

Received 27 May 2025, accepted 16 June 2025, date of publication 24 June 2025, date of current version 1 July 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3582037

RESEARCH ARTICLE

The Impact of LoRA Adapters on LLMs for Clinical Text Classification Under Computational and Data Constraints

THANH-DUNG LE^{1,2}, (Senior Member, IEEE), TI TI NGUYEN^{1,2}, (Member, IEEE),
VU NGUYEN HA^{1,2}, (Senior Member, IEEE), SYMEON CHATZINOTAS^{1,2}, (Fellow, IEEE),
PHILIPPE JOUVET^{1,3}, AND RITA NOUMEIR¹, (Member, IEEE)

¹Biomedical Information Processing Laboratory, École de Technologie Supérieure, Montreal, QC H3C 1K3, Canada

²Interdisciplinary Centre for Security, Reliability, and Trust (SnT), University of Luxembourg, 4365 Esch-sur-Alzette, Luxembourg

³CHU Sainte-Justine Research Center, CHU Sainte-Justine Hospital, University of Montreal, Montreal, QC H3T 1J4, Canada

Corresponding author: Thanh-Dung Le (thanh-dung.le@uni.lu)

This work was supported in part by the Natural Sciences and Engineering Research Council (NSERC), in part by Institut de Valorisation des données de l'Université de Montréal (IVADO), and in part by Fonds de la recherche en sante du Quebec (FRQS).

ABSTRACT Fine-tuning Large Language Models (LLMs) for clinical Natural Language Processing (NLP) poses significant challenges due to domain gap, limited data, and stringent hardware constraints. In this study, we evaluate four adapter techniques—Adapter, Lightweight, TinyAttention, and Gated Residual Network (GRN) - equivalent to Low-Rank Adaptation (LoRA), for clinical note classification under real-world, resource-constrained conditions. All experiments were conducted on a single NVIDIA Quadro P620 GPU (2 GB VRAM, 512 CUDA cores, 1.386 TFLOPS FP32), limiting batch sizes to ≤ 8 sequences and maximum sequence length to 256 tokens. Our clinical corpus comprises only 580 000 tokens, several orders of magnitude smaller than standard LLM pre-training datasets. We fine-tuned three biomedical pre-trained LLMs (CamemBERT-bio, AliBERT, DrBERT) and two lightweight Transformer models trained from scratch. Results show that 1) adapter structures provide no consistent gains when fine-tuning biomedical LLMs under these constraints, and 2) simpler Transformers, with minimal parameter counts and training times under six hours, outperform adapter-augmented LLMs, which required over 1000 GPU-hours. Among adapters, GRN achieved the best metrics (accuracy, precision, recall, F1 = 0.88). These findings demonstrate that, in low-resource clinical settings with limited data and compute, lightweight Transformers trained from scratch offer a more practical and efficient solution than large LLMs, while GRN remains a viable adapter choice when minimal adaptation is needed.

INDEX TERMS Low-rank adaptation (LoRA), adapters, LLM, clinical NLP, cardiac failure, text classification.

I. INTRODUCTION

Currently, LLMs in natural language processing (NLP) have achieved remarkable advancements, evolving significantly over recent years. Before 2017, Long Short-Term Memory Networks (LSTMs) were the state-of-the-art in language modeling, reaching impressive scales of up to a billion parameters [1]. The introduction of the Transformer model

in 2017 marked a paradigm shift, leveraging the attention mechanism to set new benchmarks in NLP [2]. This innovation laid the groundwork for models such as GPT-2 [3] and GPT-3 [4], and further studies into the scaling laws for neural language models [5]. Today, Transformer-based architectures with self-attention mechanisms, exemplified by models like GPT-4, Claude 3, and Gemini, have become the standard for LLMs [6].

In the clinical domain, the decision support system (CDSS) at CHU Sainte-Justine (CHUSJ) aims to enhance

The associate editor coordinating the review of this manuscript and approving it for publication was Carmen C. Y. Poon¹.

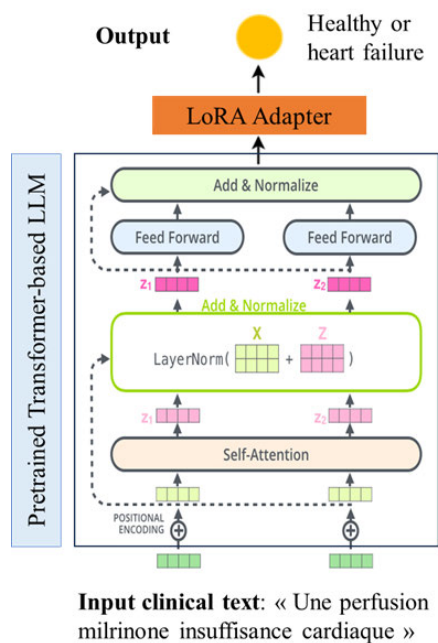


FIGURE 1. The visualized workflow for the experiment set-up with Transformer-based LLM [7] structure, and learnable layers from a LoRA Adapter, which is a lightweight auxiliary network that runs alongside the transformer layer, transforming their activations into a structured, hierarchical feature representation.

the diagnosis and management of acute respiratory distress syndrome (ARDS) in real-time by automatically analyzing data from electronic medical records, chest X-rays, and other sources. Previous research has highlighted that ARDS is often diagnosed late or missed in many patients, underscoring the need for more effective diagnostic tools [8]. Diagnosing ARDS requires identifying three main conditions: hypoxemia, chest X-ray infiltrates, and the absence of cardiac failure [9]. Furthermore, ARDS and cardiac failure frequently present with similar symptoms, making early and accurate diagnosis crucial for effective treatment strategies, particularly in critical care units like the Pediatric Intensive Care Unit (PICU). Accurately distinguishing between these conditions can significantly influence patient outcomes, potentially saving lives.

The research team at CHUSJ has developed advanced algorithms to detect hypoxemia [10], analyze chest X-rays [11], [12], and identify the absence of cardiac failure. Our research group has also extensively analyzed machine learning algorithms for detecting cardiac failure from clinical narratives using NLP techniques [13], [14]. Recent studies have demonstrated the superior performance of LLMs in handling complex tasks, such as understanding numerical attributes within clinical notes that contribute to cardiac failure, compared to traditional word embedding and deep learning methods [15], [16]. Implementing these advanced algorithms has the potential to significantly increase ARDS diagnosis rates and improve patient outcomes at CHUSJ.

However, while efforts have been made to adapt LLMs in these studies, the results have been limited, indicating

the need for further research and optimization to leverage LLM capabilities fully. Applying LLMs in clinical NLP remains challenging due to limited data availability and strict privacy regulations. Training must often be confined to protected environments within hospital servers, especially in CDSS environments that operate under constrained computational resources and inflexible data privacy policies. Despite promising results, these significant limitations persist. Consequently, this study empirically analyzes LLMs' adaptability within the CDSS framework at CHUSJ, aiming to enhance clinical decision-making and patient outcomes while navigating the challenges of data privacy and resource constraints.

In summary, as shown in Fig. 1, this study addresses the challenge of adapting LLMs for clinical note classification within the strict data, privacy, and compute constraints of the CHUSJ CDSS. Our primary objectives are to (i) empirically evaluate lightweight adapter structures (Adapter, Lightweight, TinyAttention, GRN) for fine-tuning pre-trained biomedical LLMs under these constraints, (ii) benchmark their performance against Transformer models trained from scratch on a limited 580 000-token corpus, and (iii) derive practical recommendations for deploying NLP models in resource-limited clinical settings. The main contributions of this work are:

- We conduct the first head-to-head comparison of four LoRA-equivalent adapter techniques on three biomedical LLMs (CamemBERT-bio, AliBERT, DrBERT) versus lightweight Transformers trained from scratch.
- We identify GRN as the top-performing adapter ($F1 = 0.88$) and demonstrate that simpler Transformers reach superior accuracy in under 6 GPU-hours.

The remainder of the paper is organized as follows. Section II reviews related work on adapter methods and clinical NLP. Section III describes our dataset, experimental setup, and adapter architectures. Section IV presents quantitative results and analysis. Section V discusses limitations and implications for CDSS integration and deployment. Finally, Section VI concludes and outlines future directions.

II. RELATED WORKS

One of the critical challenges with Transformer-based LLMs in clinical text classification is their difficulty in accurately interpreting short texts and their tendency to rely heavily on keywords [13]. In our recent research, we have explored various strategies to improve LLM performance in this domain. These strategies include utilizing Mixture of Experts (MoE) Transformers [17] and integrating adapters as intermediate layers to filter out irrelevant information [18]. Despite these efforts, these approaches did not surpass the performance of a simple MLP combined with a dense feature representation from an autoencoder [14]. This underperformance is attributed to a generalization gap between training and validation, especially with large models trained on small datasets. Additionally, other findings indicate that LLMs may

not consistently deliver superior results, particularly when considering accuracy, cost, and safety factors. As models become more complex and expensive, issues related to cost and accessibility become more pronounced, which are critical factors in the CDSS environment [19], [20].

Several potential approaches can be employed to address the challenges of using LLMs in clinical NLP with small, limited datasets. One effective strategy is instruction tuning, a parameter-efficient method that optimizes LLMs to follow specific instructions better, thereby aligning them to new domains [21]. Additionally, fine-tuning techniques can help unlock the capabilities of LLMs for various downstream applications, ensuring robust performance even with constrained data [22]. These strategies can significantly enhance the adaptability and effectiveness of LLMs in clinical settings where data availability is limited, ultimately improving their utility.

Two primary approaches are commonly employed in fine-tuning LLMs: full-model tuning (FMT) and parameter-efficient tuning (PET). PET includes methods such as prompt tuning and LoRA, which are especially relevant when the size of the LLM far exceeds the available fine-tuning data, a common scenario in data-limited environments [23]. Among these methods, LoRA is particularly notable due to its adaptability and ability to facilitate end-to-end customization during fine-tuning. LoRA freezes the pre-trained model weights and introduces trainable rank decomposition matrices into each layer of the Transformer architecture. This significantly reduces the number of trainable parameters required for downstream tasks. For example, compared to GPT-3 175B fine-tuned with Adam, LoRA can reduce the number of trainable parameters by 10,000 times and the GPU memory requirement by three times. LoRA performs as well as or better than traditional fine-tuning in terms of model quality on models such as RoBERTa, DeBERTa, GPT-2, and GPT-3, despite having fewer trainable parameters, higher training throughput, and no additional inference latency [24].

Adapter modules [25], [26] represent a form of LoRA efficient tuning, integrating small, newly initialized parameter modules at each transformer layer of pre-trained LLMs. These modules typically comprise a two-layer feed-forward neural network with a bottleneck structure. Specifically, the adapter structure includes (1) a down-projection layer with weights $W_{down} \in \mathbb{R}^{d \times r}$ that reduces the input h_i to a lower-dimensional space defined by the bottleneck dimension r ; and (2) an up-projection layer with weights $W_{up} \in \mathbb{R}^{r \times d}$ that projects the reduced input back to its original size. Mathematically, the adapter operation can be expressed as:

$$h_a = W_{up}^T f(W_{down}^T h_i) \quad (1)$$

where h_a is the output and $f(\cdot)$ represents the activation function. This configuration allows for efficient parameter updates during fine-tuning while maintaining the overall

structure and performance of the pre-trained LLMs. Therefore, this study aims to analyze the impact of different adapter structures, which offer minimal complexity and rapid adaptation to LLMs, for clinical NLP narrative classification. This implementation is designed to operate within constrained computational capacities, making it suitable for environments with limited computational resources. The choice of NVIDIA Quadro P620, with its significantly limited computational capabilities (512 CUDA cores and only 1.386 TFLOPS FP32 performance), imposes substantial computational constraints compared to high-performance GPUs such as the NVIDIA A100, as shown in Table 1. This selection reflects deliberate experimental conditions intended to replicate scenarios typical of resource-constrained environments, ensuring that developed models are robust and efficient under strict hardware limitations for fine-tuning LLMs directly on clinical texts.

III. MATERIALS AND METHODS

A. CLINICAL NOTES DATA AT CHUSJ

This study was conducted following ethical approval from the research ethics board at CHUSJ (protocol number: 2020-2253), and the study's design focused on identifying cardiac failure in patients within the first 24 hours of admission by analyzing admission and evolution notes during this initial period. The dataset consisted of 580,000 unigrams extracted from 5,444 single lines of short clinical narratives. Of these, 1,941 cases were positive (36% of the total), and 3,503 cases were negative. While the longest n-gram was over 400 words, most n-grams had a length distribution between 50 and 125 words. The average length of the number of characters was 601 and 704, and the average size of the number of digits was 25 and 26 for the positive and negative cases, respectively. We pre-processed the data by removing stop-words and accounting for negation in medical expressions. Numeric values for vital signs (heart rate, blood pressure, etc.) were also included and decoded to account for nearly 4% of the notes containing these values. All notes are short narratives; detailed characteristics for the notes at CHUSJ can be found in the Supplementary Materials from the study [13], [14].

In summarization, we apply the ScatterText [31] for the note visualization. In total, we have over 580,000 unigrams (n-gram) shown in Fig. 3. The figure shows the most frequent words for the positive case in the upper right corner; the most frequent words for the negative cases in the lower-left corner; and, all less frequent words for both cases are in the center. Besides, the top terms from the positive and negative cases are presented on the right-hand side. In positive cases, we quickly see that most of these terms are positively related to cardiac malfunction: milrinone or milri (milrinone), aorte or aortique valve (aortic valve). In contrast, terms such as respiratoire (respiratory), and IVRS (Infection des voies respiratoires supérieures - Virus responsible for respiratory distress) indicate respiratory syndromes.

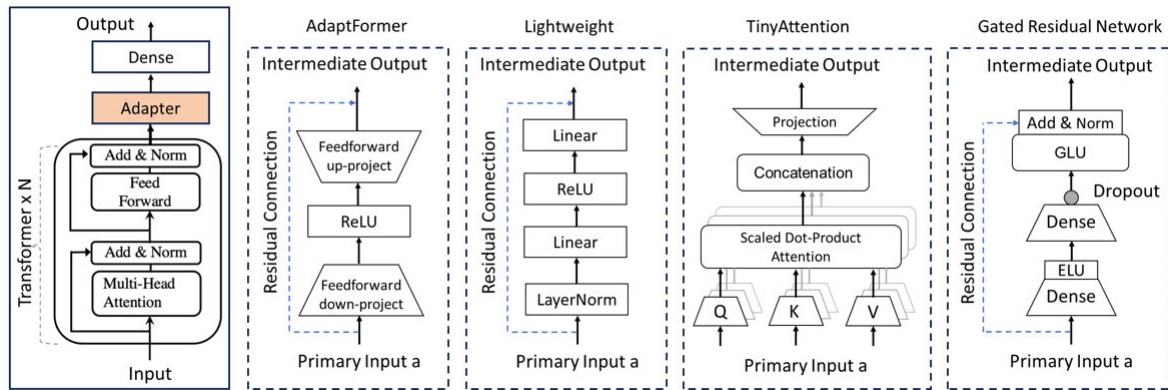


FIGURE 2. Different adapter structure, including AdaptFormer [27], Lightweight [28], TinyAttention [29], GRN [18], [30].

TABLE 1. Comparison between NVIDIA Quadro P620 and NVIDIA A100 80GB PCIe.

Specifications	NVIDIA Quadro P620	NVIDIA A100 80GB PCIe
Architecture	Pascal (GP108)	Ampere (GA100)
CUDA Cores	512	6,912
Tensor Cores	n/a	432 (3rd-gen, TF32 & FP64 support)
Memory	2 GB GDDR5	80 GB HBM2e
Computing Performance	<ul style="list-style-type: none"> FP32: 1.386 TFLOPS 	<ul style="list-style-type: none"> FP32: 19.5 TFLOPS FP16 (Tensor): 312 TFLOPS (624 w/sparsity) INT8 (Tensor): 624 TOPS (1,248 w/sparsity)
Memory Bus / Bandwidth	128-bit / 48 GB/s	~ / 2,039 GB/s
Thermal Design Power (TDP)	40 W	250 W
Key Strengths	<ul style="list-style-type: none"> Small, power-efficient (40 W TDP) 	<ul style="list-style-type: none"> Massive Tensor performance (312 TFLOPS) Multi-Instance GPU (up to 7 instances) High PCIe 4.0 bandwidth

B. BIOMEDICAL PRETRAINED LLM

In recent years, the development of biomedical pre-trained BERT-based models has significantly advanced the processing and understanding of biomedical text, particularly within the French language domain. As confirmed in [17], three notable among these models are CamemBERT-bio [32], DrBERT [33], and AliBERT [34], each tailored to address the unique challenges of biomedical text analysis. CamemBERT-bio, for instance, is designed explicitly for French biomedical data, leveraging the robust architecture of CamemBERT to deliver superior performance in this field. Similarly, DrBERT and AliBERT enhance the landscape of specialized models by offering high accuracy and efficiency in various biomedical NLP tasks. These models are exceptionally well-suited for classifying French clinical notes, having been trained on extensive French biomedical corpora. These models are particularly adept at classifying French clinical notes due to their training in extensive French biomedical corpora, which enables them to accurately capture the nuances and specialized terminology unique to French medical practice.

C. TRANSFORMER-BASED MODELS

Training Transformer models effectively with small datasets presents a significant challenge. Transformers often exhibit limitations such as a generalization gap and sharp minima when applied to small datasets [13]. Furthermore, their performance degrades on imbalanced and small clinical datasets [35]. Our recent study indicates that the Mixture-

of-Experts (MoE) Transformer [17] can mitigate some of these limitations by enhancing model performance with limited data. In this study, we will experiment with the standard Transformer and the MoE-Transformer for clinical text classification tasks to evaluate their effectiveness in handling small and imbalanced datasets.

D. ADAPTERS STRUCTURES

Employing limited computational constraints and limited data, this study limited the experiment to the following adapter structure, which is simple and scalable for effectively fine-tuning the pre-trained model as the LoRA technique [36], as shown in Fig. 2. Below we present the formal derivations for each adapter, specifying how an input token embedding $\mathbf{x} \in \mathbb{R}^d$ is transformed into an output $\mathbf{y} \in \mathbb{R}^d$.

1) ADAPTFORMER [27]

AdaptFormer is a parameter-efficient tuning module for Transformer architectures that enhances adaptability by incorporating a feedforward down-projection layer, a ReLU activation, and an up-projection layer to restore input size. It includes a residual connection to preserve the original input, improving learning without significantly increasing model complexity. By updating only the adapter modules' parameters, AdaptFormer enables effective fine-tuning while keeping the pre-trained model fixed.

$$\mathbf{h}_{\downarrow} = W_{\downarrow} \mathbf{x}, \quad W_{\downarrow} \in \mathbb{R}^{r \times d}, \quad (2)$$

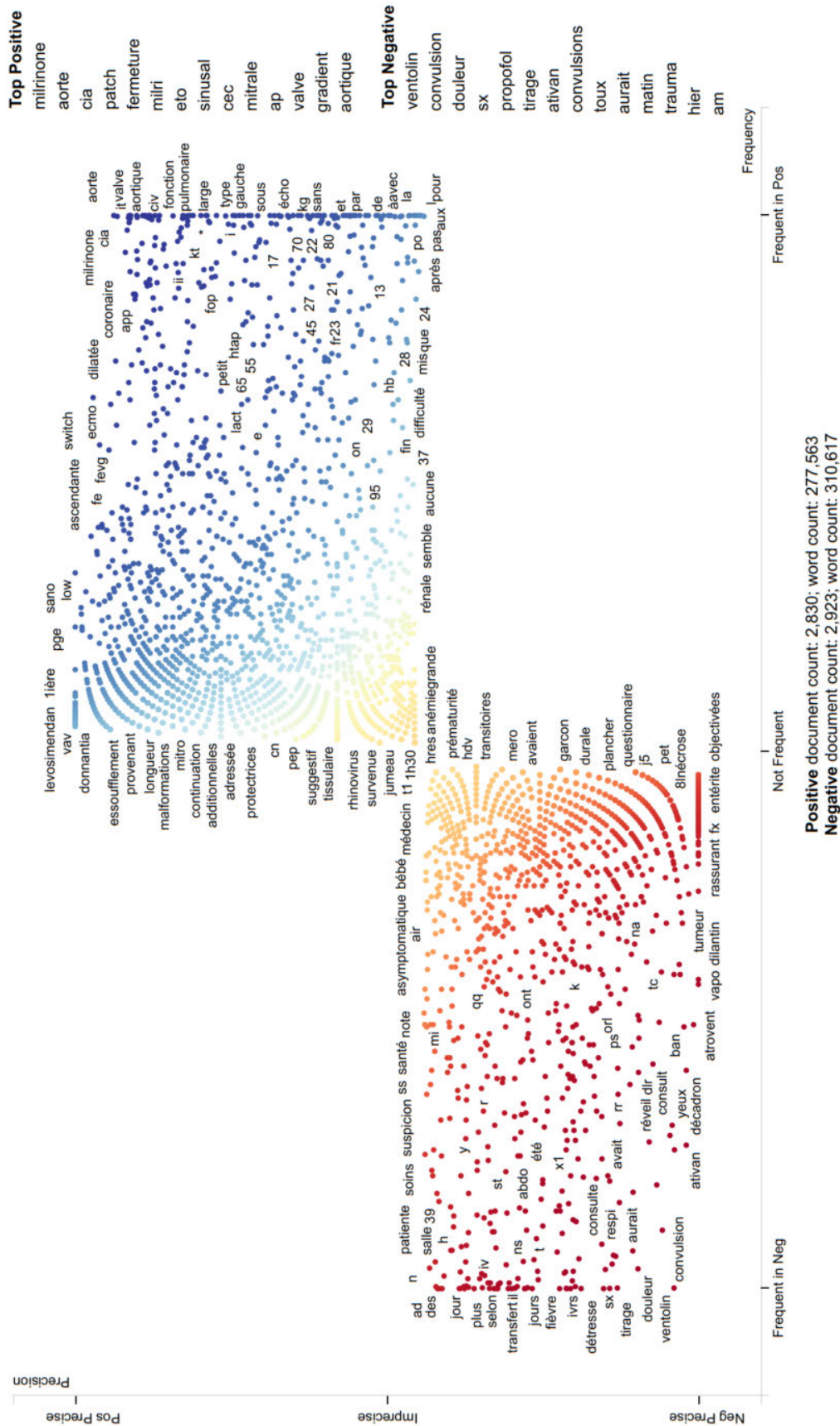


FIGURE 3. Overview of word distribution in clinical notes at CHUJ [13].

TABLE 2. Comparison of adapters for transformer with computational complexity.

Models	Structure Summarization	Complexity (FLOPs)	Highlights
AdaptFormer	<ul style="list-style-type: none"> Feedforward down-project: $M \rightarrow D$ ReLU: D Feedforward up-project: $D \rightarrow M$ 	$4 \cdot M \cdot D + D$	Least complex if D is much smaller than M
Lightweight	<ul style="list-style-type: none"> LayerNorm: M Linear: $M \rightarrow D$ ReLU: D Linear: $D \rightarrow M$ 	$4 \cdot M \cdot D + D + 2 \cdot M$	Dominated by $4 \cdot M \cdot D$; additional small terms
TinyAttention	<ul style="list-style-type: none"> Q, K, V projections: $M \rightarrow \frac{D}{h}$ Scaled dot-product attention Concatenation Projection: D 	$6 \cdot \frac{M \cdot D}{h} + 4 \cdot N \cdot D + \frac{2 \cdot N^2 \cdot D}{h} + 2 \cdot M \cdot D$	Complexity highly dependent on h and N
GRN	<ul style="list-style-type: none"> Dense: $M \rightarrow D$ ELU: D Dense: $D \rightarrow M$ Dropout: D GLU: Element-wise multiplication (gate) Add & Norm: M 	$4 \cdot M \cdot D + 4 \cdot D + 2 \cdot M$	Similar leading term as Lightweight; slightly higher additional terms

$$\mathbf{h}_{\text{act}} = \phi(\mathbf{h}_{\downarrow}), \quad \phi \in \{\text{ReLU}, \text{GELU}\}, \quad (3)$$

$$\mathbf{h}_{\uparrow} = W_{\uparrow} \mathbf{h}_{\text{act}}, \quad W_{\uparrow} \in \mathbb{R}^{d \times r}, \quad (4)$$

$$\mathbf{y} = \mathbf{x} + \mathbf{h}_{\uparrow}. \quad (5)$$

2) LIGHTWEIGHT [28]

The Lightweight adapter structure integrates a linear down-projection layer followed by a ReLU activation, a second linear layer, and a final LayerNorm for normalization. This configuration is enhanced with a residual connection to maintain the original input alongside the processed output. By focusing on linear transformations and normalization, this adapter efficiently fine-tunes the model with minimal additional parameters, ensuring lightweight adaptability.

$$\mathbf{y} = \mathbf{x} + W_{\ell} \mathbf{x}, \quad W_{\ell} \in \mathbb{R}^{d \times r} \quad (6)$$

3) TINYATTENTION [29]

The TinyAttention adapter structure incorporates scaled dot-product attention, where the primary input is split into query (Q), key (K), and value (V) components. The attention mechanism calculates attention weights and produces a weighted sum of the values, which are then concatenated and passed through a projection layer. This structure allows the model to focus on relevant input parts efficiently, enhancing the representation with minimal additional parameters.

$$\mathbf{Q}' = W_Q \mathbf{x}, \quad W_Q \in \mathbb{R}^{r \times d}, \quad (7)$$

$$\mathbf{K}' = W_K \mathbf{x}, \quad W_K \in \mathbb{R}^{r \times d}, \quad (8)$$

$$\mathbf{V}' = W_V \mathbf{x}, \quad W_V \in \mathbb{R}^{r \times d}, \quad (9)$$

$$\mathbf{A} = \text{softmax}(\mathbf{Q}' \mathbf{K}'^T / \sqrt{r}) \mathbf{V}', \quad (10)$$

$$\mathbf{y} = \mathbf{x} + W_O \mathbf{A}, \quad W_O \in \mathbb{R}^{d \times r} \quad (11)$$

4) GATED RESIDUAL NETWORKS (GRN) [18], [30]

The GRN adapter structure includes a series of dense layers. The primary input is first processed through an ELU activation function and a dense layer. The output then passes through a dropout layer and another dense layer before being gated by a gated linear unit (GLU), \odot is the element-wise Hadamard product. Finally, the gated output is added to the original input via a residual connection, followed by normalization (Add & Norm), enhancing the model's ability to learn complex representations efficiently while maintaining stability.

$$\mathbf{u} = \text{ELU}(W_1 \mathbf{x}), \quad W_1 \in \mathbb{R}^{r \times d}, \quad (12)$$

$$\mathbf{v} = W_2 \mathbf{u}, \quad W_2 \in \mathbb{R}^{d \times r}, \quad (13)$$

$$\mathbf{g} = \sigma(W_g \mathbf{v}), \quad W_g \in \mathbb{R}^{d \times d}, \quad (14)$$

$$\mathbf{h} = \mathbf{g} \odot \mathbf{v}, \quad (15)$$

$$\mathbf{y} = \text{LayerNorm}(\mathbf{x} + \mathbf{h}) \quad (16)$$

For the complexity of each adapter structure, Table 2 compares the computational complexity of different adapter structures for Transformers. AdaptFormer, with its simple feedforward layers and ReLU activation, has the least complexity when the bottleneck dimension D is significantly smaller than the model dimension M . The Lightweight adapter adds LayerNorm and utilizes linear transformations, resulting in a complexity dominated by $4 \cdot M \cdot D$. TinyAttention introduces attention mechanisms, making its complexity highly dependent on the number of heads h and sequence length N . The GRN includes dense layers, ELU activation, dropout, and a gated linear unit, leading to a complexity similar to the Lightweight adapter but with slightly higher additional terms. These approaches are particularly suitable for fine-tuning a pre-trained LLM on a limited dataset and under constrained computational capacity.

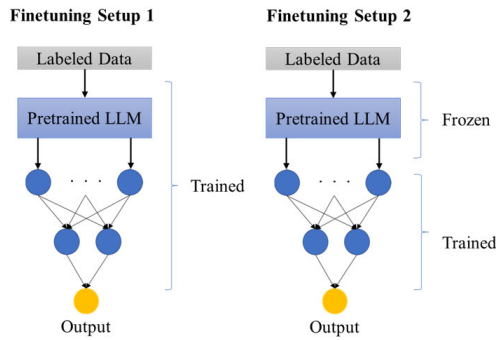


FIGURE 4. Experiment setup.

IV. EXPERIMENTAL RESULTS

We employed two fine-tuning approaches for the experimental setup, as illustrated in Figure 4. In **Fine-tuning Setup 1**, the entire pre-trained language model (LLM) is fine-tuned using labeled data, where both the pre-trained LLM and the additional layers are trained simultaneously. This method allows the model to fully adapt to the specific task by updating all parameters, potentially leading to better performance, but it is computationally intensive. In **Fine-tuning Setup 2**, the pre-trained LLM is frozen, meaning its parameters are not updated during fine-tuning, and only the additional layers appended to the LLM are trained using labeled data. This approach reduces computational requirements and mitigates the risk of overfitting, making it more suitable for scenarios with limited data and computational resources. By comparing these setups, we aim to evaluate the effectiveness and efficiency of fine-tuning strategies for clinical text classification tasks.

Table 3 compares four models: AdaptFormer, Lightweight, GRN, and TinyAttention. For each model, it outlines its specifications and the total number of parameters, including the memory footprint in megabytes (MB). AdaptFormer uses a down projection dimension of 512 and an up projection dimension of 1024, totaling 1,839,618 parameters (7.02 MB). The Lightweight model has an input dimension of 1024 with 2,890,754 parameters (11.03 MB). GRN features an input dimension 1024 with a drop-out rate of 0.5, amounting to 3,940,354 parameters (15.03 MB). Finally, TinyAttention, which includes an input dimension of 1024, four heads, and a drop-out rate of 0.25, has the highest number of parameters at 9,188,354 (35.05 MB).

All experiments were conducted on the Intel(R) Xeon(R) CPU E3-1225, 3.30GHz, 16GB RAM, and Nvidia Quadro P620 GPU, 2GB. For the implementation, experiments were implemented using the scikit-learn library [37], and Keras [38]. The data was divided into 70% training and 30% testing. Moreover, training and fine-tuning the Transformer-based model is complex. As reported by [39], model size, learning rate, batch size, and maximum sequence length are the four critical hyperparameters that significantly influence the training process of the Transformer model. In addition, we also applied dropout [40] ($p=0.25$) and GlorotNormal

kernel initializer [41], batch normalization [42], [43] are employed for models' stability. Additionally, we also apply early stopping based on the validation loss. Consequently, these hyperparameters were carefully chosen to achieve optimal performance and prevent overfitting.

To effectively assess the performance of our method, metrics including accuracy, precision, recall (or sensitivity), and F1 score were used [44]. These metrics are defined as follows:

$$\text{Accuracy (acc)} = \frac{TP + TN}{TP + TN + FP + FN} \quad (17)$$

$$\text{Precision (pre)} = \frac{TP}{TP + FP} \quad (18)$$

$$\text{Recall/Sensitivity (rec)} = \frac{TP}{TP + FN} \quad (19)$$

$$\text{F1-Score (f1)} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (20)$$

where TN and TP stand for true negative and true positive, respectively, and they are the number of negative and positive patients that are classified correctly. Whereas FP and FN represent false positive and false negative, respectively, and they represent the number of positive and negative patients that were wrongly predicted.

As shown in Fig. 5, both the GRN-Transformer and the GRN-MoE Transformer, trained from scratch, converge rapidly, exceeding 80% accuracy by the 5th epoch, and maintain a narrow train-validation gap throughout. The MoE variant achieves a higher peak training accuracy ($\approx 91.5\%$ vs. 89%) and reaches $\approx 88\%$ validation accuracy by epoch 24, indicating that expert routing effectively regularizes training. These results demonstrate that both architectures deliver strong predictive performance with minimal overfitting, making them well-suited for resource-constrained clinical tasks.

First of all, the experimental results compare the performance of various adapters applied to biomedical pre-trained LLMs (CamemBERT-bio, AliBERT, DrBERT) and Transformer-based models trained from scratch, evaluated on accuracy, precision, recall, F1 score, training time, and inference time as summarized in Table 4 to 8, respectively. In Setup 1, where full fine-tuning was applied, the baseline models achieved the highest performance across most metrics, with the GRN adapter showing competitive results. However, in Setup 2, where pre-trained weights were frozen, and only the adapters were fine-tuned, there was a significant performance decline across all adapters, demonstrating lower accuracy, precision, recall, and F1 scores. Notably, each experiment with biomedical pre-trained LLMs required extensive training times ranging from 30 to 50 hours, whereas Transformer-based models trained from scratch completed training in under an hour. This stark contrast highlights the practicality of simpler Transformer-based models for clinical NLP tasks in resource-constrained environments. While adapters like GRN can enhance performance, their benefits are diminished by the

TABLE 3. Model specifications and total parameters for adapter structures.

Model	Specifications	Total parameters
AdaptFormer	Down projection dimension 512; Up projection dimension 1024	1,839,618 (7.02 MB)
Lightweight	Input dimension 1024	2,890,754 (11.03 MB)
GRN	Input dimension 1024, drop-out at 0.5	3,940,354 (15.03 MB)
TinyAttention	Input dimension 1024, 4 heads, and drop-out at 0.25	9,188,354 (35.05 MB)

TABLE 4. Performance comparison of CamemBERT-bio with different adapters.

CamemBERT-bio	Acc (↑)	Pre (↑)	Rec (↑)	F1 (↑)	Training Time (hours) (↓)	Inference Time (s) (↓)
Setup 1						
Baseline	0.87	0.86	0.88	0.87	31.7	142
Adapter	0.84	0.85	0.83	0.84	22.5	130
Lightweight	0.84	0.83	0.82	0.82	30.8	121
GRN	0.86	0.83	0.91	0.87	27	131
TinyAttention	0.83	0.79	0.88	0.83	23	122
Setup 2						
Adapter	0.74	0.78	0.72	0.75	22.5	130
Lightweight	0.72	0.71	0.72	0.71	35.9	126
GRN	0.71	0.7	0.76	0.73	54.1	123
TinyAttention	0.71	0.7	0.72	0.71	41.6	127

Bold denotes the best values.

TABLE 5. Performance comparison of ALiBERT with different adapters.

ALiBERT	Acc (↑)	Pre (↑)	Rec (↑)	F1 (↑)	Training Time (hours) (↓)	Inference Time (s) (↓)
Setup 1						
Baseline	0.86	0.87	0.84	0.86	39.3	128
Adapter	0.78	0.72	0.88	0.79	42.8	128
Lightweight	0.84	0.82	0.85	0.84	34.8	131
GRN	0.87	0.84	0.84	0.85	30.8	127
TinyAttention	0.84	0.81	0.83	0.82	42.2	128
Setup 2						
Adapter	0.68	0.67	0.67	0.67	46.6	126
Lightweight	0.66	0.72	0.49	0.58	46.7	126
GRN	0.67	0.7	0.59	0.64	37.5	126
TinyAttention	0.67	0.67	0.7	0.68	46.2	130

Bold denotes the best values.

TABLE 6. Performance comparison of DrBERT with different adapters.

DrBERT	Acc (↑)	Pre (↑)	Rec (↑)	F1 (↑)	Training Time (hours) (↓)	Inference Time (s) (↓)
Setup 1						
Baseline	0.87	0.84	0.9	0.87	45.2	133
Adapter	0.86	0.87	0.87	0.87	38.5	126
Lightweight	0.71	0.78	0.56	0.65	41.6	132
GRN	0.86	0.84	0.88	0.86	41.3	130
TinyAttention	0.85	0.81	0.9	0.85	28.9	123
Setup 2						
Adapter	0.69	0.71	0.63	0.67	46.1	122
Lightweight	0.73	0.72	0.73	0.72	45	125
GRN	0.73	0.7	0.76	0.73	47	126
TinyAttention	0.75	0.76	0.69	0.72	47.5	123

Bold denotes the best values.

TABLE 7. Performance comparison of transformer with different adapters.

Transformer	Acc (↑)	Pre (↑)	Rec (↑)	F1 (↑)	Training Time (hours) (↓)	Inference Time (s) (↓)
Baseline	0.85	0.85	0.83	0.84	0.11	3
Adapter	0.85	0.83	0.85	0.84	0.4	2
Lightweight	0.85	0.82	0.88	0.85	1	3
GRN	0.87	0.85	0.89	0.87	0.7	3
TinyAttention	0.85	0.81	0.88	0.84	0.7	3

Bold denotes the best values.

TABLE 8. Performance comparison of MoE-Transformer with different adapters.

MoE-Transformer	Acc (↑)	Pre (↑)	Rec (↑)	F1 (↑)	Training Time (hours) (↓)	Inference Time (s) (↓)
Baseline	0.87	0.87	0.85	0.86	0.17	4
Adapter	0.84	0.78	0.92	0.84	0.4	2
Lightweight	0.84	0.76	0.95	0.84	1.2	4
GRN	0.88	0.88	0.88	0.88	0.8	3
TinyAttention	0.84	0.79	0.9	0.84	0.8	3

Bold denotes the best values.

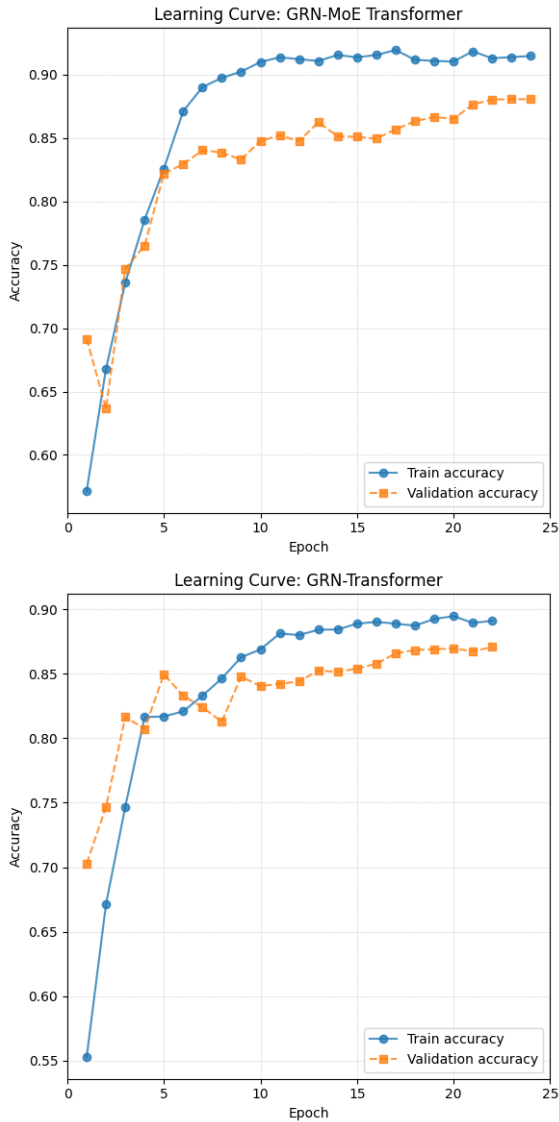


FIGURE 5. Learning curve performance of Transformer and MoE-Transformer with the GRN adapter.

substantial training times and limited improvements observed in scenarios with frozen weights and limited data.

As summarized in Table 4, 5, 6, the experimental results compare different adapters for various biomedical pre-trained models (CamemBERT-bio, AliBERT, and DrBERT, respectively) based on accuracy, precision, recall, and F1 score, with full fine-tuning as the baseline. For CamemBERT-bio, the baseline achieved high performance across all metrics,

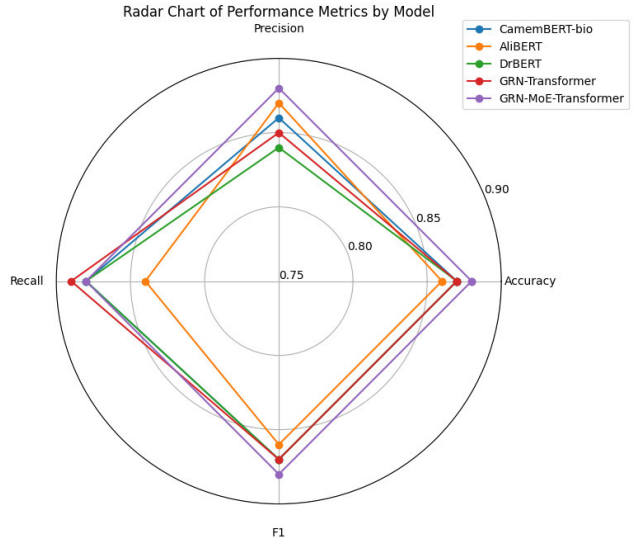


FIGURE 6. Performance comparison between biomedical pre-trained LLMs vs. Transformer-based models with different adapters.

while TinyAttention closely matched the baseline, and AdaptFormer and Lightweight showed slight reductions in recall. GRN achieved slightly higher recall than Lightweight. For AliBERT, the baseline exhibited strong performance, especially in precision. AdaptFormer and Lightweight had noticeable drops in recall but maintained high precision and accuracy. GRN provided balanced performance, and TinyAttention closely matched the baseline in accuracy and precision. For DrBERT, the baseline again delivered strong results. AdaptFormer and Lightweight showed decreased recall and F1 scores, while GRN demonstrated higher recall and comparable precision to the other adapters. TinyAttention matched the baseline in accuracy and precision with a slight decrease in recall. While full fine-tuning (baseline) provided the best performance, GRN adapters balanced performance and computational efficiency, making them suitable for scenarios with limited computational resources. However, there were no significant improvements when adapters were used to fine-tune the pre-trained model with limited data. In some cases, it degraded performance, as seen with the AdaptFormer and Lightweight adapters in AliBERT and DrBERT, respectively.

The experimental results compare the performance of Transformers and MoE-Transformers using different adapters evaluated based on accuracy, precision, recall, and F1 score, as shown in Table 7, and 8, respectively.

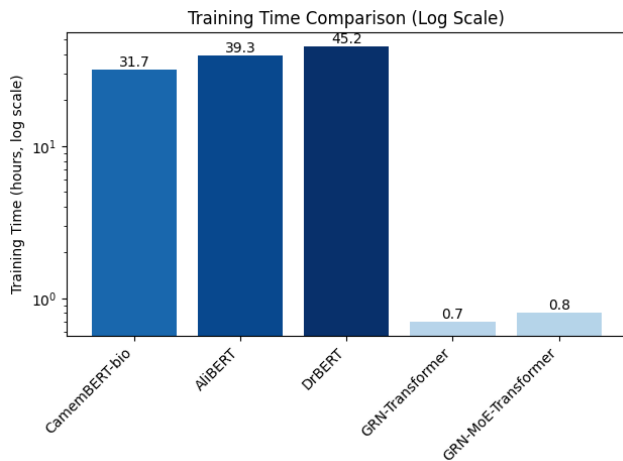


FIGURE 7. Training time comparison between biomedical pre-trained LLMs vs. Transformer-based models with different adapters.

The baseline models, trained from scratch without any adapters, provide a reference point against which the other models, also trained from scratch but with different adapters, are compared. These setups differ from using biomedical pre-trained models. From the results, two key points emerge. First, adapters help the Transformer, as all the adapters outperform the Transformer baseline. Second, with the more complex MoE-Transformer, adapters do not continually improve the MoE-Transformer baseline model; for instance, AdaptFormer and Lightweight show no significant improvement. However, both GRN and TinyAttention improve the MoE-Transformer compared to the baseline. Overall, GRN is the most effective technique, as it enhances the performance of both the Transformer and MoE-Transformer models.

Based on the results for biomedical pre-trained models and Transformer-based models trained from scratch, we compared the best performance of pre-trained models with GRN adapters to that of Transformer-based models with GRN adapters, as illustrated in Fig. 6. The results indicate no significant difference between fine-tuning the pre-trained models with adapters and applying adapters to Transformer-based models trained from scratch. This suggests that the advantage of using adapters for fine-tuning pre-trained models is unclear in scenarios with limited data. While adapters like GRN can improve model performance, their impact cannot distinguish between pre-trained models and those trained from scratch under data constraints. Overall, the benefit of employing adapters to fine-tune the pretrained LLM in limited data scenarios remains ambiguous.

The GRN is designed to enhance neural networks' capabilities by integrating sophisticated gating mechanisms. At its core is the Gated Linear Unit, which combines a linear transformation with a sigmoid gated transformation, achieved through dense layers, and performs an element-wise multiplication of the linear and gated outputs. This mechanism ensures the network can dynamically control the information

```
class Lightweight(layers.Layer):
    def __init__(self, units, activation='relu'):
        super().__init__()
        # The first linear transformation
        self.linear1 = layers.Dense(units)
        # The activation function, ReLU in this case
        self.activation = layers.Activation(activation)
        # The second linear transformation
        self.linear2 = layers.Dense(units)
        # The layer normalization
        self.layer_norm = layers.LayerNormalization(epsilon=1e-6)

    def call(self, inputs):
        # Pass the inputs through the first linear layer
        x = self.linear1(inputs)
        # Apply the activation function
        x = self.activation(x)
        # Pass the result through the second linear layer
        x = self.linear2(x)
        # Apply layer normalization
        x = self.layer_norm(x)
        # Add the original inputs to the output (residual connection)
        return inputs + x
```

FIGURE 8. Pseudocode for lightweight adapter.

```
class GatedLinearUnit(layers.Layer):
    def __init__(self, units):
        self.linear = layers.Dense(units)
        self.sigmoid = layers.Dense(units, activation="sigmoid")

    def call(self, inputs):
        return self.linear(inputs) * self.sigmoid(inputs)

class GatedResidualNetwork(layers.Layer):
    def __init__(self, units, dropout_rate):
        self.elu_dense = layers.Dense(units, activation="elu")
        self.linear_dense = layers.Dense(units)
        self.dropout = layers.Dropout(dropout_rate)
        self.gated_linear_unit = GatedLinearUnit(units)
        self.layer_norm = layers.LayerNormalization()
        self.project = layers.Dense(units)

    def call(self, inputs):
        x = self.elu_dense(inputs)
        x = self.linear_dense(x)
        x = self.dropout(x)
        if inputs.shape[-1] != self.units:
            inputs = self.project(inputs)
        x = inputs + self.gated_linear_unit(x)
        return self.layer_norm(x)
```

FIGURE 9. Pseudocode for GRN LoRA adapter.

flow, enhancing its learning capabilities. Building upon this, the GatedResidualNetwork incorporates several key components: an ELU-activated dense layer that introduces non-linearity, a linear dense layer for further transformation, and a dropout layer to prevent overfitting. The gated linear unit is central to the GRN's function, which applies the gating mechanism to the residual connections. Additionally, layer normalization stabilizes and accelerates the training process, while a projection layer ensures that the input dimensionality matches the required units. Together, these elements form a robust architecture capable of effectively managing and transforming complex input data.

In contrast, the experimental results comparing training times for biomedical pre-trained models and Transformer-based models trained from scratch reveal significant differences. As shown in Fig. 7, fine-tuning pre-trained models like CamemBERT-bio, AlibiBERT, and DrBERT with adapters takes substantially longer, ranging from 30 to 50 hours per experiment. In contrast, training transformer-based models

with GRN adapters from scratch takes less than an hour. This highlights a crucial limitation of employing adapters for pre-trained models: the extensive computational capacity and training time required. Despite the adapters not showing significant performance improvements over training Transformer-based models from scratch, they demand significantly more computational resources and time. This makes using adapters in pre-trained models less appealing, especially in scenarios with limited computational resources and time constraints.

Our extensive experiments underscore the importance of carefully designing and implementing LoRA adapters when fine-tuning LLMs with limited data. These findings are consistent with recent research in other domains, such as programming and mathematics, as reported by [45]. Their study shows that LoRA often underperforms full fine-tuning across various settings. Nonetheless, LoRA provides beneficial regularization, preserving the base model's performance on tasks outside the target domain more effectively than full fine-tuning. Moreover, LoRA offers stronger regularization compared to techniques like weight decay and dropout and supports maintaining more diverse outputs. Full fine-tuning tends to learn perturbations with a significantly higher rank (10-100 times) than typical LoRA configurations, which likely contributes to the performance differences observed. Consequently, it is crucial to exercise caution when applying LoRA adapters and fine-tuning pre-trained LLMs, particularly in sensitive domains like clinical NLP, where data privacy, limited data availability, and computational resource constraints are significant concerns. As McCoy et al. [46] caution, inserting LLM-generated text directly into medical records could undermine communication, transparency, and the quality of healthcare, underscoring the need for caution in fine-tuning LLMs in clinical settings.

V. LIMITATIONS AND FUTURE WORKS

While our study systematically evaluates adapter techniques under stringent compute (single Quadro P620) and data constraints (580,000 tokens), several limitations remain. First, due to our reliance on a single GPU, we lacked the memory capacity and compute throughput required to fine-tune large proprietary models like Deepseek and Grok [47]. Replicating those experiments would have exceeded our hardware limits, both in terms of GPU memory and acceptable training time, making such evaluations infeasible within our study's resource constraints. Second, we only compared fully trainable versus fully frozen backbones with adapters; intermediate freezing ratios (e.g., 40-80% of layers frozen) may affect convergence speed and generalization, but were not explored. Third, our performance analysis focused primarily on accuracy and training time; other metrics such as memory footprint, latency, and energy consumption in diverse hospital server environments were not measured. Finally, we did not benchmark against the latest state-of-the-art clinical NLP models (e.g., Mamba state-space

architectures [48]) or other recent LoRA variants [49], limiting our comparisons.

To address these gaps, future research should:

- Investigate hybrid fine-tuning schemes that freeze varying ratios of the LLM layers while training the rest with adapters, evaluating how varying freeze ratios influence performance, convergence, and overfitting.
- Measure GPU memory, inference latency, and energy consumption on diverse hardware (e.g., A100) to quantify adapter trade-offs, and broaden our benchmarks to include cutting-edge LLMs for a full-spectrum clinical NLP evaluation.
- Benchmark our adapter and scratch-trained Transformer models against emerging state-space architectures such as Mamba and novel LoRA variants to position our findings within the current clinical NLP landscape.

VI. CONCLUSION

Based on our comprehensive evaluation, this study concludes that employing adapter structures for fine-tuning biomedical pre-trained LLMs does not yield significant improvements in clinical NLP tasks under resource constraints. We found that simpler Transformer-based models trained from scratch perform comparably or better, especially in environments with limited computational resources and data availability. Among the adapter structures evaluated, the GRN demonstrated superior accuracy, precision, recall, and F1 score, making it the most effective adapter for enhancing clinical note classification. Furthermore, the stark contrast in training times - over 1000 hours for pre-trained LLMs versus under 6 hours for Transformer-based models - underscores the practicality of using simpler models in resource-constrained settings. This study contributes to the field by providing a viable solution for clinical NLP tasks in low-resource environments and identifying the GRN adapter as a practical approach to improve model performance without requiring extensive computational resources. Lastly, implementing the adapters with different algorithms is straightforward for reproducibility, as demonstrated by the pseudocode examples provided by Fig. 8 and 9.

ACKNOWLEDGMENT

The clinical data were provided by the Research Center at CHU Sainte-Justine Hospital. Data and reproducible codes are available upon reasonable request to Prof. Philippe Juvet, M.D., Ph.D. (e-mail: philippe.juvet.med@ssss.gouv.qc.ca).

REFERENCES

- [1] R. Jozefowicz, O. Vinyals, M. Schuster, N. Shazeer, and Y. Wu, "Exploring the limits of language modeling," 2016, *arXiv:1602.02410*.
- [2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Jun. 2017, pp. 5998–6008.
- [3] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

- [4] T. B. Brown, "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2020, pp. 1877–1901.
- [5] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," 2020, *arXiv:2001.08361*.
- [6] J. Achiam et al., "GPT-4 technical report," 2023, *arXiv:2303.08774*.
- [7] J. Alammam, "The illustrated transformer," in *The Illustrated Transformer—Jay Alammam—Visualizing Machine Learning One Concept At a Time*, vol. 27, 2018, pp. 1–2. [Online]. Available: <https://jalammam.github.io/illustrated-transformer/>
- [8] G. Bellani, J. G. Laffey, T. Pham, E. Fan, L. Brochard, A. Esteban, L. Gattinoni, F. van Haren, A. Larsson, D. F. McAuley, M. Ranieri, G. Rubinfeld, B. T. Thompson, H. Wrigge, A. S. Slutsky, and A. Pesenti, "Epidemiology, patterns of care, and mortality for patients with acute respiratory distress syndrome in intensive care units in 50 countries," *JAMA*, vol. 315, no. 8, pp. 788–800, Feb. 2016.
- [9] Pediatric Acute Lung Injury Consensus Conference Group, "Pediatric acute respiratory distress syndrome: Consensus recommendations for the pediatric acute lung injury consensus conference," *Pediatric Crit. Care Med.*, *J. Soc. Crit. Care Med. World Fed. Pediatric Intensive Crit. Care Societies*, vol. 22, no. 5, p. 299, May 2015.
- [10] M. Sauthier, G. Tuli, P. A. Juvet, J. S. Brownstein, and A. G. Randolph, "Estimated PaO₂: A continuous and noninvasive method to estimate PaO₂ and oxygenation index," *Crit. Care Explor.*, vol. 3, no. 10, p. e0546, Sep. 2021.
- [11] N. Zaglam, P. Juvet, O. Flechelles, G. Emeriaud, and F. Cheriet, "Computer-aided diagnosis system for the acute respiratory distress syndrome from chest radiographs," *Comput. Biol. Med.*, vol. 52, pp. 41–48, Sep. 2014.
- [12] M. Yahyatabar, P. Juvet, and F. Cheriet, "Dense-unet: A light model for lung fields segmentation in chest X-ray images," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2020, pp. 1242–1245.
- [13] T.-D. Le, R. Noumeir, J. Rambaud, G. Sans, and P. Juvet, "Detecting of a Patient's condition from clinical narratives using natural language representation," *IEEE Open J. Eng. Med. Biol.*, vol. 3, pp. 142–149, 2022.
- [14] T.-D. Le, R. Noumeir, J. Rambaud, G. Sans, and P. Juvet, "Adaptation of autoencoder for sparsity reduction from clinical notes representation learning," *IEEE J. Transl. Eng. Health Med.*, vol. 11, pp. 469–478, 2023.
- [15] B. Aser Lompo, T.-D. Le, P. Juvet, and R. Noumeir, "Are medium-sized transformers models still relevant for medical records processing?" 2024, *arXiv:2404.10171*.
- [16] B. Aser Lompo and T.-D. Le, "Multi-objective representation for numbers in clinical narratives: A CamemBERT-Bio-based alternative to large-scale LLMs," 2024, *arXiv:2405.18448*.
- [17] T.-D. Le, P. Juvet, and R. Noumeir, "Improving transformer performance for French clinical notes classification using mixture of experts on a limited dataset," 2023, *arXiv:2303.12892*.
- [18] T.-D. Le, C. Macabiau, K. Albert, P. Juvet, and R. Noumeir, "GRN-transformer: Enhancing motion artifact detection in PICU photoplethysmogram signals," 2023, *arXiv:2308.03722*.
- [19] J. Fields, K. Chovanec, and P. Madiraju, "A survey of text classification with transformers: How wide? How large? How long? How accurate? How expensive? How safe?" *IEEE Access*, vol. 12, pp. 6518–6531, 2024.
- [20] M. A. K. Raiaan, M. S. H. Mukta, K. Fatema, N. M. Fahad, S. Sakib, M. M. J. Mim, J. Ahmad, M. E. Ali, and S. Azam, "A review on large language models: Architectures, applications, taxonomies, open issues and challenges," *IEEE Access*, vol. 12, pp. 26839–26874, 2024.
- [21] K. Singhal, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, Jul. 2023.
- [22] B. Zhang, Z. Liu, C. Cherry, and O. Firat, "When scaling meets LLM finetuning: The effect of data, model and finetuning method," in *Proc. 12th Int. Conf. Learn. Represent.*, Apr. 2024.
- [23] G. Chen, F. Liu, Z. Meng, and S. Liang, "Revisiting parameter-efficient tuning: Are we really there yet?" in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2022, pp. 2612–2626.
- [24] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learn. Represent.*, Jan. 2021.
- [25] N. Houlsby, A. Giurugu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for NLP," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2019, pp. 2790–2799.
- [26] J. Pfeiffer, A. Rücklé, C. Poth, A. Kamath, I. Vulić, S. Ruder, K. Cho, and I. Gurevych, "AdapterHub: A framework for adapting transformers," in *Proc. Conf. Empirical Methods Natural Lang. Process., Syst. Demonstrations*, 2020, pp. 46–54.
- [27] S. Chen, C. Ge, Z. Tong, J. Wang, Y. Song, J. Wang, and P. Luo, "AdaptFormer: Adapting vision transformers for scalable visual recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2022, pp. 16664–16678.
- [28] H. Le, J. Pino, C. Wang, J. Gu, D. Schwab, and L. Besacier, "Lightweight adapter tuning for multilingual speech translation," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process. (Short Papers)*, vol. 2, 2021, pp. 817–824.
- [29] H. Zhao, H. Tan, and H. Mei, "Tiny-attention adapter: Contexts are more important than the number of parameters," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2022, pp. 6626–6638.
- [30] T.-D. Le, C. Macabiau, K. Albert, S. Chatzinotas, P. Juvet, and R. Noumeir, "Transformer meets gated residual networks to enhance photoplethysmogram artifact detection informed by mutual information neural estimation," 2024, *arXiv:2405.16177*.
- [31] J. Kessler, "Scattertext: A browser-based tool for visualizing how corpora differ," in *Proc. ACL Syst. Demonstrations*, 2017, pp. 85–90.
- [32] R. Touchent, L. Romary, and E. de la Clergerie, "CamemBERT-bio: Leveraging continual pre-training for cost-effective models on French biomedical data," 2023, *arXiv:2306.15550*.
- [33] Y. Labrak, A. Bazoge, R. Dufour, M. Rouvier, E. Morin, B. Daille, and P.-A. Gourraud, "DrBERT: A robust pre-trained model in French for biomedical and clinical domains," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, Toronto, ON, Canada, 2023, pp. 16207–16221, doi: [10.18653/v1/2023.acl-long.896](https://doi.org/10.18653/v1/2023.acl-long.896).
- [34] A. Berhe, G. Draznieks, V. Martenot, V. Masdeu, L. Davy, and J.-D. Zucker, "AliBERT: A pre-trained language model for French biomedical text," in *Proc. 22nd Workshop Biomed. Natural Lang. Process. BioNLP Shared Tasks*, Toronto, ON, Canada, Jul. 2023, pp. 223–236, doi: [10.18653/v1/2023.bionlp-1.19](https://doi.org/10.18653/v1/2023.bionlp-1.19).
- [35] C. Macabiau, T.-D. Le, K. Albert, M. Shahriari, P. Juvet, and R. Noumeir, "Label propagation techniques for artifact detection in imbalanced classes using photoplethysmogram signals," *IEEE Access*, vol. 12, pp. 81221–81235, 2024.
- [36] N. Ding, Y. Qin, G. Yang, F. Wei, Z. Yang, Y. Su, S. Hu, Y. Chen, C.-M. Chan, W. Chen, J. Yi, W. Zhao, X. Wang, Z. Liu, H.-T. Zheng, J. Chen, Y. Liu, J. Tang, J. Li, and M. Sun, "Parameter-efficient fine-tuning of large-scale pre-trained language models," *Nature Mach. Intell.*, vol. 5, no. 3, pp. 220–235, Mar. 2023.
- [37] F. Pedregosa, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011.
- [38] F. Chollet, "Keras," Apr. 2015. [Online]. Available: <https://github.com/fchollet/keras>
- [39] M. Popel and O. Bojar, "Training tips for the transformer model," 2018, *arXiv:1804.00247*.
- [40] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014.
- [41] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, Mar. 2010, pp. 249–256.
- [42] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2015, pp. 448–456.
- [43] N. Björck, C. P. Gomes, B. Selman, and K. Q. Weinberger, "Understanding batch normalization," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2018, pp. 7694–7705.
- [44] C. Goutte and É. Gaussier, "A probabilistic interpretation of precision, recall and F-score, with implication for evaluation," in *Proc. Eur. Conf. Inf. Retr.*, Jan. 2005, pp. 345–359.
- [45] D. Biderman, J. Portes, J. J. G. Ortiz, M. Paul, P. Greengard, C. Jennings, D. King, S. Havens, V. Chiley, J. Frankle, C. Blakeney, and J. P. Cunningham, "LoRA learns less and forgets less," 2024, *arXiv:2405.09673*.
- [46] L. G. McCoy, A. K. Manrai, and A. Rodman, "Large language models and the degradation of the medical record," *New England J. Med.*, vol. 391, no. 17, pp. 1561–1564, Oct. 2024.

- [47] J. Liang, Y. Lin, Z. Guo, H. Liang, J. Ni, D. Lin, J. Qi, Z. Huang, W. Wang, and J. He, "Evaluation of LLMs accuracy and application in oncology principles and practice," Jan. 2025.
- [48] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," 2023, *arXiv:2312.00752*.
- [49] S. Wang, J. Asilis, Ö. F. Akgül, E. B. Bilgin, O. Liu, and W. Neiswanger, "Tina: Tiny reasoning models via LoRA," 2025, *arXiv:2504.15777*.



THANH-DUNG LE (Senior Member, IEEE) received the B.Eng. degree in mechatronics engineering from Can Tho University, Vietnam, the M.Eng. degree in electrical engineering from Jeju National University, South Korea, and the Ph.D. degree in electrical engineering (major in applied artificial intelligence) from École de Technologie Supérieure (ETS), University of Quebec, Canada. From October 2023 to May 2024, he was a Postdoctoral Fellow with the Biomedical Information Processing Laboratory, ETS. Before that, he joined Institut National de la Recherche Scientifique, Canada, where he researched classification theory and machine learning. He is currently a Research Associate with the Interdisciplinary Center for Security, Reliability, and Trust (SnT), University of Luxembourg, focusing on applied machine learning approaches for satellite communications systems. His research interest includes applied machine learning approaches for biomedical informatics problems. He received the Merit Ph.D. Scholarship from Le Fonds de Recherche du Québec Nature et Technologies. He also received the NSERC-PERSWADE Fellowship in Canada and the Graduate Scholarship from Korean National Research Foundation, South Korea.



TI TI NGUYEN (Member, IEEE) received the B.Eng. degree in electrical engineering from Ho Chi Minh City University of Technology, Vietnam, in 2013, the M.Eng. degree in embedded systems from the University of Rennes 1, France, in 2015, and the Ph.D. degree in telecommunications from Institut National de la Recherche Scientifique (INRS), University of Quebec, Canada, in 2020. Since 2020, he has been a Postdoctoral Fellow with the SynchroMedia, École de Technologie Supérieure, University of Quebec. His current research interests include mobile edge computing and radio resource management, 5G new radio, MIMO, wireless security, and AI for wireless communications.



VU NGUYEN HA (Senior Member, IEEE) received the B.Eng. degree (Hons.) from the French Training Program for Excellent Engineers in Vietnam, Ho Chi Minh City University of Technology, Vietnam, the Addendum degree from École Nationale Supérieure des Télécommunications de Bretagne-Groupe des École des Télécommunications, Bretagne, France, in 2007, and the Ph.D. degree (Hons.) from Institut National de la Recherche Scientifique-Énergie, Matériaux et Télécommunications, University of Quebec, Montreal, QC, Canada, in 2017. From 2016 to 2021, he was a Postdoctoral Fellow with the École Polytechnique de Montréal, and then the Resilient Machine Learning Institute, École de Technologie Supérieure, University of Quebec. He is currently a Research Scientist with the Interdisciplinary Centre for Security, Reliability, and Trust, University of Luxembourg. He was a recipient of the FRQNT Postdoctoral Fellowship for International Researcher (PBEEE) awarded by the Québec Ministry of Education, Canada, in 2018 and 2019. In 2021 and 2022, he was also awarded the Certificate for Exemplary Reviews by IEEE WIRELESS COMMUNICATIONS LETTERS.



SYMEON CHATZINOTAS (Fellow, IEEE) received the M.Eng. degree in telecommunications from the Aristotle University of Thessaloniki, Greece, in 2003, and the M.Sc. and Ph.D. degrees in electronic engineering from the University of Surrey, U.K., in 2006 and 2009, respectively. He is currently a Full Professor/the Chief Scientist I and the Head of the Research Group SIGCOM, Interdisciplinary Centre for Security, Reliability, and Trust, University of Luxembourg. In parallel, he is an Adjunct Professor with the Department of Electronic Systems, Norwegian University of Science and Technology, and a Collaborating Scholar with the Institute of Informatics and Telecommunications, National Center for Scientific Research "Demokritos." In the past, he has lectured as a Visiting Professor with the University of Parma, Italy, and contributed in numerous research and development projects for the Institute of Telematics and Informatics, Center of Research and Technology Hellas, and the Mobile Communications Research Group, Center of Communication Systems Research, University of Surrey. He has authored more than 700 technical papers in refereed international journals, conferences, and scientific books, and has received numerous awards and recognitions, including the IEEE Fellowship and the IEEE Distinguished Contributions Award. He is on the Editorial Board of IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE OPEN JOURNAL OF VEHICULAR TECHNOLOGY, and *International Journal of Satellite Communications and Networking*.



PHILIPPE JOUVET received the M.D. degree in pediatrics and the M.D. degree in intensive care from Paris V University, Paris, France, in 1989 and 1990, respectively, and the Ph.D. degree in pathophysiology of human nutrition and metabolism from Paris VII University, Paris, in 2001. He joined the Pediatric Intensive Care Unit, Sainte Justine Hospital, University of Montreal, Montreal, QC, Canada, in 2004. He is currently the Deputy Director of the Research Center and the Scientific Director of the Health Technology Assessment Unit, CHU Sainte-Justine Hospital, University of Montreal. He conducts a research program on computerized decision support systems for health providers. His research program is supported by several grants from CHU Sainte-Justine Hospital, Quebec Ministry of Health, FRQS, the Canadian Institutes of Health Research (CIHR), and the Natural Sciences and Engineering Research Council (NSERC). He has published more than 160 articles in peer-reviewed journals. He gave more than 120 lectures at national and international congresses. He has a salary award for research from Quebec Public Research Agency (FRQS).



RITA NOUMEIR (Member, IEEE) received the master's and Ph.D. degrees in biomedical engineering from École Polytechnique of Montreal. She is currently a Full Professor with the Department of Electrical Engineering, École de Technologie Supérieure, University of Quebec, Canada. She is a Research Chair in artificial intelligence (AI) in healthcare/digital health and life sciences from Fonds de Recherche du Québec-Santé (FRQS). Concurrently, she is the Co-Director of the Research Cluster on AI Applied to Acute Child Care, FRQS. She has extensively worked in healthcare information technology and image processing. She has also provided consulting services in large-scale software architecture, healthcare interoperability, workflow analysis, and technology assessment for several international software and medical companies, including Canada Health Infoway. Her main research interest includes applying AI methods to create decision support systems.

...