2023).

3 Methodology

3.1 Problem Statement

Figure 2 shows the comparison between the current and proposed problem definitions. The general problem can be decomposed into two stages:

Domain-adaptive Pretraining. Given a pretrained general LLM $P_\Phi(y|x)$ with its parameters Φ and a training dataset $\mathcal{Z} = \{(x_i, y_i)\}_{i=1,\ldots,N}$. To adapt to the new domain, the model needs to update its weight iteratively from its pretrained state Φ_0 to $\Phi = \Phi_0 + \Delta\Phi$. This process of maximising the objective function can be defined as:

$$\operatorname*{argmax}_{\Phi} \sum_{(x,y)\in\mathcal{Z}} \sum_{t=1}^{|y|} \log \left(P_\Phi \left(y_t \mid x, y_{<t}\right)\right)$$

In the current paradigm, a full fine-tuning process means that the model needs to learn a $\Delta\Phi$ whose dimension is equal to the entire pretrained parameters $|\Delta\Phi| = |\Phi_0|$, which is computationally expensive.

In the proposed paradigm, we tune only small additional parameters θ such that $\Phi = \Phi_0 + \Delta\Phi(\theta)$ whose dimension is very small compared to the original parameters $|\theta| \ll |\Phi_0|$. Thus, the training objective can be redefined as:

$$\operatorname*{argmax}_{\theta} \sum_{(x,y)\in\mathcal{Z}} \sum_{t=1}^{|y|} \log \left(P_{\Phi_0+\Delta\Phi(\theta)} \left(y_t \mid x, y_{<t}\right)\right)$$

In the current paradigm, the outcome of domain-adaptive pretraining would be a clinically-adapted LLM. While in the proposed paradigm, the outcome would be the clinical PEFT component, which can be combined with the untouched pretrained general LLM for downstream applications.

Downstream Fine-tuning. In the current paradigm, the pretrained clinical LLM is fine-tuned to the downstream tasks, such as document classification tasks. Suppose that we have a pretrained clinical LLM $P_{\Phi,\Theta}$ with its domain-adapted parameters Φ and a newly initialised classifier layer Θ, as well as a training dataset $\mathcal{Z} = \{(x_i, y_i)\}_{i=1,\ldots,N}$. We want to maximise a specific loss function, such as a cross-entropy loss:

$$\operatorname*{argmax}_{\Phi,\Theta} \frac{1}{N} \sum_{i=1}^{N} y_i \log \left(P_{\Phi,\Theta} \left(x_i\right)\right)$$

In contrast, in the proposed paradigm, the fine-tuning process only updates the small additional parameters $\Delta\Phi(\theta)$ and the classifier head Θ:

$$\operatorname*{argmax}_{\theta,\Theta} \frac{1}{N} \sum_{i=1}^{N} y_i \log \left(P_{\Phi+\Delta\Phi(\theta),\Theta} \left(x_i\right)\right)$$

In fact, we can also decompose the fine-tuning into an additional "delta-updating" process:

$$\operatorname*{argmax}_{\theta,\phi,\Theta} \frac{1}{N} \sum_{i=1}^{N} y_i \log \left(P_{\Phi+\Delta\Phi(\theta)+\Delta\Phi(\phi),\Theta} \left(x_i\right)\right)$$

Similar to the Domain-adaptive Pretraining stage, the dimensions of the additional parameters θ and ϕ are very small compared to the original parameters. By updating only the additional parameters and the classifier head, the proposed paradigm reduces the computational requirements, making it more efficient and feasible, especially for clinical settings that are often resource-constrained.

3.2 Two-step LLaMA-LoRA

In this study, we propose a two-step PEFT framework (as shown on the right-hand side of Figure 2). Firstly, we introduce Clinical LLaMA-LoRA, a LoRA adapter built upon LLaMA (Touvron et al., 2023) that is adapted to the clinical domain. Secondly, we introduce Downstream LLaMA-LoRA, which is trained on top of the pretrained Clinical LLaMA-LoRA and is specifically adapted to the downstream tasks.

LLaMA models In this study, we evaluate two LLaMA models; the 7 billion parameters version of LLaMA (Touvron et al., 2023) and the 7 billion parameters version of PMC-LLaMA(Wu et al., 2023). LLaMA was pretrained with an array of texts from multiple sources, such as English CommonCrawl, Wikipedia, ArXiv, and C4 (Raffel et al.,

Dataset	# Class	Multilabel	# Train	# Valid	# Test
LOS	4	✗	30,421	4,391	8,797
MOR	2	✗	33,954	4,908	9,822
PMV	2	✗	5,666	707	706
DIAG	1,266	✓	33,994	4,918	9,829
PROC	711	✓	30,030	4,357	8,681

Table 1: Statistics and types of downstream clinical document classification tasks: length of stay (LOS), mortality (MOR), prolonged mechanical ventilation (PMV), diagnoses (DIAG), and procedures (PROC).

2020). While, PMC-LLaMA is a domain-adapted LLaMA model that was pretrained on 4.8 million biomedical academic papers from PubMed Central.

Domain-adaptive Pretraining: Clinical LLaMA-LoRA Clinical LLaMA-LoRA is trained using a combination of MIMIC-IV de-identified discharge summaries (331,794) and radiology reports (2,321,355), resulting in a collection of 2,653,149 individual clinical notes. We evaluate five PEFT techniques, which include *LoRA* (Hu et al., 2022), *Adaptation Prompt* (Zhang et al., 2023), *Prefix Tuning* (Li and Liang, 2021), *Prompt Tuning* (Lester et al., 2021), and *P-tuning* (Liu et al., 2021b).

Our approach follows the autoregressive language modelling pretraining objective employed in the original LLaMA training. To ensure compatibility with available computational resources, we use fixed model hyperparameters that allow us to fit the LLM into a single NVIDIA A100-80GB GPU (see Appendix A.1). We optimise the hyperparameters specific to each PEFT method using Gaussian Process regression for Bayesian Optimisation (Frazier, 2018) [2] with a maximum of 20 trials. The detailed hyperparameters search space can be found in Appendix A.2. During this stage, we evaluate the perplexity scores of the LLM variants.

Downstream Fine-tuning: Downstream LLaMA-LoRA We fine-tune the Clinical LLaMA-LoRA and Downstream LLaMA-LoRA to clinical document classification tasks:

- **Prolonged mechanical ventilation (PMV)**: a binary classification task to predict whether a patient will require mechanical ventilation for more than seven days (Huang et al., 2020; Naik et al., 2022).
- **In-hospital mortality (MOR)**: a binary classification task to predict whether a patient will survive during their hospital stay (van Aken et al., 2021; Naik et al., 2022).
- **Length of stay (LOS)**: a multiclass classification task to predict the length of a patient's hospital stay, categorised into four time-bins: less than three days, three to seven days, one to two weeks, and more than two weeks (van Aken et al., 2021; Naik et al., 2022).
- **Diagnoses (DIAG)**: a large-scale multilabel classification task to predict the differential diagnoses of a patient, represented by simplified ICD-9 diagnosis codes (van Aken et al., 2021).
- **Procedures (PROC)**: a large-scale multilabel classification task to predict the treatments administered to a patient, represented by simplified ICD-9 procedure codes (van Aken et al., 2021).

The label and split statistics of each dataset can be found in Table 1.

During this downstream fine-tuning process, we use fixed model hyperparameters to ensure compatibility with the available computational resources, a single NVIDIA A100-80GB GPU (see Appendix B.1). We optimise the hyperparameters specific to each PEFT method using Gaussian Process regression for Bayesian Optimisation with a maximum of 20 trials. The detailed hyperparameters search space of the PEFT method can be found in Appendix B.2.

For evaluating the performance of the model on these downstream tasks, we report the Area Under the Receiver Operating Characteristic Curve (AUROC) scores. Additionally, we report the macro-averaged AUROC score across all clinical tasks as commonly done in NLP benchmarking tasks (Wang et al., 2019; Peng et al., 2019; Gu et al., 2022).

3.3 Baseline Models

We selected baseline models that have undergone a domain-adaptive pretraining process on clinical notes (MIMIC-III). Thus, these baseline models have been designed to perform specifically on clinical data, providing comparison points for evaluating our proposed approach of two-step adaptation in downstream clinical NLP tasks. The baseline models used in the evaluation are as follows:

- **Bio+ClinicalBERT** (Alsentzer et al., 2019): Bio+ClinicalBERT is pretrained on MIMIC-III clinical notes. It is initialised from a biomedical language model called BioBERT (Lee et al., 2019), which is pretrained on biomedical research articles.

[2] Specifically, we use the W&B Sweep APIs: `https://docs.wandb.ai/guides/sweeps`

- **BlueBERT** (Peng et al., 2019): BlueBERT is pretrained on MIMIC-III clinical notes and PubMed abstracts starting from the pretrained checkpoint of BERT (Devlin et al., 2019), a general-domain language model.
- **CORe** (van Aken et al., 2021): CORe is pretrained on MIMIC-III clinical notes and biomedical articles starting from the pretrained checkpoint of BioBERT (Lee et al., 2019).
- **UmlsBERT** (Michalopoulos et al., 2021): UmlsBERT is pretrained on MIMIC-III clinical notes using the pretrained weights of Bio+ClinicalBERT with modified architecture and pretraining objective that incorporates knowledge from the Unified Medical Language System (UMLS) Metathesaurus (Schuyler et al., 1993).

4 Results and Analysis

4.1 Domain-adaptive Pretraining

The pretraining results can be found in Table 2. We employ PEFT techniques for domain-adaptive pretraining, requiring a significantly smaller number of parameters ranging from just 0.001% to 0.24% of the original model parameters. This approach substantially reduces the required computational resources and training time. We perform a full-parameter domain-adaptive pretraining of LLaMA, referred to as **Clinical LLaMA**, using four NVIDIA A100-80GB GPUs which took 49.5 hours. Instead, PEFT techniques require less than 24 hours per epoch on average with only a single GPU with a comparable perplexity score.

LoRA emerges as the best-performing PEFT method for both LLaMA and PMC-LLaMA in the clinical domain-adaptive pretraining, achieving the lowest perplexity scores of 2.244 and 2.404, respectively, which are very similar to Clinical LLaMA's perplexity score of 2.210. This pretrained LoRA is referred to as **Clinical LLaMA-LoRA** in the subsequent sections. The following experiments in downstream fine-tuning will utilise this pretrained Clinical LLaMA-LoRA.

4.2 Downstream Fine-tuning

From the downstream fine-tuning results shown in Table 3, we can decompose the analysis into multiple research questions:

Can LoRA help fine-tune LLaMA from other domains (general and biomedical) to achieve higher AUROC scores in clinical tasks? We compare the results obtained by LLaMA and LLaMA + LoRA, as well as PMC-LLaMA and PMC-LLaMA + LoRA, as presented in Table 3. The obtained results consistently demonstrate improved AUROC scores when utilising LoRA across all tasks. The macro-averaged AUROC score of LoRA-equipped LLaMA shows a notable 13.01% increase when compared to the LLaMA-only baseline. Similarly, LoRA-equipped PMC-LLaMA exhibits a 12.19% improvement in macro-averaged AUROC compared to the original PMC-LLaMA Both LLaMA and PMC-LLaMA, when equipped with LoRA, show significant AUROC score improvements in all tasks except the PMV prediction task, which is challenging for all model variants.

Furthermore, the marginal difference in AUROC scores between PMC-LLaMA and the general-domain LLaMA may be attributed to two factors. Firstly, the original LLaMA has been exposed to biomedical concepts during its pretraining, reducing the need for domain-adaptive pretraining to the biomedical domain. Secondly, clinical outcome prediction requires an understanding of how to apply biomedical knowledge in an interconnected manner to provide prognostic. We believe that biomedical pretraining may not be sufficient in providing such practical knowledge.

Can LoRA-equipped LLaMA and PMC-LLaMA perform comparably in comparison to clinically trained LMs? We compare the AUROC scores obtained by the baseline models, and LoRA-equipped LLaMA and PMC-LLaMA (see Table 3). Among the baseline models, UmlsBERT performs the best with a macro-averaged AUROC score of 72.70%. Compared to UmlsBERT, both LLaMA and PMC-LLaMA underperform with macro-averaged AUROC scores of 58.61% and 60.51%, respectively. This finding highlights the importance of clinical-specific fine-tuning.

Significant improvements can be observed in LoRA-equipped LLaMA and PMC-LLaMA, with macro-averaged AUROC scores of 71.62% and 72.70%, respectively, with noticeable improvements in the diagnoses and procedures prediction tasks. LoRA-equipped LLaMA achieves AUROC scores of 78.37% and 87.49% in the diagnoses and procedures prediction tasks, respectively, compared to 72.08% and 78.32% for UmlsBERT. This represents improvements of 6.29% in diagnoses prediction and 9.17% in procedures prediction. Improvements are also observed in the results obtained by LoRA-equipped PMC-LLaMA, outperforming

Base Model	PEFT	Trainable Params	Train Ppl	Test Ppl	GPU	Train Time (h:m:s)
Clinical LLaMA	-	6.7B (100%)	1.811	2.210	4x80GB	49:26:38
LLaMA	**LoRA**	**8.4M (0.12%)**	**1.858**	**2.244**	1x80GB	**21:37:42**
	Adaptation Prompt	1.2M (0.02%)	2.561	2.865	1x80GB	24:57:17
	Prefix Tuning	5.2M (0.08%)	2.815	2.748	1x80GB	20:11:07
	Prompt Tuning	61.4K (0.0009%)	4.846	4.007	1x80GB	23:27:28
	P-tuning	16.1M (0.24%)	2.723	3.271	1x80GB	23:49:31
PMC-LLaMA	**LoRA**	**2.1M (0.03%)**	**1.938**	**2.404**	1x80GB	**21:32:59**
	Adaptation Prompt	1.2M (0.018%)	2.374	2.867	1x80GB	23:33:10
	Prefix Tuning	2.6M (0.04%)	1.789	2.848	1x80GB	20:13:10
	Prompt Tuning	41K (0.0006%)	4.821	4.385	1x80GB	22:25:32
	P-tuning	2.2M (0.03%)	3.491	4.572	1x80GB	22:28:15

Table 2: Domain-adaptive Pretraining results of LLaMA and PMC-LLaMA trained on MIMIC-IV clinical notes with a language modelling objective. Lower perplexity scores indicate better language modelling performance. The **boldface row** indicates the model with the lowest perplexity score from each base model variant.

UmlsBERT by 6.73% in diagnoses prediction and 8.36% in procedures prediction.

Can LLaMA and PMC-LLaMA with Clinical LLaMA-LoRA achieve higher AUROC scores than the clinically trained LMs? The domain-adaptive pretraining step yields the clinically-trained LoRA adapters for LLaMA and PMC-LLaMA, denoted as **Clinical LLaMA-LoRA**. We compare the results of Clinical LLaMA-LoRA-equipped LLaMA and PMC-LLaMA with the baseline models. We evaluate Clinical LLaMA-LoRA with and without fine-tuning, referred to as "Trainable" and "Frozen" respectively.

The results indicate that Clinical LLaMA-LoRA-equipped LLaMA and PMC-LLaMA outperform the baseline models. LLaMA with a trainable Clinical LLaMA-LoRA achieves an AUROC score of 75.13%, surpassing UmlsBERT's score of 72.32%. PMC-LLaMA with a trainable Clinical LLaMA-LoRA achieves a lower AUROC score of 72.23%. LLaMA with a trainable Clinical LLaMA-LoRA also outperforms Clinical LLaMA which achieves an AUROC score of 58.86%.

These findings indicate that the Clinical LLaMA-LoRA contributes to higher AUROC scores for LLaMA and PMC-LLaMA over clinically trained LLMs, while biomedical domain-adaptive pretraining may not be necessary to improve the model's performance in the clinical settings.

Can LLaMA and PMC-LLaMA with Clinical LLaMA-LoRA achieve higher AUROC scores than the other fine-tuning variants? We examine the importance of the domain-adapted LoRA by comparing the results obtained by LLaMA and PMC-LLaMA equipped with Clinical LLaMA-LoRA against the results of LLaMA and PMC-LLaMA fine-tuning, both original and with LoRA.

Firstly, we evaluate the frozen pretrained Clinical LLaMA-LoRA. Both LLaMA and PMC-LLaMA with frozen Clinical LLaMA-LoRA do not exhibit a significant increase in performance compared to the original fine-tuning. This indicates that, despite the domain-adaptive pretraining, the limited number of trainable parameters during the downstream fine-tuning restricts the potential improvement that the model can achieve. A similar finding can also be observed in the Clinical LLaMA fine-tuning whose overall performance does not differ from the original fine-tuning. This finding is further supported by the improvement in the AUROC scores of LLaMA and PMC-LLaMA with trainable Clinical LLaMA-LoRA, which achieve 75.13% and 72.23% macro-averaged AUROC scores, respectively. These represent substantial improvements from the vanilla fine-tuning performance, 58.61% and 60.51% AUROC scores.

Can a downstream LoRA adapter improve the AUROC scores of LLaMA and PMC-LLaMA equipped with Clinical LLaMA-LoRA? By considering Clinical LLaMA-LoRA as the "delta-updating" outcome of the domain-adaptive pretraining, we can view the downstream fine-tuning process as an additional "delta-updating" step. To investigate the impact of this approach, we conduct experiments by adding a Downstream LLaMA-LoRA to LLaMA and PMC-LLaMA models that were already equipped with Clinical LLaMA-LoRA. From Table 3, we can observe that Downstream LLaMA-LoRA fails to improve the performance of LLaMA and PMC-LLaMA with frozen Clinical LLaMA-LoRA. On the other

Model	**PMV**	**MOR**	**LOS**	**DIAG**	**PROC**	**Macro Average**
BlueBERT	57.31	81.34	72.92	73.39	76.62	72.32
UmlsBERT	***58.29***	*81.83*	*73.02*	*72.08*	*78.32*	*72.70*
Bio+ClinicalBERT	54.00	72.67	72.21	76.65	83.21	71.75
CORe	52.11	71.52	64.17	72.40	84.51	69.40
Clinical LLaMA*	52.28	63.22	56.06	59.31	63.42	58.86
LLaMA∗	51.38	66.80	57.65	60.06	63.83	58.61
+ LoRA	51.65	74.89	65.70	78.37	87.49	71.62
+ Clinical LLaMA-LoRA (Frozen)	52.22	60.88	55.05	57.64	62.48	57.65
+ Downstream LLaMA-LoRA	52.31	61.72	55.16	57.70	62.58	57.90
+ Clinical LLaMA-LoRA (Trainable)	51.41	81.16	72.44	**81.97**	**88.69**	75.13
+ *Downstream LLaMA-LoRA*	*53.81*	***83.02***	***73.26***	*81.93*	*88.31*	***76.07***
PMC-LLaMA∗	53.06	66.77	57.94	60.17	64.63	60.51
+ *LoRA*	*53.84*	*78.03*	*66.14*	*78.81*	*86.68*	*72.70*
+ Clinical LLaMA-LoRA (Frozen)	51.33	67.19	58.13	63.59	68.26	60.06
+ Downstream LLaMA-LoRA	50.90	67.00	58.31	60.50	64.42	60.23
+ Clinical LLaMA-LoRA (Trainable)	52.88	75.86	65.89	79.66	86.85	72.23
+ Downstream LLaMA-LoRA	52.21	76.54	68.42	78.67	87.08	72.58

Table 3: AUROC scores in clinical downstream document classification tasks. The macro-averaged AUROC score is calculated by taking the average of AUROC scores across all tasks. The **boldface cell** indicates the highest AUROC score in a column, the *row in italic* indicates the variant with the highest macro-averaged AUROC in its category. + *LoRA* denotes applying LoRA on top of the pretrained LLM without domain-adaptive pretraining. + *Clinical LLaMA-LoRA* denotes applying Clinical LLaMA-LoRA that is domain-adaptively pretrained on top of the pretrained LLM. + *Downstream LLaMA-LoRA* denotes applying Downstream LLaMA-LoRA on top of the LLM + Clinical LLaMA-LoRA. *Frozen* means that the parameters are not trainable, while *Trainable* means that the parameters are trainable. ∗ Due to restricted computing resources, the fine-tunings of Clinical LLaMA, LLaMA, and PMC-LLaMA were constrained to only training the final classification layer.

hand, improvement can be observed when adding Downstream LLaMA-LoRA to LLaMA with trainable Clinical LLaMA-LoRA. This combination of LLaMA with trainable Clinical LLaMA-LoRA and Downstream LLaMA-LoRA achieves the highest macro-averaged AUROC score of 76.07%. The macro-averaged AUROC score of Clinical LLaMA-LoRA was almost similar to that of PMC-LLaMA with LoRA, suggesting similar efficacy between Clinical LLaMA-LoRA and the full fine-tuning process that PMC-LLaMA has undergone. Moreover, Clinical LLaMA-LoRA offers the advantage of reduced computational resources and training time, which is aligned with the requirements of practical implementation in clinical settings.

Overall, our proposed method manages to achieve better performance in comparison to clinically trained models. We also provide a comparison with the state-of-the-art method of PMV, mortality, and length of stay predictions, called BEEP (Naik et al., 2022), which leverages retrieval augmentation method to provide more contextual information to the model during inference. The comparison is only partial as BEEP models were not evaluated on the diagnosis and procedure prediction tasks. As shown in Appendix C, our best-performing model achieves a 70.03% averaged AUROC score, which is slightly worse compared to the best-performing BEEP model with 72.26% averaged AUROC score. However, it is worth noting that our proposed method and the state-of-the-art method are complementary to each other. Hence, future work may explore the possibility of combining the two approaches.

5 Conclusions

In this study, we propose a two-step PEFT framework. We introduce Clinical LLaMA-LoRA, a LoRA (Hu et al., 2022) adapter built upon LLaMA (Touvron et al., 2023). Then, we introduce Downstream LLaMA-LoRA, a task-specific adapter that is trained on top of the pretrained Clinical LLaMA-LoRA. The fusion of the two adapters achieves an AUROC score of 76.07% macro-averaged across all clinical NLP downstream tasks, which represents a 3.37% improvement over the best-performing clinical LLM. Our proposed framework achieves improvement in performance while reducing the computational requirements, which is suited for clinical settings that are often constrained by their computational power.

Limitations

This study presents a two-step PEFT framework aimed at effectively adapting LLMs to diverse clinical downstream applications. However, the evaluation of our model was restricted to MIMIC-based datasets, which are constrained to English and obtained exclusively within the Commonwealth of Massachusetts, United States of America. Consequently, despite the promising efficacy demonstrated by our proposed method, it would have been advantageous to directly assess its performance across diverse hospital systems spanning other geographical locations and languages. This would enable a more comprehensive understanding of its applicability and generalizability. However, it is essential to acknowledge that conducting such an analysis would require working within a trusted research environment and obtaining the necessary permissions to access the relevant datasets.

It is crucial to recognise the restrictions imposed on accessing internal clinical datasets, as they limit our ability to evaluate the effectiveness of our approach across different care provider systems. Therefore, we encourage care providers to conduct internal experiments within their trusted research environment to ensure the efficacy of our proposed method within their specific use cases should they adopt this approach.

Despite the demonstrated performance improvements, the proposed model may still be susceptible to spurious correlations. Predicting patient outcomes solely based on clinical notes presents significant challenges due to the other factors that may not be captured within those notes. For instance, the length of a patient's in-hospital stay is not solely correlated with their diagnoses and disease progression. Factors such as the patient's insurance status, which is not typically mentioned in clinical notes, can severely impact the duration of a patient's stay. Therefore, we encourage end users of such clinical LLMs to consider additional measures to ensure predictions that reflect a holistic view of the patient's situation, instead of relying solely on the predictions of LLMs.

Ethics Statement

In this study, we use MIMIC-based datasets obtained after completing the necessary training. These datasets comply with de-identification standards set by the Health Insurance Portability and Accountability Act (HIPAA) through data cleansing. Due to privacy concerns, we refrain from including direct excerpts of the data in the paper. We also refrain from publicly sharing the pretrained checkpoints.

While our model demonstrates effectiveness, it is important to acknowledge the risks associated with relying solely on clinical outcome prediction models. There are crucial pieces of information that can be found beyond the scope of clinical notes. Considering the potential impact on patient health outcomes, it is crucial to exercise caution when utilising these clinical LLMs. Therefore, we propose that the PEFT adapter generated by our framework, in conjunction with the pretrained LLM, should be used as an aid rather than a replacement for trained clinical professionals.

Acknowledgements

APG was supported by the United Kingdom Research and Innovation (grant EP/S02431X/1), UKRI Centre for Doctoral Training in Biomedical AI at the University of Edinburgh, School of Informatics. PM was partially funded by the European Union's Horizon 2020 research and innovation programme under grant agreement no. 875160, ELIAI (The Edinburgh Laboratory for Integrated Artificial Intelligence) EPSRC (grant no. EP/W002876/1), an industry grant from Cisco, and a donation from Accenture LLP; and is grateful to NVIDIA for the GPU donations. BA was partially funded by Legal and General PLC as part of the Advanced Care Research Centre and by the Artificial Intelligence and Multimorbidity: Clustering in Individuals, Space and Clinical Context (AIM-CISC) grant NIHR202639. For the purpose of open access, the authors have applied a Creative Commons attribution (CC BY) licence to any author-accepted manuscript version arising. This work was supported by the Edinburgh International Data Facility (EIDF) and the Data-Driven Innovation Programme at the University of Edinburgh.

We also thank Yahan Li, Ajitha Rajan, and Bruce Guthrie for their feedback on the manuscript.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Peter I. Frazier. 2018. A tutorial on bayesian optimization. *CoRR*, abs/1807.02811.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *Proceedings of the 36th International Conference on Machine Learning*, page 2790–2799. PMLR.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Kexin Huang, Abhishek Singh, Sitong Chen, Edward Moseley, Chih-Ying Deng, Naomi George, and Charolotta Lindvall. 2020. Clinical XLNet: Modeling Sequential Clinical Notes and Predicting Prolonged Mechanical Ventilation. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 94–100, Online. Association for Computational Linguistics.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1.

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Eric Lehman, Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J. Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, and Emily Alsentzer. 2023. Do We Still Need Clinical Language Models?

Eric Lehman and Alistair Johnson. 2023. Clinical-T5: Large Language Models Built Using MIMIC Clinical Text.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bo Li, Yezhen Wang, Tong Che, Shanghang Zhang, Sicheng Zhao, Pengfei Xu, Wei Zhou, Yoshua Bengio, and Kurt Keutzer. 2020. Rethinking distributional matching based domain adaptation. *CoRR*, abs/2006.13352.

Xiang Lisa Li and Percy Liang. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Yikuan Li, Ramsey M. Wehbe, Faraz S. Ahmad, Hanyin Wang, and Yuan Luo. 2022. Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences. *CoRR*, abs/2201.11838.

Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun Li, Hejie Cui, Xuchao Zhang, Tianjiao Zhao, Amit Panalkar, Wei Cheng, Haoyu Wang, Yanchi Liu, Zhengzhang Chen, Haifeng Chen, Chris White, Quanquan Gu, Carl Yang, and Liang Zhao. 2023. Beyond one-model-fits-all: A survey of domain specialization for large language models. *CoRR*, abs/2305.18703.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021a. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *CoRR*, abs/2110.07602.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021b. GPT understands, too. *CoRR*, abs/2103.10385.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6). Bbac409.

Bhavitvya Malik, Abhinav Ramesh Kashyap, Min-Yen Kan, and Soujanya Poria. 2023. UDAPTER - efficient domain adaptation using adapters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2249–2263, Dubrovnik, Croatia. Association for Computational Linguistics.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, and Sayak Paul. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft.

George Michalopoulos, Yuanxin Wang, Hussam Kaka, Helen Chen, and Alexander Wong. 2021. Umls-BERT: Clinical domain knowledge augmentation of contextual embeddings using the Unified Medical Language System Metathesaurus. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1744–1753, Online. Association for Computational Linguistics.

Aakanksha Naik, Sravanthi Parasa, Sergey Feldman, Lucy Lu Wang, and Tom Hope. 2022. Literature-augmented clinical outcome prediction. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 438–453. Association for Computational Linguistics.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67. Citation Key: JMLR:v21:20-074.

Sebastian Ruder, Jonas Pfeiffer, and Ivan Vulić. 2022. Modular and Parameter-Efficient Fine-Tuning for NLP Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 23–29, Abu Dubai, UAE. Association for Computational Linguistics.

P L Schuyler, W T Hole, M S Tuttle, and D D Sherertz. 1993. The UMLS Metathesaurus: Representing different views of biomedical concepts. *Bulletin of the Medical Library Association*, 81(2):217–222.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Kumar Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Schärli, Aakanksha Chowdhery, Philip Andrew Mansfield, Blaise Agüera y Arcas, Dale R. Webster, Gregory S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle K. Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2022. Large language models encode clinical knowledge. *CoRR*, abs/2212.13138.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Andrew Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Agüera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle K. Barral, Dale R. Webster, Gregory S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards expert-level medical question answering with large language models. *CoRR*, abs/2305.09617.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Betty van Aken, Jens-Michalis Papaioannou, Manuel Mayrdorfer, Klemens Budde, Felix Gers, and Alexander Loeser. 2021. Clinical Outcome Prediction from Admission Notes using Self-Supervised Knowledge Integration. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 881–893, Online. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-llama: Further finetuning llama on medical papers. *CoRR*, abs/2304.14454.

Xi Yang, Aokun Chen, Nima M. Pournejatian, Hoo Chang Shin, Kaleb E. Smith, Christopher

Parisien, Colin Compas, Cheryl Martin, Anthony B. Costa, Mona G. Flores, Ying Zhang, Tanja Magoc, Christopher A. Harle, Gloria P. Lipori, Duane A. Mitchell, William R. Hogan, Elizabeth A. Shenkman, Jiang Bian, and Yonghui Wu. 2022. A large language model for electronic health records. *npj Digit. Medicine*, 5.

Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. 2023. LLaMA-Adapter: Efficient Fine-tuning of Language Models with Zero-init Attention.

Rongsheng Zhang, Yinhe Zheng, Xiaoxi Mao, and Minlie Huang. 2021. Unsupervised domain adaptation with adapter. *CoRR*, abs/2111.00667.

A Hyperparameters for the Domain-adaptive Pretraining

A.1 Fixed Model Hyperparameters

Hyperparameter	Value
Learning rate	3e-4
Warmup steps ratio	0.06
Maximum sequence length	512
Gradient accumulation step	4
Batch size	10

Table 4: Fixed model hyperparameters for language modelling pretraining. These hyperparameters remain unchanged to fit LLaMA into a single GPU.

A.2 PEFT Hyperparameters Optimisation Search Space

PEFT	Hyperparameter	Search space
LoRA	r alpha dropout	[2, 4, 8, 16] [4, 8, 16, 32] [0.0, 0.1, 0.2]
Prefix Tuning	num virtual tokens prefix projection	[1, 5, 10, 15, 20] [true, false]
Prompt Tuning	num virtual tokens prompt init	[1, 5, 10, 15, 20] [text, random]
P-Tuning	num virtual tokens reparameterisation hidden size num layers dropout	[1, 5, 10, 15, 20] ["MLP", "LSTM"] [64, 128, 256, 768] [1, 2, 4, 8, 12] [0.0, 0.1, 0.2]
Adaptation Prompt	adapter length adapter layers	[5, 10] [10, 20, 30]

Table 5: The search space for PEFT Hyperparameters optimisation runs during the domain adaptation fine-tuning with language modelling objective. Each PEFT technique has a specific set of hyperparameters to tune, we selected the combination of hyperparameters which has the lowest perplexity score.

Specifically for Prompt Tuning, we use a common prompt initialisation text "Finish this clinical note:".

B Hyperparameters for the Downstream Fine-tuning

B.1 Fixed Model Hyperparameters

Hyperparameter	Value
Learning rate	5e-5
Warmup steps ratio	0.06
Maximum sequence length	512
Gradient accumulation step	10
Batch size	10

Table 6: Fixed model hyperparameters for the clinical downstream fine-tuning. These hyperparameters remain unchanged to fit LLaMA into a single GPU.

B.2 PEFT Hyperparameters Optimisation Search Space

PEFT	Hyperparameter	Search space
LoRA	r alpha dropout	[2, 4, 8, 16] [4, 8, 16, 32] [0.0, 0.1, 0.2]

Table 7: The search space for PEFT Hyperparameters optimisation runs during the downstream fine-tuning. Each PEFT technique has a specific set of hyperparameters to tune, we selected the combination of hyperparameters which has the highest AUROC score.

C Comparison with BEEP (Naik et al., 2022)

Model	**PMV**	**MOR**	**LOS**	**Avg**
BEEP	*59.43*	*84.65*	*72.71*	*72.26*
Our method	53.81	83.02	73.26	70.03

Table 8: AUROC scores in a subset of the clinical downstream document classification tasks. The macro-averaged AUROC score is calculated by taking the average of AUROC scores across this subset of tasks. The *row in italic* indicates the model variant with the highest macro-averaged AUROC.

We compared our method with the state-of-the-art clinical outcome prediction model, BEEP (Naik et al., 2022), which leverages a retrieval augmentation technique to enhance the predictive capabilities of clinical language models. A small caveat is that BEEP focused on three downstream tasks: prolonged mechanical ventilation, mortality, and length of stay predictions. We selected the best-performing solution from BEEP, UmlsBERT with weighted voting retrieval augmentation, based on the averaged AUROC score to compare with our solution. While BEEP outperforms our approach, particularly in the prediction of PMV, it is crucial to emphasise that our method achieves its predictions without relying on retrieval augmentation. Future work may explore using retrieval augmentation on top of our proposed method.

D Training Configurations

We use HuggingFace's Transformers (Wolf et al., 2020) and PEFT (Mangrulkar et al., 2022) libraries for the experiments. All LLaMA-based models are trained on one NVIDIA A100-80GB GPU, while the baseline models are trained on a single NVIDIA GeForce GTX 1080 Ti-16GB GPU.

E Artefacts

The pretrained baseline models including BioClinicalBERT (Alsentzer et al., 2019), BlueBERT (Peng et al., 2019), and CORe (van Aken et al., 2021) were released under the Creative Commons designation CC0 1.0 Universal license, whereas UmlsBERT (Michalopoulos et al., 2021) was released under the MIT license. LLaMA (Touvron et al., 2023) was released under a noncommercial license.

MIMIC-III and MIMIC-IV dataset was released under the PhysioNet Credentialed Health Data License 1.5.0 and can only be accessed after one finishes the CITI Data or Specimens Only Research training[3].

[3] https://physionet.org/about/citi-course/