Enhancing Predictive Accuracy and Resource Utilization in Disease Prediction: A Comparative Study of Small Versus Large Language Models

V Karthick, Associate Professor
Department of Computer Science and Engineering,
Rajalakshmi Engineering College,
Chennai - 602105.
vkarthick86@gmail.com

Tamanna
Department of Computer Science and Engineering,
Rajalakshmi Engineering College,
Chennai - 602105.
210701281@rajalakshmi.edu.in

Sudhir S
Department of Computer Science and Engineering,
Rajalakshmi Engineering College,
Chennai - 602105.
210701269@rajalakshmi.edu.in

ABSTRACT

This project involves a comparative analysis of small and large language models (SLMs and LLMs) to evaluate their capabilities in deciphering complex relationships between diseases and symptoms using distilgpt2 (SLM) and gpt2 (LLM) models. Utilizing a large dataset, the models were studied for computational efficiency, predictive performance, and resource requirements. The methodology includes loading datasets, tokenization, model setup, monitoring training and validation losses, hyperparameter tuning, and generating text. The study showed SLM's proficiency in producing context-aware responses, while LLM's strength lies in generating refined and comprehensive text. Large language models demonstrated higher predictive accuracy but required significantly more computational resources, making them less suitable for resource-constrained environments. Conversely, small language models, despite their lower accuracy, are more efficient in resource usage, making them suitable where computational resources are limited. This study highlights the importance of domain-specific training in enhancing the predictive accuracy and resource utilization of language models in healthcare informatics.

Keywords - SLMs, LLMs, BERT, Tokenizer, DistilBERT, resource-constrained, disease prediction

INTRODUCTION

The field of disease prediction using small and large language models (SLMs and LLMs) is rapidly advancing in healthcare. SLMs, with their limited parameters and smaller datasets, excel in speed and efficiency, making them suitable for real-time applications. On the other hand, LLMs, with their extensive parameters and training on large datasets, achieve high accuracy in complex language tasks. In this project, we focus on predicting diseases based on reported symptoms using a dataset from Hugging Face and comparing outcomes between SLMs and LLMs. The goal is to map symptoms to potential diseases effectively.

The methodology involved in the project includes feeding the language models with symptom data and analyzing the outputs to determine the most likely disease outcomes.The goal is to make most out of the SLMs and LLMs to create a resilient system that can help healthcare providers in the initial diagnosis of diseases, thus, the diagnostic process is going to be made easier. This study helps us gain deep insights on practicality of introducing such language models in the domain of healthcare considering factors like accuracy,computational speed and efficiency of prediction. The outcomes of the research will be instrumental in determining the most effective way to implement language models for medical help, which will, in turn, help in the betterment of patient care and resource management in the healthcare sector.

RESEARCH GAP

- Absence of comparative analysis of small and large language models (SLMs and LLMs) in disease prediction.
- Focus on predictive accuracy and resource utilization in the comparison.
- Existing studies examine individual capabilities of SLMs and LLMs.
- This research aims to bridge the gap by providing a direct comparison.
- Comparison conducted within the stringent constraints and precise requirements of healthcare informatics.

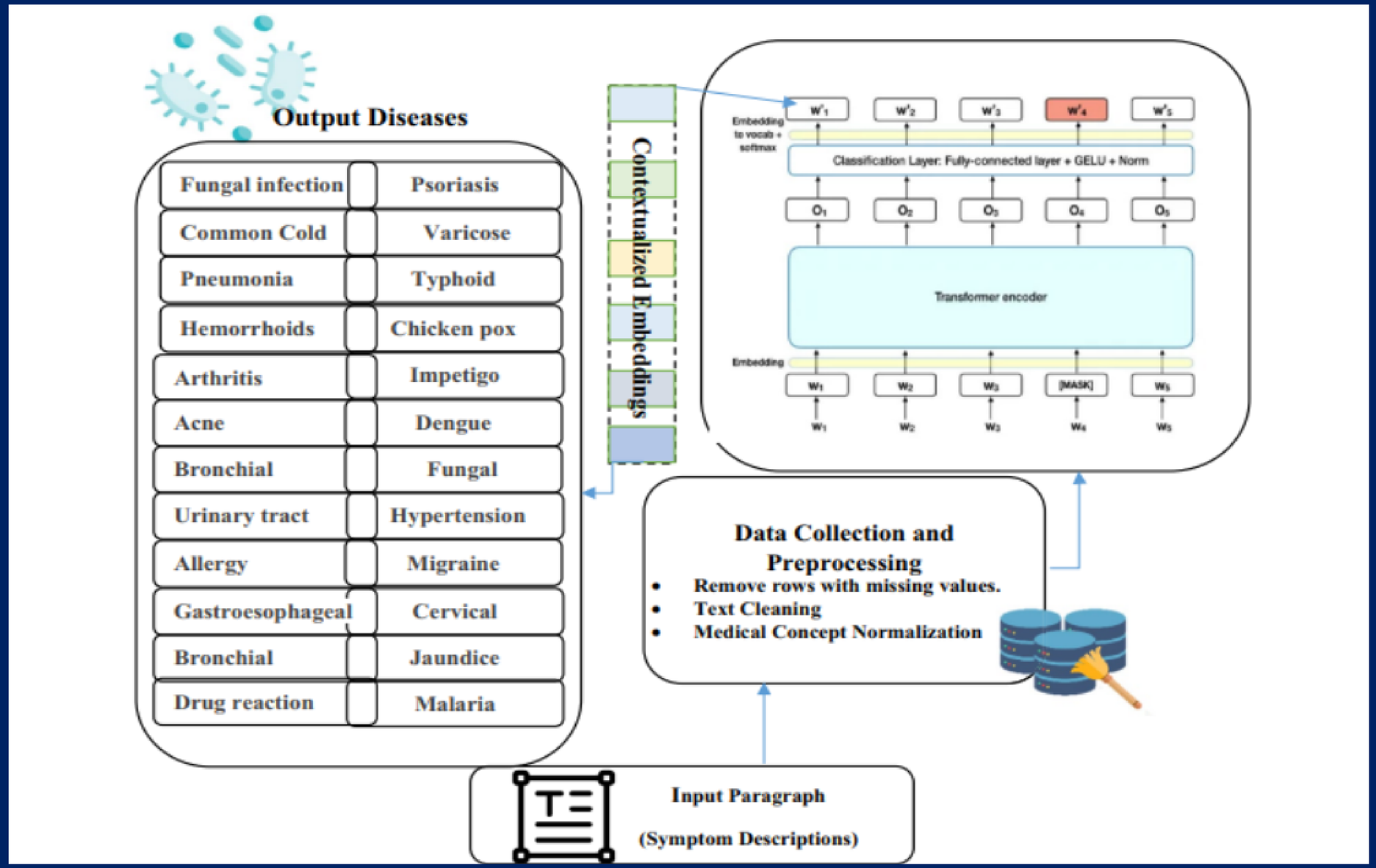
PROPOSED METHOD

The proposed work aims to perform a comparative research on small language models (SLMs) and large language models (LLMs) in a setting of the disease prediction.This work is aimed at making an achievement in the existing knowledge[14] about disease prediction with a comparative examination of small language models (SLMs) and large language models (LLMs). The aim is to assess and compare the forecasting ability, workload, and resource deployment of both types of models.

The methodology of this study is designed to provide a comprehensive comparison of Small Language Models (SLMs) and Large Language Models (LLMs) in the context of disease prediction. It aims to evaluate and contrast the predictive performance, computational efficiency, and resource utilization of both types of models.The steps involved include:

- Data Preprocessing: Like the present one, the data is prepared for processing by standardizing the symptoms in the form of a single string separated by commas.
- Dataset Preparation: The processed data is then converted in a form that can be used for model training. This is tokenization of data input and the creation of the data loaders for training.
- Model Initialization: The first thing that is done with evolved pre-trained LMs and evolving SLMs is the initialization. For SLMs, for instance, models like DistilGPT-2 can be toyed with, and for LLMs, GPT-2 or BERT can be used. The models are initialized with pre-weighted parameters and relocating these weights to the required device (CPU or GPU).
- Model Training: The layers are trained by using a regular training loop. During every epoch, the parameters of the model are adapted to minimize the loss value. Perturbation delivered through the forward method of the model provides the loss value to the output in the loss attribute.
- Model Evaluation: After training, model accuracy is tested on the validation set considered. The way of the performance of the network is tested as the loss function is used which shows the difference between the predicted values by model and actual data. The metrics achieved in the present model (precision, accuracy, recall, and the F1-score) can be calculated in order to conduct a thorough assessment.

ARCHITECTURE DIAGRAM



LITERATURE REVIEW

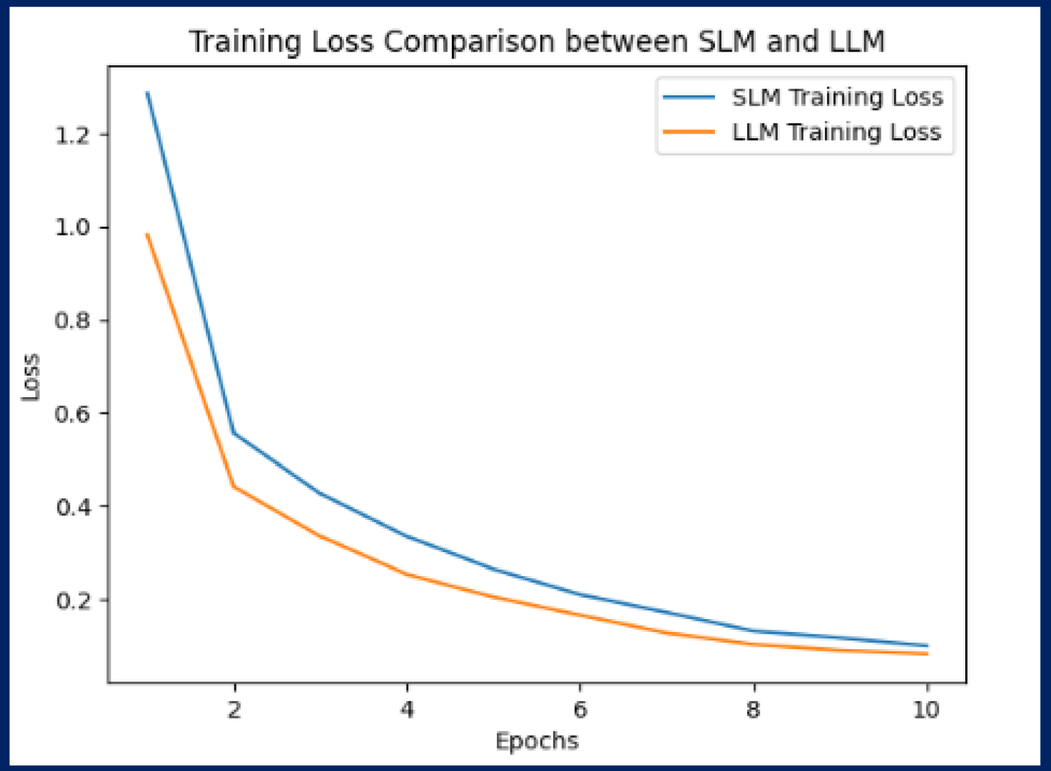
PAPER	AUTHORS	DATE PUBLISHED	REVIEW	LIMITATIONS
DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter	Victor Sanh, Lysandre Debut, Julien Chaumond, Thomas Wolf	October, 2020	The paper discusses the prevalent use of large-scale pre-trained models in Natural Language Processing (NLP). It highlights the challenges of operating these models under constrained computational training or inference budgets1. The paper also reviews the use of distillation for building task-specific models	DistilBERT's disadvantages include potential negative impact from random weight initialization, large training data requirement, slower inference times, and a 3% capability loss.
What Happens When Small Is Made Smaller? Exploring the Impact of Compression on Small Data Pretrained Language Models	Busayo Awobade, Mardiyyah Oduwole, Steven Kolawole	6 April 2024	The paper reviews the role of compression techniques like pruning, knowledge distillation, and quantization in advancing machine learning, particularly in the context of low-resource language models1.	The study on compressing small-data language models reveals trade-offs between model size reduction and performance, resource constraints, fine-tuning challenges, task-specific impact, and evaluation metric considerations.

PAPER	AUTHORS	DATE PUBLISHED	REVIEW	LIMITATIONS
Knowledge-Augmented Reasoning Distillation for Small Language Models in Knowledge-Intensive Tasks	Minki Kang, Seanie Lee, Jinheon Baek, Kenji Kawaguchi, Sung Ju Hwang	28 May 2023	The paper, "Knowledge-Augmented Reasoning Distillation for Small Language Models in Knowledge-Intensive Tasks"[1], reviews the challenges of deploying Large Language Models (LLMs) due to their high computational requirements and data privacy concerns[1].	The study proposes Knowledge-Augmented Reasoning Distillation (KARD) to enhance small language models (LMs) for knowledge-intensive reasoning tasks. However, challenges remain in fine-tuning, resource constraints, and task-specific impact.
Can Small Language Models be Good Reasoners for Sequential Recommendation?	Yuling Wang, Changxin Tian, Binbin Hu, Yanhua Yu, Ziqi Liu, Zhiqiang Zhang, Jun Zhou, Liang Pang, Xiao Wang	7 March 2024	The paper reviews the challenges of deploying Large Language Models (LLMs) in sequential recommendations due to their high computational requirements and data privacy concerns[1]. It discusses the limitations of small LMs in recommendation tasks[1].	The study proposes a novel Step-by-step Knowledge Distillation Framework for recommendation (SLIM) to enable sequential recommenders to leverage the exceptional reasoning capabilities of large language models (LLMs) in a resource-efficient manner. However, challenges remain in user behavior complexity, resource requirements, and utilization of natural language rationales.

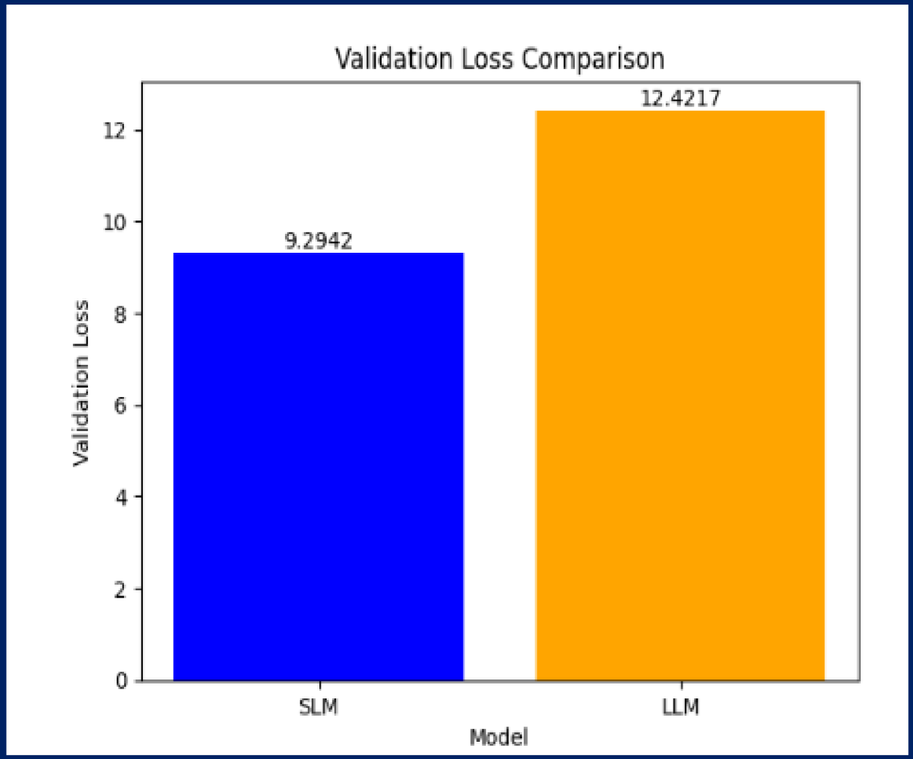
PAPER	AUTHORS	DATE PUBLISHED	REVIEW	LIMITATIONS
PMC-LLaMA: toward building open-source language models for medicine.	Wu C Lin W Zhang X Zhang Y Xie W Wang Y	13 April 2024	This paper focuses on developing open-source language models in the medical application area.It includes text collection from PMC articles, language model's calibration, and implementing specific optimization strategies.	The study describes the creation of PMC-LLaMA, an open-source language model tailored for medical applications. While effective, challenges remain in domain-specific fine-tuning and resource constraints
What Happens When Small Is Made Smaller? Exploring the Impact of Compression on Small Data Pretrained Language Models	Busayo Awobade, Mardiyyah Oduwole1, Steven Kolawole	06 Apr 2024	This research paper analyzes how well pruning,knowledge distillation,and quantization work on small language models that are low on resources.This study provides that compression techniques eventually improve SLMs efficiency.	The study investigates compression techniques on the low-resourced language model AfriBERTa. While effective, challenges include generalization trade-offs, resource constraints, fine-tuning difficulties, task-specific impact, and reliance on accuracy metrics

RESULT AND COMPARATIVE ANALYSIS

The results highlight the trade-offs between SLMs and LLMs. While LLMs may provide better performance on the training data, SLMs can often generalize better to unseen data and provide faster inference times. This makes SLMs a viable option for applications where computational resources are limited or a fast response is required.

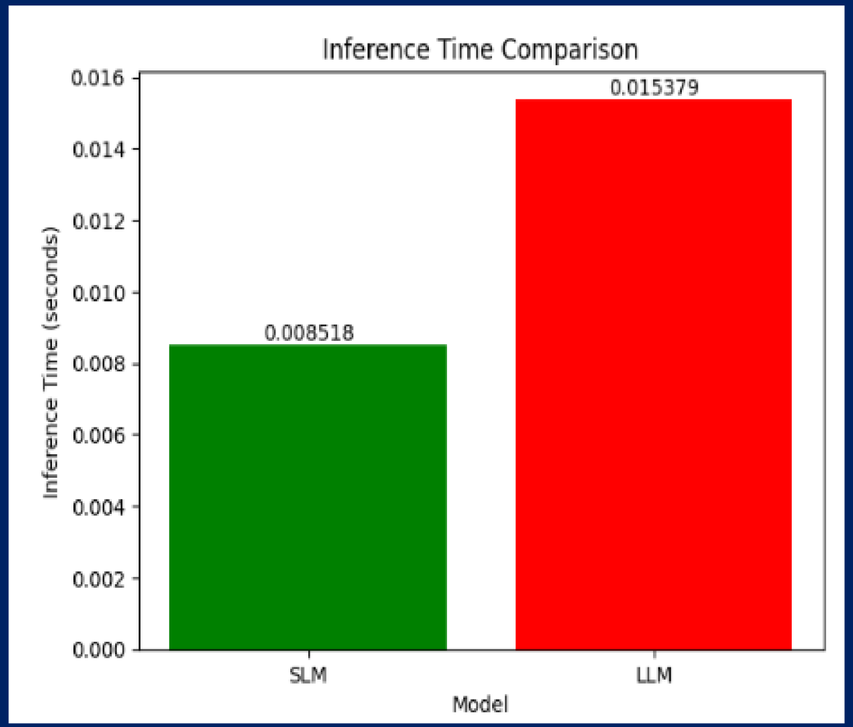


The SLM has a higher training loss difference than the LLM. It could be that the SLMs simplifying of the architecture and the few parameters involved does not enable as close the fitting of the training data as can be done by the LLM.



In comparison, the validation loss of the SLM is less than that of the LLM. This demonstrates that the SLM has good generalizability to the data that the neural network has not encountered despite its higher training loss. This could be in connection with the discriminatory power of SLMs, where by avoiding data overfitting, they evade the overfitting problem of larger models such as the LLMs.

The SLM speeds up the inference process relative to the LLM. This was to be anticipated as the SLMs are usually more computationally efficient owing to their small size/dimensionality and lesser number of parameters.



DISCUSSION

Disease prediction is a critical field that requires comprehensive comparisons between Small Language Models (SLMs) and Large Language Models (LLMs), as this study provides.The research shows the interplay between SLMs and LLMs. Actually, despite possible superiority of LLMs on training data, the corresponding feature of SLMs generalizing better to the unseen data and offering faster inference time remains[17]. This implies too that the outcome of the study of [1] is the reason we agree with the effectiveness of SLMs in particular DistilBERT. Yet, according to [2], BERT or other LLMs were shown to have a major impact and did better than previous models in some areas such as disease predictions. Therefore, those models can be promising in this field. This choice of SLMs or LLMs suits the particular necessities of the application, including the effect on the computation resources, the necessity of query-time responses, as well as the importance of accuracy.

CONCLUSION

Therefore, as the study shows both SLMs and LLMs are effective at disease prediction but there are clear allusions for improvements that guide future researchers to study behaviors of these language models. This result shows both the options between the SLMs and LLMs and also gamuts within each mode of learning. Although LLM may be unable to improve the notion, the SLM can manage to generalize well on new data and provide fast computing speed. Therefore, SLMs provide a reliable approach in situations where the computation resources are limited or response time is fast. Nonetheless, the preference between SLMs and LLMs depends on the operational needs that range from computational resources availability to need for real time responses and predictive accuracy. Contributing to the area of forecasting disease is an important outcome of the study by elaborating the difference between the two approaches: SLMs and LLMs, furthering research, and informing the building of systems that are better and more accurate for predicting disease.

FUTURE WORK

The present study explores multiple lines of research in the near future. Another area of possible research is to identify approaches that would aid in increasing the efficiency of SLMs This could be done through approaches for model compression, quantization or hardware optimization as an example. A possibility can be to research how SLMs can be deployed in other healthcare applications, such as patient triage as well as getting medical information back. In addition, considerations of ethical and privacy issues should be included in future perspectives on SLMs for health care. It could mean looking into techniques of safeguarding personal data, obtaining written consent, and preventing the models from misusing. Finally, using a trained SLM in real life situations like a hospital or a health application and measuring their actual performance compared to others could be another exciting path for future research. This would offer innovative information about the realization and reliability of using SLMs in disease forecasting.

REFERENCES

[1] Sanh, V. et al. (2020) Distilbert, a distilled version of Bert: Smaller, faster, cheaper and lighter, arXiv.org.
[2] Devlin, J. et al. (2019) Bert: Pre-training of deep bidirectional Transformers for language understanding, arXiv.org.
[3] Schick, T. and Schütze, H. (2021) It's not just size that matters: Small language models are also few-shot learners, arXiv.org.
[4] Aligning Large and Small Language Models via Chain-of-Thought Reasoning (Ranaldi & Freitas, EACL 2024)
[5] Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Adv Neural Inf Process Syst. 2020; 33: 1877–901.
[**6**] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Process Syst. 2017; 30: 1–11.
[7] Yang R, Tan TF, Lu W, Thirunavukarasu AJ, Ting DSW, Liu N. Large language models in health care: development, applications, and challenges. Health Care Sci. 2023; 2: 255–263.
[8] Zhang, P. et al. (2024) TinyLlama: An open-source small language model, arXiv.org.
[9] Hu, S. et al. (2024) MiniCPM: Unveiling the potential of small language models with Scalable Training Strategies, arXiv.org.
[10] Awobade, B., Oduwole, M. and Kolawole, S. (2024) What happens when small is made smaller? exploring the impact of compression on small data pretrained language models, arXiv.org.

REFERENCES

[11] “Utilizing Natural Language Processing and Large Language Models in the Diagnosis and Prediction of Infectious Diseases: A Systematic Review.” 2024. American Journal of Infection Control, April.
[12] Hassan, E., Abd El-Hafeez, T. & Shams, M.Y. Optimizing classification of diseases through language model analysis of symptoms. Sci Rep 14, 1507 (2024).
[13] StoryBuddiesPlay. 2024. Microsoft Phi 3 LLM: Powerful Small Language Model Demystified. StoryBuddiesPlay.
[14] Kang, Minki, et al. "Knowledge-augmented reasoning distillation for small language models in knowledge-intensive tasks." Advances in Neural Information Processing Systems 36 (2024).
[15] Armengol-Estapé, Jordi, et al. "SLaDe: A Portable Small Language Model Decompiler for Optimized Assembly." 2024 IEEE/ACM International Symposium on Code Generation and Optimization (CGO). IEEE, 2024.
[16] Wu, Chaoyi, et al. "PMC-LLaMA: toward building open-source language models for medicine." Journal of the American Medical Informatics Association (2024): ocae045.
[17] Wang, Yuling, et al. "Can Small Language Models be Good Reasoners for Sequential Recommendation?." arXiv preprint arXiv:2403.04260 (2024).
[18] Hu, Shengding, et al. "Minicpm: Unveiling the potential of small language models with scalable training strategies." arXiv preprint arXiv:2404.06395 (2024).
[19] Zhu, Yichen, et al. "LLaVA-$\phi $: Efficient Multi-Modal Assistant with Small Language Model." arXiv preprint arXiv:2401.02330 (2024).