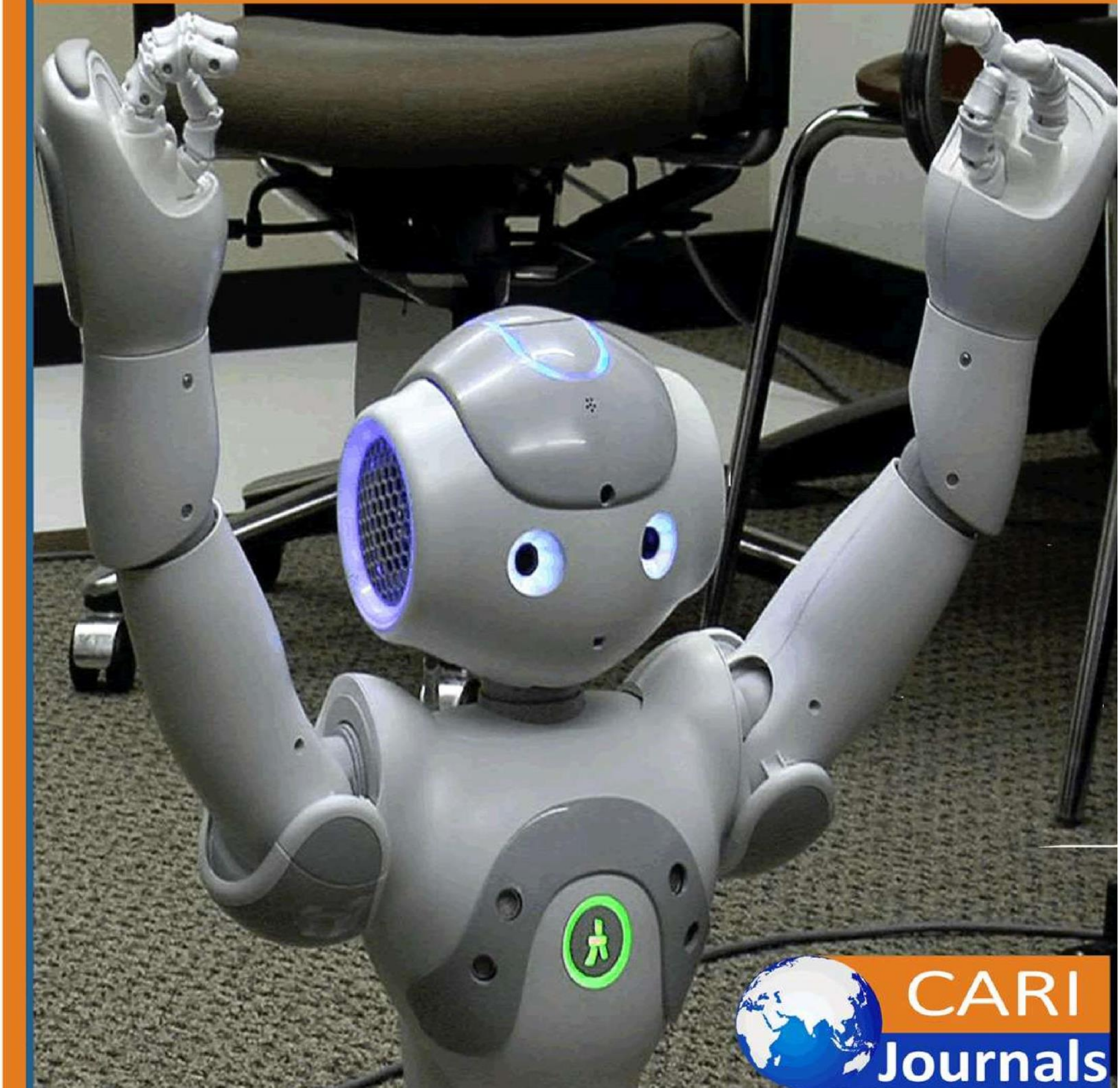
International Journal of Computing and Engineering (IJCE)

Improving Medical Abstract Classification Using PEFT-LoRA Fine-Tuned Large and Small Language Models



CARI Journals

International Journal of Computing and Engineering
ISSN 2958-7425 (online)
Vol. 6, Issue No. 6, pp. 68 - 75, 2024 www.carijournals.org

Improving Medical Abstract Classification Using PEFT-LoRA Fine-Tuned Large and Small Language Models

[1*]Dr. Rahul Kavi, [2]Jeevan Anne

[1,2]Independent Researcher

https://orcid.org/0000-0001-7137-9108

Accepted: 30th Oct, 2024, Received in Revised Form: 14th Nov, 2024, Published: 23rd Nov, 2024

Abstract

Designing intelligent systems to classify text in the medical domain is a challenging task. There is a shortage of openly available medical datasets (due to HIPPA-related strict regulations on protected health information for patients). In this paper, we explore the application of Open Source Medical LLMs (such as Meditron LLM), generic Large Language Models (such as LLAMA2), and Small Language Models (such as Phi2) on medical text classification (medical abstract dataset). We show that PEFT approaches such as LoRA can perform very well in classifying medical text, which involves interpreting patient conditions and symptoms and determining what medical problems the patients have. These approaches (based on Large and Small Language Models) have outperformed the current state of the results on medical abstracts corpus. In addition to medical LLMs, the open-source generic LLMs can be adapted to solving classification tasks on medical text and perform nearly as well as the specialized medical LLMs. SLMs can be a serious contender for solving domain-specific classification tasks (e.g., medical literature). This shows that carefully selecting the training data and fine-tuning positively impacts classification accuracy, precision, and recall. Generic Language Models such as LLAMA2 (LLM) and Phi2 (SLM) weren't specifically trained with medical text. Medical LLMs such as Meditron outperform LLAMA2 and Phi2 in precision and accuracy. This is quite evident as Meditron was originally trained on medical text. The (micro averaged) F1 score for the fine-tuned Meditron model is 0.64. This is superior compared to fined-tuned LLAMA2 of 0.58 and Phi2 of 0.62. We see that Phi2 can outperform LLAMA2 with fewer number of parameters. The approaches used in this work can be extended to other medical text classification problems in the medical domain.

Keywords: *Machine Learning, Natural Language Processing, Medical Text Analysis, Language Models.*

1. Introduction

Supervised learning approaches usually outperform unsupervised learning approaches when labeled data is available. In the medical domain, there is an immense curiosity to explore the usage of LLMs to assist medical professionals (like every other domain). By leveraging large datasets such as PubMed (containing citations and abstracts of Bio-Medical papers) [20] and Medical Guidelines [21], LLMs have been built to solve problems in the medical research community [22] [23]. One such recent and popular LLM is Meditron [24]. It is freely available on HuggingFace Hub for download and use. The Meditron LLM is based on a continued pre-trained LLAMA2 model (on large medical text datasets). After this pre-training process, the LLM has shown incredible performance in answering queries related to the medical domain (e.g., MedQA [26], MedMCQA [27], etc.). This model has outperformed LLAMA-like models on such benchmarks. The LLAMA2 [25] model was trained on large publicly available datasets and is designed to be a generic model (which can understand natural language). Microsoft Phi2 [14] is another popular Small Language Model (SLM) model that was trained on textbook-like data (not specifically medical text). This study uses the Meditron 7B, LLAMA2 7B, and Phi2 model with LoRA (PEFT) finetuning. PEFT is an increasingly popular approach for finetuning language models for downstream tasks. LoRA [16] is one of the well-known PEFT approaches that is widely used. This approach works by learning partial weights (instead of updating all the language model weights). This approach is practical in adapting a generic purpose language model to different (but related tasks).

We have chosen one specific medical LLM, a generic LLM, and an SLM for this study to compare performance on medical datasets. This ensures that the problem of medical abstract classification is tackled with various available language models (with different computing requirements). This work explored the Microsoft Phi2 model as an alternative, as Meditron and LLAMA2 have high computing requirements. The Phi2 model was trained on textbook data and performed much better on multi-step reasoning tasks than the LLAMA2 model. This was achieved with significantly fewer parameters (2.7 billion) than LLAMA2. The Phi2 model has a high potential to be fine-tuned on specialized tasks such as medical text classification. The LLAMA2 model trained on 7B parameters was explored in this work. This model was groundbreaking work when it was released. This was completed with GPT3.5 LLM in MMLU tasks. However, this model is large compared to Phi2. The Meditron LLM was based on continued pre-training of the LLAMA2 model (on medical text datasets). The Meditron LLM is generally considered superior in the medical domain compared to generic language models. We compare the results of this study with other unsupervised approaches like zero-shot and similarity-based approaches [18]. The dataset used in this study has 14438 samples under different classes of diseases/medical conditions. These are categorized as *Neoplasms*, *Digestive system diseases*, *Nervous system diseases*, *cardiovascular diseases,* and *General pathological diseases*. The dataset is available with a train and test split of ratio 0.79 to 0.21. Zero-shot approaches based on DistilBERT, BART, and DeBERTa have shown the performance of F1 scores (micro-averaged): 0.25, 0.56 and 0.57 respectively [18]. Our LoRa fine-tuned language models LLAMA2, Phi2, and Meditron perform 0.61, 0.62, and 0.67, respectively. This demonstrates a better approach to solving the problem of medical abstract classification.

2. Related Work

There is extensive research in the application of machine learning in the medical domain. However, data availability is hindered due to government regulation of the health industry and privacy concerns protecting medical professionals and patients. The PubMed database comprises extensive bio-medical literature, life science journals, and online books. This is useful to NLP train models as this text is available at PubMed Central (provided by the U.S. National Institute of Health's National Library of Medicine). The Meditron model was pre-trained on the PubMed dataset (selected abstracts, medical guidelines, etc.). This helps the Meditron model to identify and understand medical context and terminology. In the medical literature, there are popular approaches based on BERT [1][2][3][4][5]. These approaches have proven more popular than traditional approaches such as TF-IDF. The features from pre-trained BERT are obtained and passed to a classification algorithm to predict input clinical text [6]. Features obtained from a pre-trained BERT capture context and meaning within a sentence (unliked TF-IDF, which treats each word differently and focuses on word frequency) [7][8][9][10][11]. The Phi2 model released by Microsoft Research has been used in several settings for efficient feature extraction [12][13][14]. These features can be used as embeddings, and classification can be performed on them. Parameter Efficient Fine-Tuning approaches such as LoRA are well-researched techniques to fine-tune large generic models to solve specific language generation/classification tasks. This results in relatively easy fine-tuning (with less compute and memory requirements) for downstream tasks [15][16][17]. BERT and Language Models extract embeddings from text that capture context and meaning. However, LLMs cluster similar words better than classical models such as BERT [19].

For this study, we don't explore traditional approaches such as TF-IDF and focus solely on language model-based strategies. This study compares a small language model (Phi-2), a large language model (LLAMA2), and a medical text pre-trained language model (Meditron) on medical text classification.

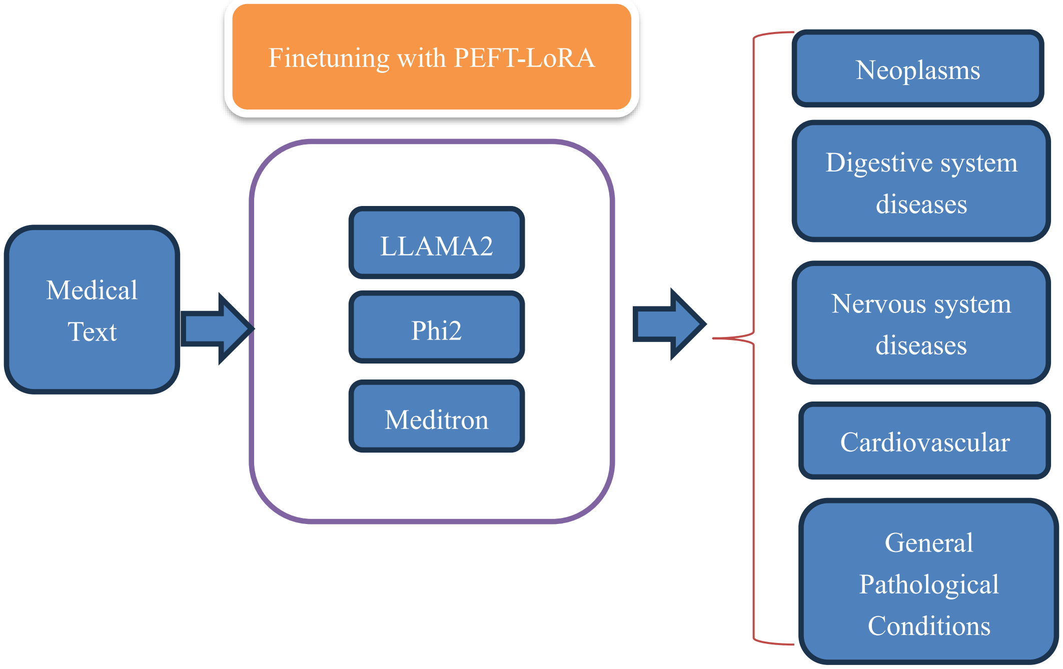

Finetuning with PEFT-LoRA
Medical Text
LLAMA2
Phi2
Meditron
Neoplasms
Digestive system diseases
Nervous system diseases
Cardiovascular
General Pathological Conditions

Figure 1: Experiment setup

3. Methodology

In this section, we describe our experimental setup. In Figure 1, we see how the experimental setup is configured. This diagram also shows how training/inference is run. We obtain the medical text from [18]. The dataset comprises five classes and is divided into train and test sets. This dataset consists of text describing the medical abstract of a patient's condition, and the label is that of the disease that this text is classified into. There are five kinds of classes in this dataset. These are Neoplasms, Digestive system diseases, Nervous system diseases, Cardiovascular diseases, and General Pathological Conditions. The dataset with the class distribution is described in Table 1. We can see from Table 1 that all classes aren’t distributed equally (based on frequency). For this reason, we employed weighted cross entropy to train our classifier (for Phi2, LLAMA2, and Meditron). The frequency of the class count determines the weights of the disease class. This ensures that there is no bias towards a particular class by the classifier. We use Stochastic Gradient Descent with a 0.0004 learning rate, batch size of 4, a momentum of 0.9, and train for three epochs. We trained Meditron for one epoch only as it has been pre-trained on medical text (and it needs no extra training). A weight decay of 0.01 is also added. This regularization effect prevents overfitting and improves the model's generalization. The model was trained on an A100 GPU (40GB). The experimental setup used HuggingFace APIs to load pre-trained Language Models such as Meditron, LLAMA2, and Phi2. The training took roughly 1 hour each. Smaller models, such as Phi2, took significantly less time. This shows that smaller LLMs can be trained much faster with PEFT-LORA (as compared with larger models). The HuggingFace APIs also provide an easy-to-use interface to train a given LLM with the PEFT-LoRA approach.

Table 1: Data distribution among classes

Disease or Class	Train Set	Test Set
Neoplasm	2530	633
Digestive system diseases	1195	299
Nervous system disease	1540	385
Cardiovascular diseases	2441	610
General Pathological Conditions	3844	961
Total	**11550**	**2888**

4. Results

In this section, the results for these experiments are described. The results for the training and inference are shown in Table 2. We can see from Table 2 that the Microsoft Phi2 model with fewer parameters performs better than LLAMA2. However, the Meditron model outperforms LLAMA2 and Phi2 as it was pre-trained on medical text from PubMed. This gives Meditron a significant advantage over other models in the medical domain. The Meditron LLM tokenizer is customized for medical text (that it was trained with). This results in much better performance compared to LLAMA2 (which was largely trained on internet data) and Phi2 (largely trained on textbook data). Meditron has significantly better precision, recall scores (we note these scores very sensitive in medical literature).

Table 2: Precision, Recall and F1 scores on medical abstracts datasets

Model	F1-Score	Precision	Recall
Meditron	0.67	0.67	0.67
LLAMA2	0.61	0.61	0.61
Phi2	0.62	0.62	0.62

5. Conclusions

PEFT-LoRA approaches are instrumental in customizing (fine-tuning) generic and medical LLMs to various datasets. Meditron-7B model is a strong LLM for medical abstract classification tasks. It has much better precision and recall scores than LLAMA2 and Phi2. This approach can also be extended to other medical classification tasks (including Named Entity Recognition). Other models like Microsoft Phi2 have shown better than expected performance with fewer trainable parameters. However, Meditron-7B outperforms it in every aspect as the tokenizer is customized to medical text, and the model is much better at handling medical text than others.

6. Recommendation

Working in sensitive areas such as medical text classification, Meditron-7 B-like models are much more effective than models (LLAMA2) trained on generic internet data. Such models should be used with applications demanding high precision and recall scores.

References

[1] Gasmi, K. (2022, September). Improving bert-based model for medical text classification with an optimization algorithm. In *International Conference on Computational Collective Intelligence* (pp. 101-111). Cham: Springer International Publishing.

[2] Khadhraoui, M., Bellaaj, H., Ammar, M. B., Hamam, H., & Jmaiel, M. (2022). Survey of BERT-base models for scientific text classification: COVID-19 case study. *Applied Sciences*, *12*(6), 2891.

[3] Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., ... & Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, *3*(1), 1-23.

[4] Mascio, A., Kraljevic, Z., Bean, D., Dobson, R., Stewart, R., Bendayan, R., & Roberts, A. (2020). Comparative analysis of text classification approaches in electronic health records. *arXiv preprint arXiv:2005.06624*.

[5] Qasim, R., Bangyal, W. H., Alqarni, M. A., & Ali Almazroi, A. (2022). A Fine-Tuned BERT-Based Transfer Learning Approach for Text Classification. *Journal of healthcare engineering*, *2022*(1), 3498123.

[6] Lenivtceva, I., Slasten, E., Kashina, M., & Kopanitsa, G. (2020). Applicability of machine learning methods to multi-label medical text classification. In *Computational Science–ICCS 2020:*

20th International Conference, Amsterdam, The Netherlands, June 3–5, 2020, Proceedings, Part IV 20 (pp. 509-522). Springer International Publishing.

[7] Gema, A. P., Minervini, P., Daines, L., Hope, T., & Alex, B. (2023). Parameter-efficient fine-tuning of llama for the clinical domain. *arXiv preprint arXiv:2307.03042.*

[8] Xu, L., Xie, H., Qin, S. Z. J., Tao, X., & Wang, F. L. (2023). Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. arXiv preprint arXiv:2312.12148.

[9] Hu, Z., Wang, L., Lan, Y., Xu, W., Lim, E. P., Bing, L., ... & Lee, R. K. W. (2023). Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models. *arXiv preprint arXiv:2304.01933*.

[10] Clarke, C., Heng, Y., Tang, L., & Mars, J. (2024). PEFT-U: Parameter-Efficient Fine-Tuning for User Personalization. *arXiv preprint arXiv:2407.18078*.

[11] Wu, C., Lin, W., Zhang, X., Zhang, Y., Xie, W., & Wang, Y. (2024). PMC-LLaMA: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, ocae045.

[12] Zhu, Y., Zhu, M., Liu, N., Ou, Z., Mou, X., & Tang, J. (2024). LLaVA-$\phi $: Efficient Multi-Modal Assistant with Small Language Model. *arXiv preprint arXiv:2401.02330*.

[13] Valade, F. (2024). Accelerating Large Language Model Inference with Self-Supervised Early Exits. *arXiv preprint arXiv:2407.21082*.

[14] Microsoft Research. Phi-2: The surprising power of small language models. Microsoft Research Blog, December 2023.

[15] Pu, G., Jain, A., Yin, J., & Kaplan, R. (2023). Empirical analysis of the strengths and weaknesses of PEFT techniques for LLMs. *arXiv preprint arXiv:2304.14999.*

[16] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685.*

[17] Han, Z., Gao, C., Liu, J., & Zhang, S. Q. (2024). Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608.*

[18] Schopf, T., Braun, D., & Matthes, F. (2022, December). Evaluating unsupervised text classification: zero-shot and similarity-based approaches. *In Proceedings of the 2022 6th International Conference on Natural Language Processing and Information Retrieval (pp. 6-15).*

[19] Freestone, M., & Santu, S. K. K. (2024). Word Embeddings Revisited: Do LLMs Offer Something New?. *arXiv preprint arXiv:2402.11094.*

[20]National Library of Medicine. PubMed. ***https://pubmed.ncbi.nlm.nih.gov/.*** *Accessed Octobe r 8, 2024.*

[21] World Health Organization. Clinical guidelines. *https://www.who.int/guidelines. Accessed October 8, 2024.*

[22] García-Ferrero, I., Agerri, R., Salazar, A. A., Cabrio, E., de la Iglesia, I., Lavelli, A., ... & Zaninello, A. (2024). Medical mT5: an open-source multilingual text-to-text LLM for the medical domain. arXiv preprint arXiv:2404.07613.

[23] Song, Y., Zhang, J., Tian, Z., Yang, Y., Huang, M., & Li, D. (2024). LLM-based privacy data augmentation guided by knowledge distillation with a distribution tutor for medical text classification. *arXiv preprint arXiv:2402.16515*.

[24] Chen, Z., Cano, A. H., Romanou, A., Bonnet, A., Matoba, K., Salvi, F., ... & Bosselut, A. (2023). Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

[25] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

[26] Jin, D., Pan, E., Oufattole, N., Weng, W. H., Fang, H., & Szolovits, P. (2021). What disease does this patient have? a large-scale open domain question answering dataset from medical exams. Applied Sciences, 11(14), 6421.

[27] Pal, A., Umapathi, L. K., & Sankarasubbu, M. (2022, April). Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning* (pp. 248-260). PMLR.