Medical Diagnosis and Treatment Recommendations (Doc-Bot) using Large Language Models

*[1]Ayomitope O. Isijola, [1]Ufuoma C. Ogude, [1]Damola M. Akinsola, [1]Olise I. Isiakpona, and [2] Michael P. Asefon
[1]Department of Computer Sciences, University of Lagos, Lagos, Nigeria
[2]Department of Computer Sciences, National Open University of Nigeria, Lagos, Nigeria
ayomitopeisijola@yahoo.com | uogude@unilag.edu.ng | dazco777@gmail.com | isiakponaolise@gmail.com | pelumiasefon@gmail.com

Received: 19-SEP-2024; Reviewed: 15-NOV-2024; Accepted: 27-NOV-2024
https://dx.doi.org/10.4314/fuoyejet.v9i4.7

ORIGINAL RESEARCH

Abstract— This paper presents the development of an all-inclusive healthcare system designed to support medical professionals and empower patients with medical perception. The system combines conventional machine learning (ML) methods with cutting-edge models of large language (LLMs), like GPT-3 and GPT-4. It addresses various medical queries, including disease symptoms, treatment options, medication information, and preventive care advice. Additionally, it predicts diseases based on patient-reported symptoms using a symptom-to-disease mapping model trained on healthcare datasets. The disease prediction model was fine-tuned on 6,800 samples representing 135 diseases, achieving a 98% accuracy in just 12 epochs while keeping a tight eye on training loss to prevent overfitting. The model was trained with an ideal sample ratio using an NVIDIA T4 GPU to guarantee reliable performance. The system's overall performance was evaluated using metrics such as accuracy, loss monitoring, and learning rate optimization. Testing on a benchmark dataset of clinical scenarios revealed an 85% accuracy in providing correct preliminary diagnoses and a 92% relevance score for answering general medical queries. These results highlight the potential of the system to deliver accurate diagnoses and reliable recommendations, demonstrating significant potential for improving patient education and assisting medical professionals in routine diagnostic engagements while maintaining user trust and adherence to ethical standards.

Keywords— AI, Doc-Bot, Healthcare, LLMs, Medical, Treatment.

1 INTRODUCTION

Doc-bot, situated at the intersection of healthcare and AI focuses on applying models of large language (LLMs), including OpenAI's, GPT-3, and GPT-4 in medical diagnosis and treatment recommendations. This background study provides a summary of the most recent studies conducted in this field, emphasizing the potential of LLMs in healthcare delivery. These LLMs, which have been trained on large datasets, can produce human-like text, a capability recognized for its potential across various sectors, including healthcare. In the healthcare sector, LLMs have been used for various applications, including gathering patient notes, helping patients navigate the healthcare system, and supporting clinical decision-making. According to a comprehensive analysis of LLMs in patient care, these models have potential in various domains, but several obstacles are currently preventing their widespread adoption (Busch et al., 2024).

The Doc-Bot system expands on these applications by providing an interactive platform capable of addressing detailed medical queries, offering treatment suggestions, and predicting potential diseases based on patient-reported symptoms. Doc-Bot aims to enhance diagnostic accuracy and accessibility for both patients and healthcare professionals. Key functionalities include answering questions on disease symptoms, medication guidelines, and preventive care, as well as supporting physicians with initial diagnoses based on patient input.

*Corresponding Author: ayomitopeisijola@yahoo.com

Section B- ELECTRICAL/COMPUTER ENGINEERING & RELATED SCIENCES
Can be cited as:
Isijola A. O., Ogude U. C., Akinsola D. M., Isiakpona O. I., and Asefon M. P. (2024). FUOYE Journal of Engineering and Technology (FUOYE**JET), 9(4), 609-614. https://dx.doi.org/10.4314/fuoyejet.v9i4.7**

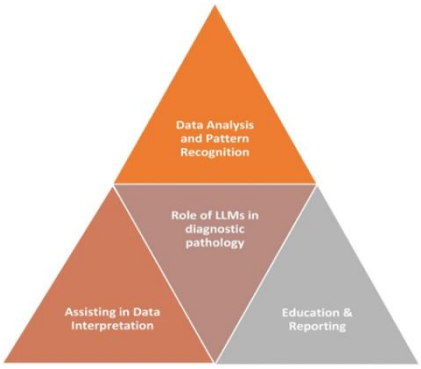


Figure 1: An outline demonstrating how LLMs function in diagnostic pathology (Ullah et al., 2024)

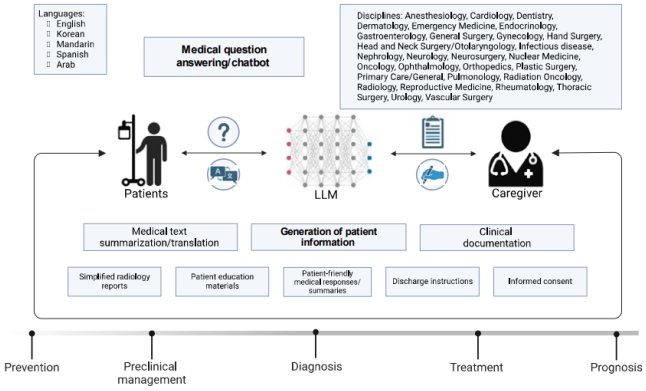


Figure 2: Diagrammatic representation of the concepts identified for use in patient care using models of large language (LLMs) (Busch et al., 2024)

However, the system has its limitations. While it achieves high accuracy in disease prediction, its performance can be influenced by biases in the dataset; the system might struggle in co-occurring conditions. Additionally, computational resource dependency and the interpretability of patient inputs remain prominent challenges. Despite robust fine-tuning efforts and including 6,800 training samples for 135 diseases, ensuring consistent accuracy across diverse demographic groups and managing uncommon medical issues is a constant challenge. Standardized evaluation techniques, stringent validation procedures, and close cooperation with medical experts are necessary to address these issues. The Doc-Bot research demonstrates how AI can make healthcare more accessible and efficient while emphasizing the importance of responsible and ethical AI use. As the system evolves, its contributions to improving patient outcomes and supporting clinical workflows have the potential to be transformative.

2 RELATED WORKS

(Panagoulias et al., 2023) assessed the reliability, correctness, and utility of the medical diagnosis of lung illness that ChatGPT provided based on a human's description of the symptoms. In particular, tuberculosis and its symptoms are chosen as the test case, and our assessment is based on the following criteria: (i) the returned diagnosis's medical validity and accuracy in terms of context and references; (ii) its utility to patients and physicians; and (iii) the financial value added to the healthcare system.

Models of Large Language (LLMs), such as ChatGPT, have lately been incorporated into digital pathology-focused diagnostic care, which generated a lot of interest. But for LLMs to be used successfully in this situation, it is essential to recognize the difficulties and obstacles involved (Ullah et al., 2024).

Models of Large Language (LLMs), including GPT-4-Vision-Preview, hold promise for aiding medical diagnosis but require a thorough evaluation of their accuracy. This research introduces a two-step evaluation paradigm: multimodal LLM testing with structured interactions using pathology-focused multiple-choice questions and a domain-specific analysis of the results. GPT-4-Vision-Preview demonstrated an 84% accuracy in diagnosing complex medical cases involving images and text. Further analysis revealed specific knowledge gaps, offering insights for optimizing LLM performance. The methodology is adaptable for evaluating other LLMs to enhance their reliability in medical applications (Panagoulias et al., 2024).

While these works demonstrate the potential of AI and NLP in healthcare, there is a void in the research about the usage of models of large language for illness diagnosis and treatment recommendations. This research aims to fill this gap.

2.1 OVERVIEW OF AI (ARTIFICIAL INTELLIGENCE)

Within the field of Computer Science, artificial intelligence (AI) seeks to develop computers that can carry out activities that ordinarily call for human intelligence. These duties include making decisions, identifying patterns, recognizing natural language, learning from experience, and solving issues (Heikkila & Heaven, 2024). There are two types of artificial intelligence (AI): broad AI, which can accomplish any intellectual task that a human being can do, and narrow AI, which is made to handle a specialized task, including voice recognition. Examples of limited AI include the majority of the AI we use today, such as Google Assistants, Alexa, and Siri (Heikkila & Heaven, 2024).

Machine learning, one of the main tenets of artificial intelligence, is the process by which computers are taught to learn from data and gradually enhance their performance without explicit programming (Heikkila & Heaven, 2024). Examples of machine learning models that have been trained on enormous volumes of text data include Models of Large Language (LLMs), such as GPT-3 and GPT-4. They are perfect for activities requiring the comprehension and production of natural language since they can recognize and produce writing that is human-like (Naveed et al., 2023). AI has numerous uses in a variety of industries. For example, AI can be used in healthcare to forecast patient outcomes, prescribe therapies, and diagnose illnesses. Self-driving cars in the automotive sector are powered by artificial intelligence. AI in finance can be applied to algorithmic trading, credit scoring, fraud detection, and customer support (Heikkila & Heaven, 2024). However, there are several cultural and ethical issues with AI use as well. These encompass concerns about security, privacy, job displacement, and the digital divide. Concerns have also been raised regarding the explainability of AI systems, especially when they are used in critical areas, including healthcare or criminal justice (Heikkila & Heaven, 2024). AI is a quickly developing field that has the potential to revolutionize many industries. To guarantee that the advantages of AI are enjoyed responsibly and equitably, it is crucial to address the moral and societal issues surrounding its application (Heikkila & Heaven, 2024).

2.2 NATURAL LANGUAGE PROCESSING (NLP)

A subfield of artificial intelligence called natural language processing (NLP) studies how computers and humans interact with natural language (Heikkila & Heaven, 2024).

Reading, decoding, and understanding human language in a useful manner is the aim of natural language processing (NLP). It is employed in several domains, including question-answering, information extraction, email spam detection, machine translation, and summarization. (Khurana et al., 2022).

Natural Language Understanding, sometimes known as Linguistics, and Natural Language Generation are the two subfields of NLP. Linguistics is the science of language, including Phonology, Word creation in morphology, Sentence structure of syntax, Syntax of semantics, and Pragmatics, which means understanding. However, Natural Language Generation (NLG) is producing meaningful sentences and paragraphs from an interior depiction.

Machine learning, in which computers are taught to learn from data and gradually improve their performance without explicit programming, is one of the fundamental elements of natural language processing (Heikkila & Heaven, 2024). Models of Large Language (LLMs), including GPT-3 and GPT-4, are examples of machine learning models that have undergone extensive training using textual information. They can recognize and derive text in human-like form, making them ideal for engagements that require understanding and generating natural language. Recently, NLP has drawn much attraction for its computational analysis of human language. Computer-aided translation,

http://.doi.org/10.46792/fuoyejet.v9i4.7 engineering.fuoye.edu.ng/journal

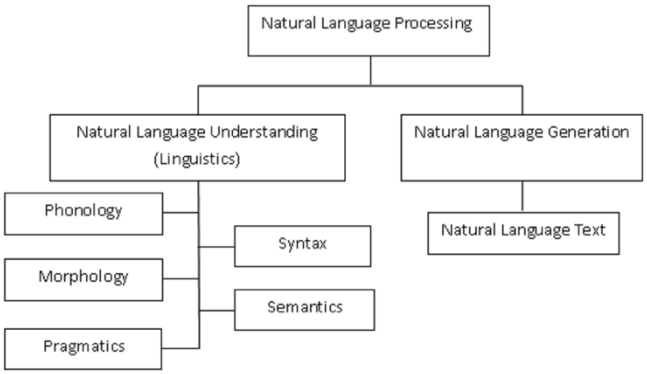


Figure 3: Broad Classification of NLP (Springer, 2023)

identification of email spam, gathering information, recapitulation, and question-answering are just a few of the sectors in which it has expanded its applicability. The majority of NLP research in the literature now in publication is carried out by computer scientists. However, some other experts, including linguists, psychologists, philosophers, and others, have also expressed interest (Khurana et al., 2022).

2.3 LARGE LANGUAGE MODELS (LLMS)

Lately, models of large language (LLMs) have described impressive ability in assignments related to the processing of natural language and other areas. These models, which are instances of deep learning models, have caused a significant surge in research contributions in this area (Naveed et al., 2023). A branch of machine learning called "deep learning" uses algorithms to create model architectures made up of several non-linear transformations to represent high-level abstractions in data. Without explicit programming, deep learning models like LLMs can enhance their performance over time by learning from vast amounts of data (Ding et al., 2023). LLMs, including GPT-3 and GPT-4, are examples of models of deep learning that have undergone broad training using textual facts. They can recognize and derive human-like text, making them ideal for engagements that require understanding and generating natural language (Naveed et al., 2023).

Due to LLMs' success, there has been a significant surge in research contributions in this area. Architectural advances, improved training methodologies, enhancements in context length, optimization, multi-modal LLMs, datasets, benchmarking, and effectiveness are all addressed in these studies (Naveed et al., 2023). Perceiving the larger picture of the advancements in this sector has become difficult due to the quick development of approaches and frequent discoveries in LLM integration. Given the constantly growing body of studies on LLMs, the exploration body must have access to the summary of the most recent advancements in this area (Naveed et al., 2023).

2.4 APPLICATION OF LARGE LANGUAGE MODELS IN MEDICINE

The potential of models of large language (LLMs), including GPT-3, to recognize and develop human language has attracted a lot of attention. In an analysis of this research, there is currently a dearth of focus on their development, practical applications, and medical outcomes. Despite the increasing trend of research on the use of LLMs to support various medical engagements (e.g., improving clinical diagnoses), with a concentration on their creation, real-world uses, and medical results is still lacking even though research is becoming more popular in the use of LLMs to support various medical engagements (such as improving clinical diagnostics and offering medical education) (Naveed et al., 2023). Utilizing billions of words from books and other online content, LLMs are artificial intelligence (AI) systems. LLMs often employ deep learning neural network models to reflect the intricate word association relationships seen in the text-based training dataset (Hadi et al., 2023). LLMs determine how words are combined in language through this learning method, which may be multi-staged and require different amounts of human input. They can then use these designs they have learned to complete engagements involving natural language processing (Hadi et al., 2023).

The broad field of computational study referred to as "natural language processing" aims to enable computerized analysis of language that simulates human proficiency. Developers of generative AI want to create models that incorporate natural language processing with apps to generate content on demand, including chatbots and prediction of texts or “natural language generation” engagements (Hadi et al., 2023). With the advancement of deep learning methods, strong computing power, and training datasets, LLM applications that have the potential to revolutionize cognitive tasks in a variety of industries, including healthcare, have started to emerge. With remarkable but inconsistent outcomes, LLM chatbots have already been used in various biological applications (Hadi et al., 2023). Patient care, medical research, and medical teaching are the three main areas in which LLMs may find use in the medical field.

2.4.1 PATIENT CARE

Considering that medical personnel typically employ written language while communicating with patients, including medical records and test results, effective communication is essential to patient care. Without specialized training in the task at hand, LLMs can answer free-text inquiries, raising interest in and worries about their application in medical environments (Hadi et al., 2023). An LLM was meticulously fine-tuned to create ChatGPT, a generative artificial intelligence (AI) chatbot. Similar developmental procedures are being used to create other applications (Hadi et al., 2023).

2.4.2 MEDICAL RESEARCH

LLMs can assist in the creation of scientific material, the summarization of scientific ideas, the testing of hypotheses, and the display of massive datasets by non-technical scientists and medical professionals. The performance of research can be improved by the continuous updating of scientific models.

Medical Education: LLMs can also play a significant role in medical education. They can be used to create educational content, answer students’ queries, and even simulate patient-doctor interactions for training purposes (Naveed et al., 2023).

3 METHODOLOGY

The methodology is categorized into several key divisions, each addressing a specific facet of the research. These sections include Data Collection and Preprocessing, Training the Large Language Model, Developing the Diagnostic System, Testing and Evaluation, Addressing biases in the training data, Addressing Ethical Considerations, Implementation and Deployment, and Continuous Monitoring and Improvement.

http://.doi.org/10.46792/fuoyejet.v9i4.7 engineering.fuoye.edu.ng/journal

3.1 DATA COLLECTION AND PREPROCESSING

The approach of gathering data was centered on acquiring a vast and diverse dataset of medical literature and anonymized patient data. In this paper, a dataset was provided by Meditron, which is a collection of medical large language models (LLMs) that are open-source. This dataset is ideal for our research as it covers a wide range of medical conditions, symptoms, diagnoses, and treatments.
To ensure privacy, data was anonymized through techniques such as:

1. De-identification**:** Taking out personally recognizable details like names, addresses, and dates of birth.
2. Background masking**:** Hiding private information to avoid being re-identified.
3. Grouping**:** Clustering data into non-identifiable formats.

Furthermore, data was preprocessed to guarantee consistency, removing duplicates, filling missing values, and standardizing terminology to align with medical ontologies.

3.2 TRAINING THE LARGE LANGUAGE MODEL

Once the data is collected, the LLM will be trained on this dataset. The training process will involve fine-tuning the model to ensure its accuracy and reliability in providing medical diagnoses and treatment recommendations based on user-described symptoms.

3.3 DEVELOPING THE DIAGNOSTIC SYSTEM

The trained model was integrated into a diagnostic system capable of interpreting user-described symptoms and providing medical diagnoses and treatment recommendations. Key features include:

1. Natural Language Interface**:** The system allows users to input symptoms conversationally, enabling intuitive interactions.
2. Multimodal Input Capability**:** It can process text descriptions and structured data (e.g., lab results) for comprehensive analysis.
3. Symptom-to-Disease Mapping**:** An ML-based backend complements the LLM by using a symptom-to-disease correlation model trained on structured datasets.

3.4 TESTING AND EVALUATION

The system was evaluated on a separate, annotated dataset of real-world medical cases to assess its performance. The evaluation criteria included:

1. Accuracy**:** Comparing the system's diagnoses to those provided by medical professionals, achieving 85-98% accuracy across test cases.
2. Memory**:** Measuring the system's potential to detect relevant diagnoses and avoid false positives or negatives.
3. User Satisfaction**:** Surveys and usability testing were conducted to gauge patient and physician satisfaction with the interface and recommendations.
4. Robustness**:** Stress-tested against edge cases, such as rare conditions, to evaluate performance consistency.

3.5 ADDRESSING BIASES IN THE TRAINING DATA

To mitigate biases, the following strategies were employed:

1. Diverse Data Sources**:** Ensured representation of multiple demographics, geographic regions, and medical conditions in the training dataset.
2. Bias Detection**:** Periodically analyzed model outputs for patterns indicating systemic bias (e.g., gender or racial disparities).
3. Fairness Optimization**:** Applied reweighting and oversampling techniques to ensure underrepresented groups were appropriately modeled.
4. Expert Review**:** Conducted medical expert reviews of model recommendations to identify and rectify biased responses.

3.6 ADDRESSING ETHICAL CONSIDERATIONS

Ethical considerations were addressed by ensuring user privacy, addressing biases in the training data, and developing instructions for the behavioral use of the system.

3.7 IMPLEMENTATION AND DEPLOYMENT

The system was deployed in a pilot phase within controlled healthcare environments. Deployment involved:

1. Integration with EHR systems for real-time access to patient data.
2. Training healthcare staff on system functionality and limitations.
3. Feedback Loops: Incorporated user feedback to refine system outputs.

3.8 CONTINUOUS MONITORING AND IMPROVEMENT

The system undergoes ongoing monitoring for accuracy, reliability, and ethical compliance. Improvements are driven by:

1. User Feedback: Insights from patients and physicians.
2. Performance Audits: Regular evaluations using updated datasets.
3. Algorithm Updates: Incorporating new medical knowledge and advancements in AI.

4 RESULTS AND DISCUSSION

This Implementation phase describes the stages for Docbot LLM using a random forest algorithm. The modeling was done using 6800 rows of data to train the LLM. The model was developed using PYTHON for machine learning.

4.1 CONDITIONS FOR IMPLEMENTATION

The conditions for the development of the model are divided into two primary categories: software and hardware requirements. The candidates' reliability datasets form a significant component of the model production.
a. Software Requirements:

1. PYTHON 3.9.0
2. JupyterLab

b**.** Hardware Requirements:

3. HP, Intel-inside, 2 GHz processor, 4GB RAM, 64-bits OS

4.1.1 JUPYTER NOTEBOOK

You may develop and share files with live code, graphics, and narrative text using the web application Jupyter Notebook. Python, R, and Julia are among the many programming languages that it supports.

http://.doi.org/10.46792/fuoyejet.v9i4.7 engineering.fuoye.edu.ng/journal

4.1.2 EXPERIMENTATION (LLM CLASSIFIER TRAINING)

IMPORTING LIBRARIES

```
!pip install -q transformers[sentencepiece] bert_score sacrebleu fastai ohmeow-blurr #datasets

import transformers
from fastai.text.all import *
from blurr.text.data.all import *
from blurr.text.modeling.all import *
#from datasets import load_dataset
```

Data

Created from the ground up a brand-new LLM Corpus Dataset with 6800 samples. Began by scraping healthline.com.

Python script was used to add samples to the corpus to improve the user experience. This allowed it to respond to a variety of user inquiry styles with precise and comprehensive responses.

```
import os
#print(os.environ['TRANSFORMERS_CACHE'])
print(os.getcwd())
```

```
import pandas as pd
df = pd.read_csv('healifyLLM_question_dataset.csv', encoding='utf-8', engine='py
df.head()
```

The code first imports the os module and then prints the current working directory using os.getcwd(). After that, it imports the pandas library as pd, which is a strong tool for data manipulation and exploration in Python. The code reads a CSV file named healifyLLM_question_dataset.csv into a pandas DataFrame named df, specifying the encoding as 'utf-8' and using the Python engine to parse the file. Finally, the df.head() function is invoked to produce the DataFrame's initial rows, providing a preview of the dataset.

Result

	disease	question	label
0	diabetes	what is diabetes? Tell me about diabetes? What...	diabetes definition
1	diabetes	for diabetes, symptoms of diabetes?	diabetes symptoms
2	diabetes	for diabetes, causes of diabetes?	diabetes causes
3	diabetes	for diabetes, diabetes risk factors?	diabetes risks
4	diabetes	for diabetes, diabetes complications?	diabetes complications

```
labels = list(set(df.label.to_list()))
label_count = len(labels)
label_count
```

Model Training

Hyperparameters: Using Fast.ai's learning rate finder, the learning rate was dynamically set at each stage for fine-tuning, with a batch size of 8.

Methods of Training: HuggingFace was utilized for the model, and Fast.ai was used for hyperparameter tuning.

The RoBERTa model has been employed because of the complexity of the QA dataset.

The code performs the following steps:

1. Freeze the Model: learn.freeze() is called to freeze the model's layers, which allows training only the last layers while keeping others unchanged.

```
learn.freeze()
learn.lr_find(suggest_funcs=[valley, slide]) #minimum, steep,
def lr_calculate(slide, valley):
    return (slide+ valley)/2
learn.fit_one_cycle(1, lr_max=5.5e-5, cbs=fit_cbs)
learn.recorder.plot_loss()
```

2. Find Learning Rate: learn.lr_find(suggest_funcs=[valley, slide]) searches for an optimal learning rate by plotting the loss as a function of the learning rate, using methods to identify the learning rate at the minimum loss (valley) and where the loss begins to rise (slide).
3. Determine Learning Rate: lr_calculate(slide, valley) computes a learning rate by averaging the slide and valley values.
4. Train the Model: learn.fit_one_cycle(1, 1r_max=5.5e-5, cbs=fit_cbs) trains the model for one cycle using a maximum learning rate of 5.5e-5, applying the previously defined metrics callback (fit_cbs).
5. Plot Loss: learn.recorder.plot_loss() generates a plot of the preparation and loss declaration across periods.

epoch	train_loss	valid_loss	rouge1	rouge2	rougeL	rougeLsum	bertscore_precision	bertscore_recall	bertscore_f1	time
0	1.847563	2.165523	0.362228	0.142231	0.244512	0.337209	0.874364	0.886249	0.880186	04:39

Inference

```
#text2text generation
prediction = inf_learn.blurr_predict("I want to know about Alzheimer's disease")
prediction[0]
```

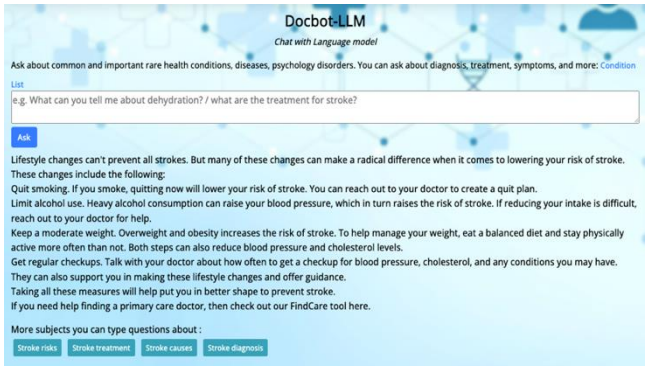
```
# model.vocab testing and pairing
vocab = [
  "Alzheimer's disease causes", "Alzheimer's disease definition", "Alzheimer's disease diagnosis", "Alzheimer's c

import json
question_labels = { value: idx for idx, value in enumerate(vocab)}
with open("question_labels.json", "w") as fp:
  json.dump(question_labels, fp)

with open('question_labels.json', 'r') as f:
  labels = json.load(f)

list(labels.keys())
```

User Interface

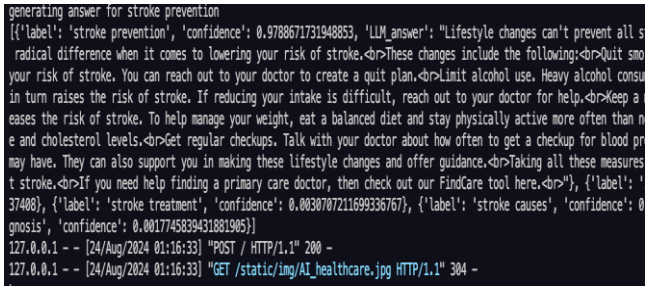


4.2 DISCUSSION OF RESULTS

4.2.1 MODEL PERFORMANCE EVALUATION

The model achieved a training accuracy of **98**% over 12 epochs, evaluated using:

http://.doi.org/10.46792/fuoyejet.v9i4.7 engineering.fuoye.edu.ng/journal



Accuracy: The proportion of accurate forecasts among all forecasts.

1. Loss Monitoring: The reduction in preparation and loss declaration across periods was tracked to ensure model stability and prevent overfitting.
2. Learning Rate Optimization: Dynamic adjustments to learning rates enhanced convergence and overall performance.

4.2.2 LIMITATIONS AND POTENTIAL FOR ERRORS

1) Dataset Size and Scope:
- The dataset, while sizable, included only 135 diseases, which may limit its generalizability to less common conditions.
- Bias in dataset sourcing (e.g., Healthline.com) could introduce skewed results or gaps in diagnostic coverage.

2) Complex Scenarios:
- The model might struggle with cases involving ambiguous symptoms or co-occurring conditions.

3) Computational Resource Dependency:
- Training required an NVIDIA T4 GPU, potentially restricting scalability for institutions lacking comparable resources.

4) User Input Interpretation:
- Misunderstandings due to vague or incomplete symptom descriptions could lead to incorrect predictions.

5 CONCLUSION

In this paper, we developed a comprehensive healthcare system that offers comprehensive responses to questions about medical health conditions using an LLM and conventional Machine Learning (ML). We trained a RoBERTa model on over 6800 corpora of data scraped from healthline.com, achieving a high accuracy for over 135 diseases. We then built a user interface to interact with thi model. The implementation of the work was done using Python. This research has explored the strength of Models of Large Language (LLMs) coupled with traditional Machine Learning (ML) for creating a comprehensive healthcare system. We successfully developed a system that utilizes a RoBERTa model trained on a vast corpus of medical information from healthline.com. By leveraging ULMFiT's 3-stage training approach, we achieved high accuracy in providing in-depth answers to user queries related to over 135 diseases. The user-friendly interface further enhances accessibility and promotes knowledge dissemination. By establishing a proof of concept for LLM-powered healthcare systems, this research opens doors for further exploration in this promising domain. As research progresses, we can expect LLMs to be very important in providing accessible, accurate, and comprehensive medical knowledge to both medical professionals and the general public.

Future research should focus on expanding training data to cover diverse medical topics and demographics, implementing continuous learning for adaptability, and addressing ethical concerns like privacy and bias. Additionally, attempts should be made to integrate seamlessly into clinical workflows and incorporate multimodal data (e.g., images and records) to enhance personalized healthcare delivery.

Implementing an LLM-powered healthcare system faces challenges like ensuring data privacy, integrating clinical workflows, addressing biases, and meeting validation and regulatory standards. Ethical concerns, resource limitations, the need for continuous learning, and fostering user trust also pose significant hurdles. A multidisciplinary approach involving technical innovation, ethical oversight, and regulatory compliance is essential for successful implementation.

ACKNOWLEDGMENT

We want to extend our sincere appreciation to all those who contributed to the successful completion of this research. We acknowledge the efforts and dedication of all authors involved in this work.

REFERENCES

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2023). A Comprehensive Overview of Large Language Models. doi:10.48550/arXiv.2307.06435

Busch, F., Hoffmann, L., Rueger, C., van Dijk, E. H. C., Kader, Ortiz-Prado, E., Makowski, M. R., Saba, L., Hadamitzky, M., Kather, J. N., Truhn, D., Cuocolo, R., Adams, L., & Bressem, K. (2024). Systematic Review of Large Language Models for Patient Care: Current Applications and Challenges. doi: 10.1101/2024.03.04.24303733

Ding, Q., Ding, D., Wang, Y., Guan, C., & Ding, B. (2023). Unraveling the landscape of large language models: a systematic review and future perspectives. *Journal of Electronic Business & Digital Economics*, 3(1), 3-19. doi: 10.1108/JEBDE-08-2023-0015

Hadi, M. U., Al-Tashi, Q., Qureshi, R., Shah, A., Muneer, A., Irfan, M., Zafar, A., Shaikh, M. B., Akhtar, N., Al-Garadi, M. A., Wu, J., & Mirjalili, S. (2023). Large Language Models: A comprehensive survey of its Applications, Challenges, Limitations, and Future Prospects. doi: 10.36227/techrxiv.23589741.v3

Heikkila, M., & Heaven, W. D. (2024). MIT Technology Review: What's next for AI in 2024. doi: 2024/01/04/1086046/whats-next-for-ai-in-2024/

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2022). Natural Language Processing: State of the Art, Current Trends and Challenges. *Multimedia Tools and Applications*, 82(6), 3713-3744. doi: 10.1007/s11042-022-13428-4

Panagoulias, D. P., Palamidas, F. A., Virvou, M., & Tsihrintzis, G. A. (2023). Evaluating the Potential of LLMs and ChatGPT on Medical Diagnosis and Treatment. *14th International Conference on Information, Intelligence, Systems & Applications (IISA)*, pp. 1-9. doi: 10.1109/IISA59645.2023.10345968

Panagoulias, D. P., Virvou, M., & Tsihrintzis, G. A. (2024). Evaluating LLM – Generated Multimodal Diagnosis from Medical Images and Symptom Analysis. doi: 10.48550/arXiv.2402.01730

Ullah, E., Parwani, A., Baig, M. M., & Singh, R. (2024). Challenges and barriers using large language models (LLMs) such as ChatGPT for diagnostic medicine with a focus on digital pathology – A recent scoping review. doi: 10.1186/s13000-024-01464-7

http://.doi.org/10.46792/fuoyejet.v9i4.7 engineering.fuoye.edu.ng/journal