ENHANCING PREDICTIVE ACCURACY AND RESOURCE UTILIZATION IN DISEASE PREDICTION : A COMPARATIVE STUDY OF SMALL VS LARGE LANGUAGE MODELS

MINI PROJECT REPORT

Submitted by

Tamanna **210701281**

in partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

RAJALAKSHMI ENGINEERING COLLEGE, CHENNAI

ANNA UNIVERSITY:: CHENNAI 600 025

APRIL 2024

RAJALAKSHMI ENGINEERING COLLEGE, CHENNAI

BONAFIDE CERTIFICATE

Certified that this Report titled "**Enhancing Predictive Accuracy and Resource Utilization in Disease Prediction: A Comparative Study of Small Versus Large Language Models**" is the bonafide work of **"Tamanna (210701281)"** who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Rahul Chiranjeevi. V
Assistant Professor,
Department of Computer Science and Engineering,
Rajalakshmi Engineering College,
Chennai – 602015

Submitted to Mini Project Viva-Voce Examination held on ___________________

Internal Examiner **External Examiner**

ABSTRACT

This project involves a comparative analysis of small language and large language models where it evaluates the models' capabilities in deciphering complex relationships between diseases and their symptoms using the distilgpt2(SLM) and gpt2(LLM) models. We utilized a large dataset where both small and large language models were used to study diseases and their symptoms. The models were evaluated based on their computational efficiency, predictive performance and the resources required for the operation. The methodology includes loading datasets, tokenization and model setup, monitoring training and validation losses, hyperparameter tuning to optimize the models and eventually generating text. The study showed SLM's proficiency in producing context-aware responses whereas LLM's shows its strength in generating refined and comprehensive text. The outcome showed the ability of large language models to possess higher predictive accuracy and they demanded significantly higher computational resources as compared to small language models, which may not be advised to use in resource-constrained environment. In contrast, small language models have efficient resource usage despite their low accuracy make their application more pertinent where computational resources are limited. This study paves the way for understanding the blueprint for health informatics practitioner highlighting the importance of domain-specific training in enhancing predictive accuracy and resource utilization of language models. It also underscores the requisite for a new approach to deployment of language models in healthcare settings.

ACKNOWLEDGEMENT

Initially we thank the Almighty for being with us through every walk of our life and showering his blessings through the endeavour to put forth this report. Our sincere thanks to our Chairman **Mr. S.MEGANATHAN, B.E, F.I.E.**, our Vice Chairman **Mr. ABHAY SHANKAR MEGANATHAN, B.E., M.S.,** and our respected Chairperson **Dr. (Mrs.) THANGAM MEGANATHAN**, **Ph.D.,** for providing us with the requisite infrastructure and sincere endeavoring in educating us in their premier institution.

Our sincere thanks to **Dr. S.N. MURUGESAN, M.E., Ph.D.,** our beloved Principal for his kind support and facilities provided to complete our work in time. We express our sincere thanks to **Dr. P. KUMAR, Ph.D.**, Professor and Head of the Department of Computer Science and Engineering for his guidance and encouragement throughout the project work. We convey our sincere and deepest gratitude to our internal guide, **Rahul Chiranjeevi V** Professor, Department of Computer Science and Engineering. Rajalakshmi Engineering College for his valuable guidance throughout the course of the project.

Tamanna-210701281

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

SLM	Small Language Models
LLM	Large Language Model
BERT	Bidirectional Encoder Representations from Transformers
LSTM	Long Short Term Memory
GPT-2	Generative Pre-trained Transformer 2
CoT	Chain of thought

CHAPTER 1
INTRODUCTION

1.1 GENERAL

The area of investigation of small and large language models in disease prediction is a rapidly growing field of study that has significant potential for the future of healthcare. The SLMs with their few parameters and smaller datasets are characterized by their speed and efficiency in processing, thus, they are suitable for the applications where computational resources are limited making them suitable for real-time applications. The other way around, large language models (LLMs) are characterized by their huge number of parameters and the intensive training on the large datasets, thus, they can perform the complex language tasks with high accuracy and are capable of understanding complex symptom descriptions and providing more accurate disease associations.

1.2 OBJECTIVE

The aim of the study is to:

- To find out the linkage between diseases and symptoms by finding out how efficient their response is and its predictive performance.

- The resources needed by language models to carry out the functions needs to be specified.

- In order to improve the predictive performance of the model, it is necessary to understand the magnitude of domain-specific training.

- To understand the practicality of these models in real world environments, it is mandatory to deploy and put them to use.

1.3 EXISTING SYSTEM

The existing system involves using GPT-2 which is a large language model for prediction of disease based on the symptoms and vice versa. Through bespoke fine-tuning they have been trained on domain specific data to find the diseases based on the input symptoms [7]. They were trained on massive datasets to capture patterns and apply the medical knowledge. A well-known algorithm in this area is the Medical Concept Normalization-Bidirectional Encoder Representations from Transformers (MCN-BERT), which is a very accurate tool for disease prediction from symptom descriptions[12].They tackled this challenge by harnessing the power of language models like BERT that can learn by finding out intricate connections between the symptoms and the diseases .This model can enhance their disease predictive accuracy by getting trained on large medical data thereby improving domain-specific knowledge. Bidirectional LSTM layers were studied to investigate the influence of their use on the comprehension of the contextual relations between symptoms and diseases. The technique of hyperparameter optimization using Hyperopt was employed to increase the model's performance and thus, the model will be able to generalize well to the new data. In the existing system MCN-BERT model was use that followed the below steps:

- **Data collection and preprocessing -** This step involved collecting input datasets which consisted of symptoms mapped to their corresponding disease labels. The data was preprocessed by using a medical tokenizer where symptoms were diligently tokenized to enhance the contextual understanding of medical terms.

- **BERT model and tokenizer initialization -** The model was fed with pre-trained model weights which contain detailed linguistic knowledge that enhance the model's ability to grasp intricate medical terms. BERT models were architectured in such a way to include multiple transformer encoder layers and a specialized tokenizer was initialized. This tokenizer seamlessly converted the symptom description to tokens that were in turn integrated into the model as input.

- **Model Training** - The course of MCN-BERT training includes a thorough analysis of all stages to guarantee the correct learning process. Firstly, batches of tokenized symptom description and corresponding disease labels are organized as in the equation below

  BatchData=PrepareBatches(TokenizedSymptoms, DiseaseLabels)

  where BatchData and PrepareBatch are the prepared training batches, and function for PrepareBatches orchestrates this preparation. The following is the Model Forward Pass in which we propagate the batches through the BERT model using its forward pass mechanism.

- **Model Evaluation** - Well-known evaluation metrics, including accuracy, precision, recall, and F1-score, were used to assess the models' classification performance.

  Accuracy = TP + TN / TP + FP + TN + FN

  Precision = TP / TP + FP

  Recall = TP / TP + FN

  F1 − score = 2 ∗ (10) (Precision ×

  Recall) / (Precision + Recall)

1.4 PROPOSED SYSTEM

The proposed work aims to perform 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 research will further utilize these SLMs such as DistilGPT-2 and LLMs which includes the GPT-2 or BERT which have been shown to perform impressively in natural language processing tasks in diverse ways. These models will be adjusted to the data set of symptoms and illnesses, their performance will be evaluated based on the level that they can correctly predict diseases from the symptom descriptions. Similarly, the study will evaluate both models for inference time in addition to the predictive accuracy parameter. This approach will help to determine a comparative study of resource usage by both SLMs and LLM. Besides, the study will use the comparison of the loss function of the both models to measure their predicting accuracy level. The proposed work will, therefore, prove to be a vital asset in the field of disease prediction as it will offer an exhaustive appraisal of SLMs and LLMs. The findings from this experiment might serve as a guide for later studies in the field and help create better and comprehensive disease prediction models. This work is an innovative and well separated concept that not only considers the precision of the models but also the computational efficiency and the resource utilization. This makes it a very important contribution to the field of disease prediction.

CHAPTER 2
LITERATURE SURVEY

[1]Victor Sanh, Lysandre Debut, Julien Chaumond, Thomas Wolf in this paper attempted to create a small language model that is equally efficient without compromising its performance. It mentions the issues they usually face while training large language models when the resources and data available are limited . Their approach of using knowledge distillation during the pre-training phase is notable and shows promising results. Their work provides a motivation to develop more efficient small language models with high accuracy and better performance.

[2]In this paper, the authors are proving the efficiency of BERT in different NLP cases, such as question answering and language inference tasks.The BERT model was put to test and it was observed how it outweighed the performance of other models. BERT's capability to perform complex language processing tasks with utmost efficiency eventually leads to the understanding that it can similarly perform disease prediction with high accuracy by interpreting the intricate details of connection between symptoms and diseases.

This study[3] by Timo Schick and Hinrich Schütze analyzed the performance small language models compared to large language models.They showed how small models trained with few parameters as compared to large models can perform equally in predicting the result. By converting textual inputs into cloze questions and leveraging gradient-based optimization, these "greener" models achieve impressive natural language understanding.

This research by Leonardo Ranaldi and André Freitas in [4] tried to develop a method to bridge the gap in reasoning skills between small and large language models.With the help of Instruction-tuning-CoT method they provide SLMs with ability to perform multi step controlled reasoning when elicited with the CoT mechanism.

This paper [5] by Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, and others show how scaling up the models improve their performance sometimes even rivaling prior state-of-the-art fine-tuning approaches.They trained GPT-3 with 175 billion parameters and tested their performance without any fine tuning.They demonstrated how GPT-3 produced strong results on translation,question-answering and other reasoning based challenges.

[7]The paper "Large language models in health care: vulnerabilities, risks, and challenges" discusses the LLMs opportunities in healthcare. It briefly touches upon the problems of using machine learning systems in biomedical and clinical fields being applied in clinical environments, and those challenges which need to be solved in order to widely adopt their usage in this field.

This paper introduces us to TinyLlama which is a compact language model pre-trained on approximately 1 trillion tokens for around 3 epochs[8] It significantly outperforms existing open-source language models with comparable dimensions.

In [9] Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun explores the potential of Small Language Models as resource-efficient alternatives to Large Language Models.While focusing on SLMs, our approach exhibits scalability in both model and data dimensions for future LLM research.

In [10], Awobade, Oduwole, and Kolawole investigate the effect of compression techniques on small language models. Practical strategies such as pruning, knowledge distillation, and quantization have been applied on AfriBERTa, a low-resource and small-data BERT system for Koine Greek. These techniques were performed to find out the which one had more effect in improving the model's efficiency
The research gap in this study is the absence of the comparative analysis of small and large language models (SLMs and LLMs) majorly in disease prediction in terms of their predictive accuracy and resource utilization. Even though there are studies on the individual capabilities of SLMs and LLMs,this research aims to bridge the gap by giving a direct comparison within the stringent constraints and precise requirements of healthcare informatics.

This paper [11] showed NLPs effectiveness varies across diseases.They conducted a search across major databases including PubMed, Embase, Web of Science, and Scopus, up to December 2023, using keywords related to NLP, LLM, and infectious diseases.

In [12] the research explores the use of LLMs and deep learning techniques for predicting disease from symptoms.It analyzed two Medical Concept Normalization—Bidirectional Encoder Representations from Transformers (MCN-BERT) models and a Bidirectional Long Short-Term Memory (BiLSTM) model and demonstrated how accurately they predict the output.

The research in [14] by Minki Kang, Seanie Lee, Jinheon Baek, Kenji Kawaguchi, and Sung Ju Hwang devised a new approach to performing knowledge intensive tasks in small language models. By leveraging the use of external-knowledge resources ,they utilize knowledge-augmented reasoning distillation to improve the performance of SLMs.

This paper [16] 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.

This paper by Wang et al. introduced a novel framework called Step-by-step knowledge distillation framework for recommendation (SLIM)[17] .In this research paper they tried to elevate the reasoning capabilities of large language models to work in a resource efficient manner by enabling sequential recommenders. They distill knowledge from large models and feed it into small models based on user behavior sequences.

In this paper [18] addresses the concerns associated with developing large language models with up to trillion parameters. To address these limitations, they present MiniCPM which consists of: 1. 2B and 2. 4B non-embedding parameter variants. It is worth to note that in each of these categories these Small Language Models (SLMs) themselves perform very well and are on a par with 7B-13B LLMs.

In [19] they introduced LLaVa-Phi which is an multi modal assistant that harnessed the power of small language models Phi-2.It showed that even small language models even those with 2.7B parameters can excellently engage in complex dialogues that combine text and visual elements trained on quality datasets.

This research paper [20] 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.

CHAPTER 3
SYSTEM DESIGN

3.1 DEVELOPMENT ENVIRONMENT

3.1.1 HARDWARE SPECIFICATIONS

This project uses minimal hardware but in order to run the project efficiently without any lack of user experience, the following specifications are recommended:

Table 3.1.1 Hardware Specifications

PROCESSOR	Intel Core i5
RAM	4GB or above (DDR4 RAM)
GPU	Intel Integrated Graphics
HARD DISK	6GB
PROCESSOR FREQUENCY	1.5 GHz or above

3.1.2 SOFTWARE SPECIFICATIONS

The software specifications in order to execute the project has been listed down in the below table. The requirements in terms of the software that needs to be pre- installed and the languages needed to develop the project has been listed out below.

Table 3.1.2 Software Specifications

FRONT END	HTML, CSS, Bootstrap, JavaScript
BACK END	Python, Django
FRAMEWORKS	Pytorch, Tensor Flow
SOFTWARES USED	Visual Studio, Jupyter Notebook

3.2 SYSTEM DESIGN

3.2.1 ARCHITECTURE DIAGRAM

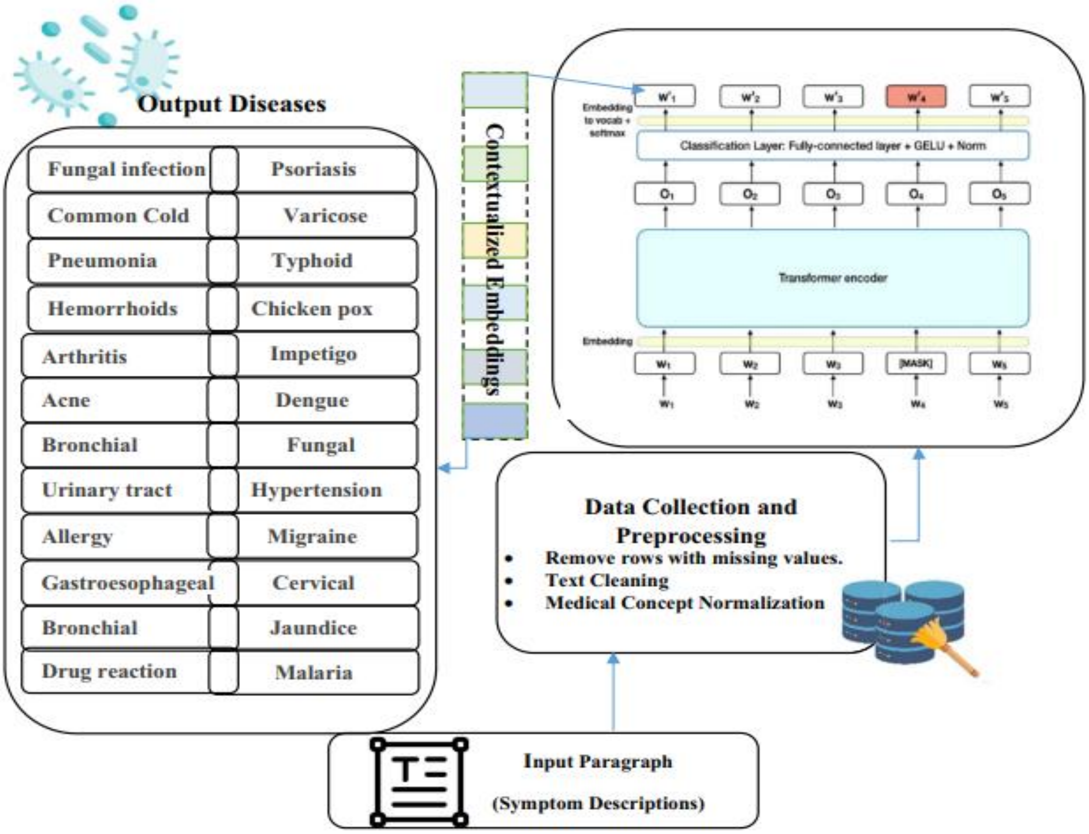


Fig 3.2.1 Architecture Diagram

APPROACH: UNDERSTANDING SLMs AND LLMs

Small Language Models (SLMs):

- **Architecture**: Language Models with a reduced context are called Small Language Models (SLMs). SLMs describe a neural network with fewer layers. They often utilize less numbers of layers and parameters, thus ensuring faster training and inference time.
- **Training:** SLM that are built by only using smaller datasets are less computationally expensive. This training method is optimal for situations, where there is no data available in high volume to work with.
- **Efficiency:** The lean construction or architecture of the SLMs is what leads to fast operational times and maximum efficiency[9].They can process language tasks rapidly, making them ideal for real-time applications.
- **Application:** Due to their resource efficiency, SLMs are being effectively implemented in situations where computational resources are a limitation or a real-time analysis is necessary.

Large Language Models (LLMs):

- **Architecture:** Large Language Models (LLMs) are based on a neural network of a great complexity that involves a deep learning approach. They are characterized by a higher number of trainable parameters that give them the ability to express complex patterns in data.
- **Training:** LLMs are trained with the aid of massive data that has an enormous variety. This procedure is an essential but energy-prodigal part of the model that enables it to carry out complicated language tasks.
- **Predictive Performance:** LLMs are well-known for their high accuracy and generating output that is not only detailed but also of good quality. Such language mastery helps to establish disease correlations with greater precision.
- **Resource Requirements**: The elaborate features of LLMs bear the tradeoff of more computing resources. They need more powerful hardware and consume more energy, thus in the places of resource scarcity, their performance can be low

CHAPTER 4

PROJECT DESCRIPTION

4.1 MODULE DESCRIPTION

4.1.1 DATA PRE-PROCESSING:

Like the present one, the data is prepared for processing by standardizing the symptoms in the form of a single string separated by commas.

4.1.2DATASET 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.

4.1.3MODEL 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).

4.1.4MODEL 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.

4.1.5 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.

4.1.6 INFERENCE TIME MEASUREMENT:

Both such models' inference time is measured. It meant, feeding an input text into both the models and measuring the time, how fast these models generated the output.

4.1.7 INFERENCE TIME COMPARISON:

Two models are compared in terms of the inference times and a corresponding bar chart is used to illustrate this.

4.1.8 VALIDATION LOSS COMPARISON:

Bar graphs are employed to show the validation losses of the two models under testing. Moreover, the losses are compared.

CHAPTER 5
IMPLEMENTATION AND RESULTS

5.1 IMPLEMENTATION

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.
In this comparison:

Training Loss: 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.

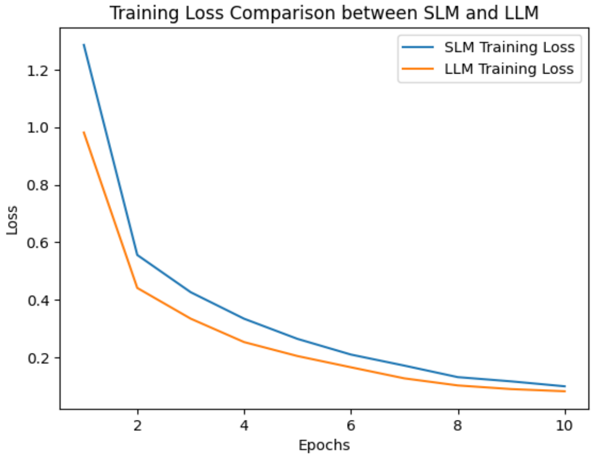


Fig.5.1.1 Training Loss Comparison

Validation Loss: 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.

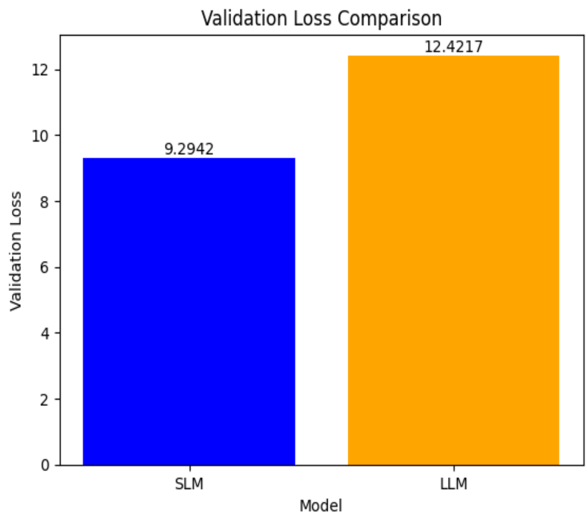


Fig.5.1.2 Validation Loss Comparison

Inference Time: 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.

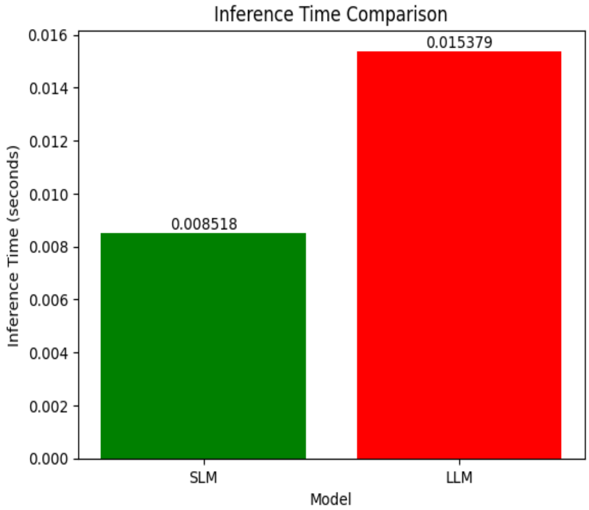


Fig.5.1.3 Inference Time Comparison

5.2 RESULTS

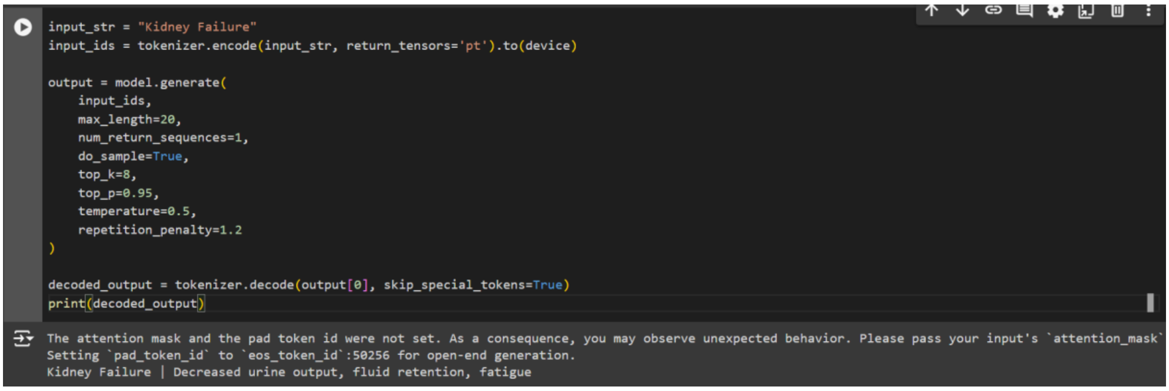
The comparison of the models based on the above three values gave the following results:

Model	Training Loss	Validation Loss	Inference Time
SLM	Higher	9.2942	0.008518 secs
LLM	Lower	12.4217	0.015379 secs

Table .5.2.1 Model Comparison

5.3 OUTPUT SCREENSHOTS

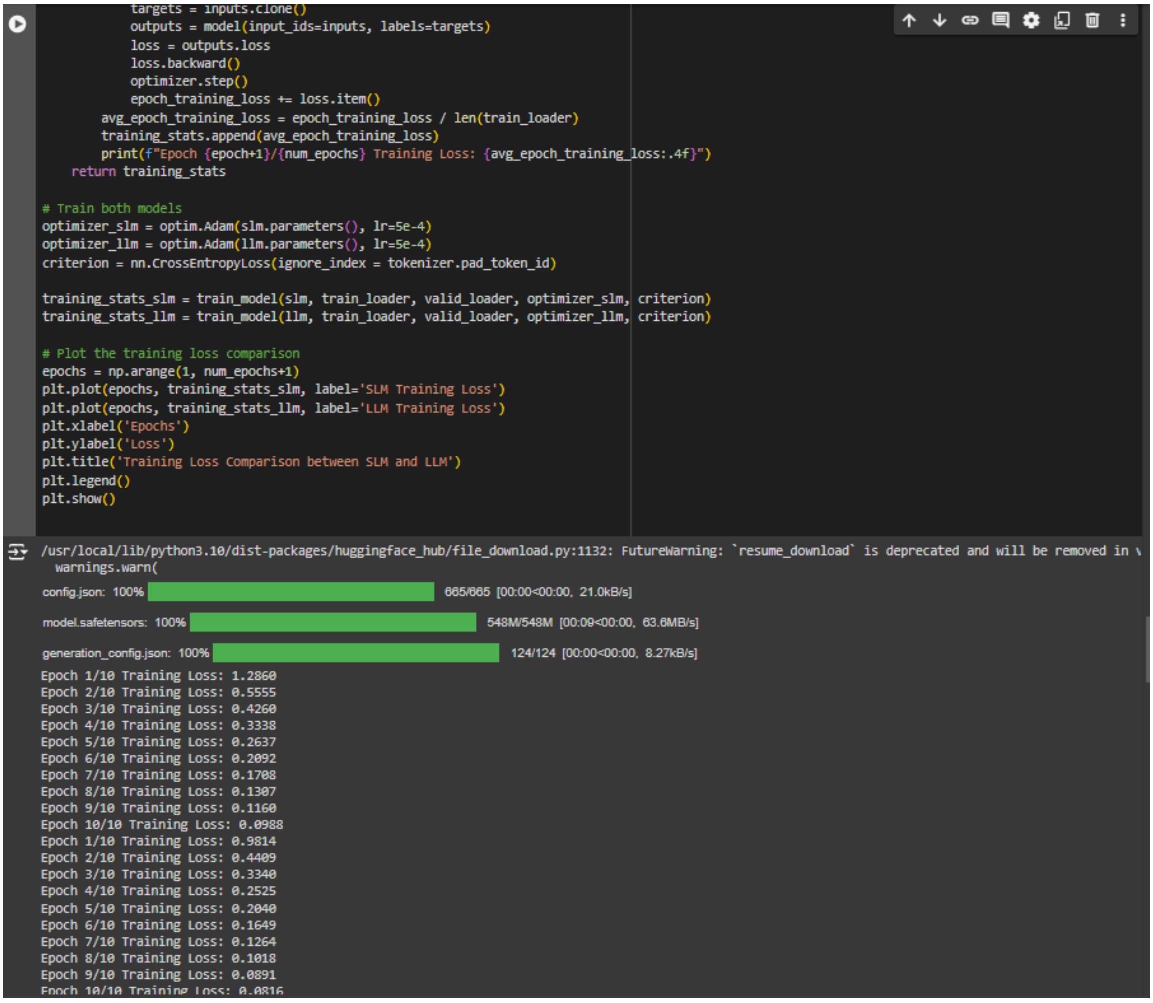
The below screenshot shows the prediction of diseases from symptoms and vice versa.



```
input_str = "Kidney Failure"
input_ids = tokenizer.encode(input_str, return_tensors='pt').to(device)

output = model.generate(
    input_ids,
    max_length=20,
    num_return_sequences=1,
    do_sample=True,
    top_k=8,
    top_p=0.95,
    temperature=0.5,
    repetition_penalty=1.2
)

decoded_output = tokenizer.decode(output[0], skip_special_tokens=True)
print(decoded_output)

The attention mask and the pad token id were not set. As a consequence, you may observe unexpected behavior. Please pass your input's `attention_mask`
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
Kidney Failure | Decreased urine output, fluid retention, fatigue
```

Fig.5.3.1 Prediction of symptom and disease

The below screenshot shows the comparison of performance of SLMs and LLMs based on certain parameters.



```
            targets = inputs.clone()
            outputs = model(input_ids=inputs, labels=targets)
            loss = outputs.loss
            loss.backward()
            optimizer.step()
            epoch_training_loss += loss.item()
        avg_epoch_training_loss = epoch_training_loss / len(train_loader)
        training_stats.append(avg_epoch_training_loss)
        print(f"Epoch {epoch+1}/{num_epochs} Training Loss: {avg_epoch_training_loss:.4f}")
    return training_stats

# Train both models
optimizer_slm = optim.Adam(slm.parameters(), lr=5e-4)
optimizer_llm = optim.Adam(llm.parameters(), lr=5e-4)
criterion = nn.CrossEntropyLoss(ignore_index = tokenizer.pad_token_id)

training_stats_slm = train_model(slm, train_loader, valid_loader, optimizer_slm, criterion)
training_stats_llm = train_model(llm, train_loader, valid_loader, optimizer_llm, criterion)

# Plot the training loss comparison
epochs = np.arange(1, num_epochs+1)
plt.plot(epochs, training_stats_slm, label='SLM Training Loss')
plt.plot(epochs, training_stats_llm, label='LLM Training Loss')
plt.xlabel('Epochs')
plt.ylabel('Loss')
plt.title('Training Loss Comparison between SLM and LLM')
plt.legend()
plt.show()

/usr/local/lib/python3.10/dist-packages/huggingface_hub/file_download.py:1132: FutureWarning: `resume_download` is deprecated and will be removed in v
  warnings.warn(
config.json: 100% 665/665 [00:00<00:00, 21.0kB/s]
model.safetensors: 100% 548M/548M [00:09<00:00, 63.6MB/s]
generation_config.json: 100% 124/124 [00:00<00:00, 8.27kB/s]
Epoch 1/10 Training Loss: 1.2860
Epoch 2/10 Training Loss: 0.5555
Epoch 3/10 Training Loss: 0.4260
Epoch 4/10 Training Loss: 0.3338
Epoch 5/10 Training Loss: 0.2637
Epoch 6/10 Training Loss: 0.2092
Epoch 7/10 Training Loss: 0.1708
Epoch 8/10 Training Loss: 0.1307
Epoch 9/10 Training Loss: 0.1160
Epoch 10/10 Training Loss: 0.0988
Epoch 1/10 Training Loss: 0.9814
Epoch 2/10 Training Loss: 0.4409
Epoch 3/10 Training Loss: 0.3340
Epoch 4/10 Training Loss: 0.2525
Epoch 5/10 Training Loss: 0.2040
Epoch 6/10 Training Loss: 0.1649
Epoch 7/10 Training Loss: 0.1264
Epoch 8/10 Training Loss: 0.1018
Epoch 9/10 Training Loss: 0.0891
Epoch 10/10 Training Loss: 0.0816
```

Fig.5.3.2 Comparison of model's performance

CHAPTER 6

CONCLUSION AND FUTURE ENHANCEMENTS

6.1 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 gamut’s 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.

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.

6.2 FUTURE ENHANCEMENTS

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.

[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).

[20] Awobade, Busayo, Mardiyyah Oduwole, and Steven Kolawole. "What Happens When Small Is Made Smaller? Exploring the Impact of Compression on Small Data Pretrained Language Models." arXiv preprint arXiv:2404.04759 (2024).

[21] Magister, Lucie Charlotte, et al. "Teaching small language models to reason." arXiv preprint arXiv:2212.08410 (2022).

[22] Fu, Yao, et al. "Specializing smaller language models towards multi-step reasoning." International Conference on Machine Learning. PMLR, 2023.