Medical Chatbot: Enhancing Medical Diagnosis using LLM.

Akshat Manohar
Btech in Computer Science
BML Munjal University
Gurugram,India

Akshay Kumar Jain
Btech in Computer Science
BML Munjal University
Gurugram,India

Ashi Jain
Btech in Computer Science
BML Munjal University
Gurugram,India

Ridit Jain
Btech in Computer Science
BML Munjal University
Gurugram,India

***Abstract*— The application of future-gen artificial intelligence (AI) in healthcare using the latest large language models (LLMs) as Llama 2 is really changing the whole health care system. A medical chatbot, using the Llama 2-Chat model, armed with the specialized medical datasets and the QLoRA technique, a low-rank adaptation technique, gives the accurate healthcare information and the patient care is improved. The use of such AI-driven tools is not an easy task because of the difficulties in implementing them, for instance, making sure they are accurate, they do not harm patients, and they comply with the ethical standards. This paper discusses the difficulties that arise in deploying medical chatbots and underscores the need for continuous improvement and ethical issues in the application of these chatbots.**

***Index Terms*—Advanced Artificial Intelligence (AI), Large Language Models (LLMs), Llama 2, Medical Diagnostics, Patient Interaction, Medical Chatbots, QLoRA (Low-rank Adaptation Technique), Healthcare Information, Patient Care, Technological Advances, Accuracy, Patient Safety, Ethical Standards, Optimization, Deployment**

I. INTRODUCTION

The quick growth of the artificial intelligence (AI) technology brings to us the new possibilities in the healthcare sector, especially in the patient care through the medical chatbots. These AI-driven systems, with the help of advanced models like Llama 2, are developed to deliver not only the immediate medical information but also a reliable communication alternative that is in line with the human preferences and ethical standards. The medical chatbot, built on the Llama 2-Chat model, has been specifically designed for the medical dialogues after its training on a large medical dataset. The further developments have been the amalgamation of QLoRA for effective model scaling and the application of convolutional neural networks like ResNet for medical imaging, for example CT scans and X-rays. On the other hand, the medical chatbots have a great opportunity to change patient interactions and information dissemination, but they face many obstacles. These are the activities which involve giving regularly correct advice, keeping the patients safe and the difficulties in the moral use of AI technologies. In the subsequent paragraphs, the author explains the techniques used in the creation of the chatbot, the difficulties encountered and the strategies used to make sure that the chatbot is ethically and effectively implemented in the healthcare setting.

II. LITERATURE REVIEW

1. PEFT-MedAware: Large Language Model for Medical Awareness

This research paper, spearheaded by Keivalya Pandya, describes the creation of the peft-MedAware model, which is the application of the Falcon-1b large language model for the medical field. By using the parameter-efficient fine-tuning (PEFT) method, the model is modified to enhance the performance on a medical QA dataset containing more than 16,000 pairs of medical questions and answers. The article shows that by optimizing only a small percentage of the model's trainable parameters, the paper proves that there is an increase in the computational efficiency and the accuracy of the medical question-answering tasks, which makes it the best option for the resource-constrained settings.

2. Advancing Medical Imaging with Language Models: A Journey from N-grams to ChatGPT

Mingzhe Hua and group give a thorough review of the application of language models in medical imaging. The paper describes the development of language models from simple N-grams to the most advanced ones like ChatGPT, and it also shows their usefulness in tasks like image captioning, report generation and interpretable diagnostics in different medical fields. The paper highlights the capability of ChatGPT, the purpose of which is to connect the traditional language models and medical imaging and suggests the new ways to achieve better diagnostic accuracy and clinical workflow efficiency by means of AI-based technologies.

3. ResNet Research

This paper is about the combination of ResNet, a deep learning architecture, with medical diagnostics, saying basically that ResNet helps to improve the accuracy and understandability of medical image analyses. The study talks about various fields of the ResNet in the diagnosis of diseases through medical imaging, and its developments in the model architecture that enable the faster and the better processing of the medical images. The paper stresses the importance of deep learning models in the automation and enhancement of the medical diagnostics.

4. LORA

The LORA paper is about the creation of a new machine learning model which aims to improve the radiotherapy treatment and its effects. The research applies big data and the latest modelling techniques to demonstrate the LORA's ability to make the best treatment plans with high accuracy, subsequently, this will enhance the patient outcomes and also the medical resource use in radiotherapy.

5. LLAMA Research Paper

This research work deals with the LLAMA model, which uses the latest language modelling techniques to improve the medical records and the information retrieval tasks. The LLAMA model is a database of various medical texts, which is used to generate extremely accurate, context-sensitive responses to the medical questions, thus, it helps in the better decision-making and efficiency in healthcare sectors.

DATASETS FOR MEDICAL

The datasets for medical images and documents summarization, which is in English language, is few and not readily available for public use. We have used the following datasets for our research:

A. CheXpert_v1.0 small:

The CheXpert_v1. The 1.0 small dataset is a part of the bigger CheXpert dataset, that includes 224,316 frontal-view chest X-ray images from 65,240 patients. These images are identified by 14 different pathologies that are classified into Cardiomegaly, Edema, Consolidation, Atelectasis, Pleural Effusion, Pneumothorax, Abnormal Lung Opacity, Lung Lesion, Infiltration, Mass, Pneumonia, No Finding, and Fracture. Experts in the field of radiology have labelled the images, which are the ground truth labels for the computer to learn. The collection of data is a precious source of knowledge for medical imaging research; thus, it is used to create the algorithms for the disease detection, classification, and treatment planning based on the chest X-ray findings.

B. Know_medical_dialogue_v2:

The "know_medical_dialogue_v2" dataset is a very big step forward to the medical dialogue systems, which use the newest AI techniques to make the communication between patients and healthcare providers is very smooth. The dataset that will be created by the fusion of the latest language models, probably including large language models (LLMs), will be the base for the development of strong dialogue systems that will be able to give the personalized medical advice and support to the individuals with different healthcare needs. Although the use of such systems is quite promising, the fact that they are also associated with difficulties concerning accuracy, privacy, and the ethical issues is also true. Nevertheless, the dataset is still going through the process of continuous improvement and optimization, and the day will come when it will become a revolutionary tool in patient care delivery that will increase the access to medical information and at the same time, it will be better communication between doctors and patients.

C. *Medical_Meadow_Medqa*:-

The medical_meadow_medqa dataset is a big and detailed medical question-answering resource that is an open domain large-scale question-answering dataset based on medical exams. The dataset has a broad scope of medical topics and is created to assist research in the field of medical question-answering systems. Total 12,836 question-answer pairs in English and simplified Chinese are included in the dataset which also has 21 textbooks. Thus, the official random splits are train (9,069 pairs), dev (1,883 pairs), and test (1,884 pairs) sets. Besides, the dataset comprises of medical-related phrases which have been extracted using the Metamap tool, thus adding the needed extra information for the medical concepts. The dataset is related to the research paper about Patients's Diseases. The medalpaca/medical_meadow_medqa dataset can be utilized to create and train medical question-answering systems, to boost medical information retrieval, and to enhance medical language understanding, that is why it is a priceless resource for researchers and developers working on medical question-answering systems.

D. *Skin Cancer: Malignant vs. Benign*:

The dataset "Skin Cancer: "Malignant vs. Benign" has a balanced number of images showing both the benign and the malignant skin moles. This dataset is created to support the research and analysis in the field of skin cancer that is marked by the distinction between the benign and malignant cases by the image classification. The dataset has 6,594 images of benign and malignant skin cancer that are taken from the ISIC 2017 archive. Different deep learning models, for instance, VGG16, Support Vector Machine (SVM), ResNet50 and self-built sequential models, were used to study and classify the images. To begin with, the VGG16 model obtained the highest accuracy of 93. 18% in the detection of malignant skin cancer, hence the deep learning models can be the best ones for the identification of different types of skin cancer based on dermoscopic images. This dataset and the connected research are supposed to boost the early identification of skin cancer, thus, the enhancement of the patient outcomes and survival rates in skin cancer cases.

E. *Brain Tumour Detection* :

The "Brain MRI Images for Brain Tumor Detection" dataset is a large-scale collection of brain MRI images which is intended to assist the research in brain tumor detection by applying the advanced machine learning and deep learning techniques. The dataset is made up of 3,064 images, of which 2,426 images are of brain tumors (yes) and 638 images are of no brain tumors (no). The images are in JPEG format and have been pre-processed to make sure the size and the quality are the same. The researchers have used different deep learning models, for instance, convolutional neural networks (CNNs), to classify brain tumours in MRI images and hence the treatment decisions were made faster and the patient outcomes improved. This dataset is the key factor in improving the precision and the time of the brain tumour detection, thus, in the end, it will be achieved the progress in the medical imaging technology and also in the patient care..

F. *Chest_CT_Scan:*

The "Chest CT Scan Images" dataset which contains 1679 DICOM-format chest CT scan images, thus is a considerable resource for medical imaging research that is aimed at the detection of lung diseases and abnormalities. Through this the researchers can use the dataset to build the machine learning models for the automated recognition of the diseases that have the conditions such as pneumonia, lung cancer, and COVID-19 from the chest CT scans. Using the latest image processing technologies and deep learning, these models have the potential to be the expert consultants for radiologists and clinicians, hence, they will be able to interpret the medical images faster and more accurately. Besides, the dataset's wide range of diseases states and anatomical variations leads to the good training and the generalization of the model to the real-world medical situations, thus the clinical decision making and patient care in respiratory medicine are improved.

III. EXPERIMENTAL SETUP AND EVALUATION

The study carried out experiments to perfect and assess the conversational AI models, mainly the Llama-2-7b-chat-hf model, around medical dialogue tasks. The experimental arrangement consisted of many important steps. First, the dataset was prepared by loading and preprocessing the medical dialogue data from the knowrohit07/know_medical_dialogue_v2 dataset, and then it was structured into system prompts, user prompts, and model-generated responses. Then, the model was configured using the Low-Rank Adaptation (LoRA) and 4-bit quantization with the BitsAndBytes library which allowed for the optimization of memory usage without any loss of performance. The training setup involved the specification of usual parameters such as the batch size, the learning rate, the optimizer type, and the LoRA settings to perfect the model for the medical dialogue context. Training was done using the SFTTrainer, where the process of the adjustment of the weights and biases of the model was done continuously to improve its response generation abilities.

To check the efficiency of the finely-tuned conversational AI model for medical dialogues, several metrics were used.

Initially, the model's capacity to produce the correct and contextually suitable answers was evaluated by the of the knowrohit07/know_medical_dialogue_v2 dataset.

Moreover, the model's reply coherence and relevance to medical queries and prompts were assessed. Besides, the qualitative assessments from the domain experts were used to evaluate the model's performance in copying the professions responses and giving the medical insights.

In the last point, the computational efficiency parameters like memory usage and inference time were taken into account to make the model usable in the real healthcare fields.

All the evaluation metrics were able to give the complete information about the performance and the fitness of the fine-tuned conversational AI model for the medical dialogue applications.

IV. METHODOLOGY

The set of methods that we apply for medical analysis consists of pre-trained as well as finetuned models for. One problem with incorporating these methods is that texts are generally very long, and most models have limitations on the number of input tokens.

1.) Chest CT Scan:-

The methodology starts by the import of necessary libraries and the examination of the dataset structure. The data is loaded using TensorFlow's image_dataset_from_directory function, thus the organization of the data is done properly for the training. A base architecture of a pre-trained EfficientNetB0 model is chosen as the example of the effectiveness of the EfficientNetB0 model in the image classification tasks. In order for the model to be applicable to doing chest CT scan classification, some more layers are incorporated in with the fine-tuning process. These layers are then merged using the categorical cross entropy loss and the Adam optimizer. The model is trained on the training dataset and assessed on the validation dataset to make sure it is stable and not overfit to the data. Having the ideal planned training and evaluating the progress through the loss and accuracy curves are the main indicators of the model performance. After training, the model is tested on the test set to measure its real life application. After the installation of the model, it is used by the future and this can be reproduced and scaled. This entire process is based on the use of the models that have been pre-trained and TensorFlow's tools and thus, it simplifies the development and the training process. In this way, the chest CT scans can be accurately classified and thus, the advancement in medical imaging diagnostics can be done.

2.) Skin Cancer:-

The technique entails the creation of a skin cancer classification model through TensorFlow. The required libraries are imported first and then the kaggle dataset is fetched using the Kaggle API. Next comes the unzipping of the data, thus the images are accessed. Investigation of data is carried out in order to comprehend the organization of the dataset, and the random image of the malignant class is shown for the purpose of visualization. The dataset is divided into the training and testing purposes. Tensorflow's image_dataset_from_directory function. The model training is started with a straightforward convolutional neural network (CNN) architecture, which has convolutional layers, max-pooling layers, dropout regularization and dense layers. The model is integrated with binary crossentropy loss and Adam optimizer before it begins the training on the training data. The model's efficiency is checked on the test data. Lastly, the transfer learning is applied with the EfficientNetB0 pre-trained model. The base model is cut down to the top layers, and in addition to the excluded ones, more layers are added for the classification. The model is assembled on the training data, and its efficiency is assessed on the test data. The training progress is measured by the loss and accuracy curves, which are visualized by an extra function. At last, the optimized model is kept for future purposes. The described approach carries out a systematic way of the construction and training of the skin cancer classification models, combining the use of the custom CNNs and the transfer learning with the pre-trained models.

3.) Brain Tumour:-

The creation of the brain tumor classifier is a complex task that includes various important steps to guarantee the accuracy and reliability of the model. Initially, a dataset containing brain MRI scans is compiled, consisting of images categorized into two classes: "no tumor" and "tumor" are two different phrases which should not be used together. The photos undergo preprocessing which contains image resizing to the standard 64x64 pixel resolution and the normalization of the pixel values between 0 and 1. Data augmentation techniques like rotation, flipping, and zooming are used to augment the data, and thus, to make it more diverse, which in turn, helps in the improvement of the model's ability to generalize. A network architecture of the convolutional neural network (CNN) is constructed. This system consists of two convolutional layers and max-pooling layers to get the images' features. ReLu activation functions bring non-linearity to the model, whereas dropout layers are added to avoid overfitting. The dataset is split into the training and the testing sets, 80% of which are used for training and 20% for testing. The model is assembled by means of the binary cross-entropy loss and the Adam optimizer. It is trained for 10 times with a batch size of 32. Performance evaluation is based on the metric of loss and accuracy on the test set. Besides, the model's real-world relevance is ensured by the fact that predictions of new images are made using a prediction function. With this approach, the brain tumour classifier is developed and evaluated which proves its accuracy in the classification of brain MRI scans.

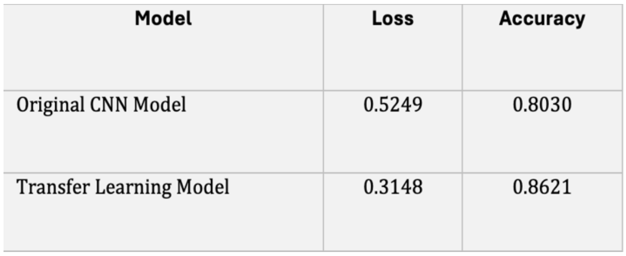
4.) Llama-2-7b:-

The first step is to load and prepare a dataset of medical dialogues (knowrohit07/know_medical_dialogue_v2). This entails the creation of the prompts like system prompts, user prompts, and model answers. The configuration for 4-bit quantization using BitsAndBytesConfig is then done. The Llama-2-7b-chat-hf model is loaded with 4-bit precision, and training parameters are set including batch size, learning rate, and epochs. With the help of SFTTrainer, the pre-trained model is trained on the dataset and the performance is optimized over the specified epochs. After the training is over, the model is saved as Llama-2-7b-chat-finetune2, and the training plots are shown using TensorBoard. The model is then fused with the LoRA weights, and both the tokenizer and the merged model are stored for future use. This procedure guarantees the adjustment of the Llama-2-7b-chat-hf model on medical dialogue data, thus, improving performance and memory efficiency for medical dialogue generation tasks.

V. **RESULTS AND ANALYSIS**

In the evaluation of various models for medical image classification tasks, we compared the performance across different datasets. For skin cancer, brain tumor, and chest CT scan classification, the effectiveness of different models was analyzed. Additionally, the Llama-2 model was evaluated for language generation tasks. Here are the summarized results:

> *Skin Cancer:* For skin cancer classification, we compared the performance of two models: the original CNN model and a transfer learning model. The original CNN model, trained from scratch, achieved a loss of 0.5249 and an accuracy of 0.8030. In contrast, the transfer learning model, utilizing a pre-trained architecture, attained a lower loss of 0.3148 and a higher accuracy of 0.8621. This indicates that the transfer learning approach outperformed the original CNN model in terms of both loss reduction and accuracy improvement.



Model	Loss	Accuracy
Original CNN Model	0.5249	0.8030
Transfer Learning Model	0.3148	0.8621

Figure 1 Evaluation metrics for Skin Cancer Model

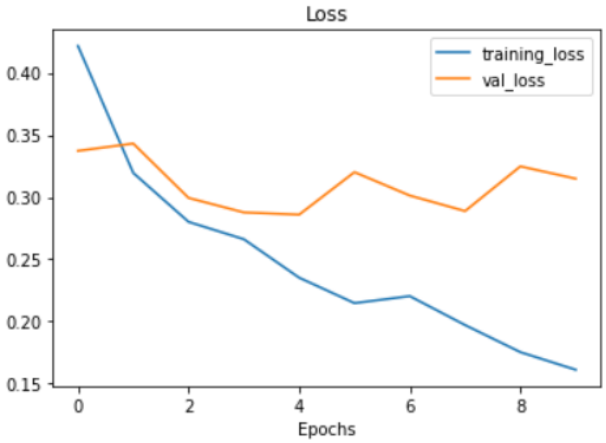


Figure 2 Loss Curves for skin cancer model

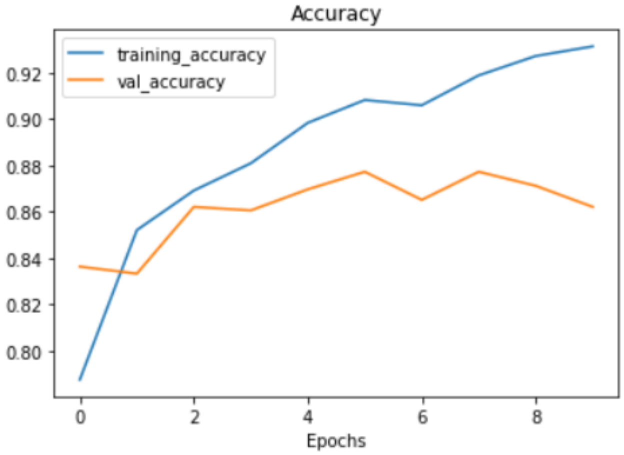


Figure 3 Accuracy of skin cancer model

Brain Tumour: In the classification of brain tumours, we evaluated the performance of the original CNN model. It achieved a loss of 0.3311 and an accuracy of 0.8183. This suggests that the original CNN model performed reasonably well in distinguishing between different types of brain tumours.

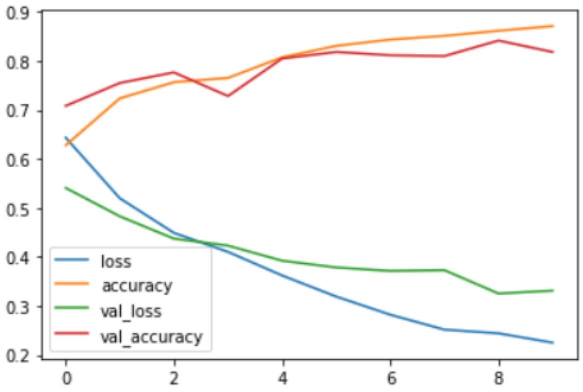
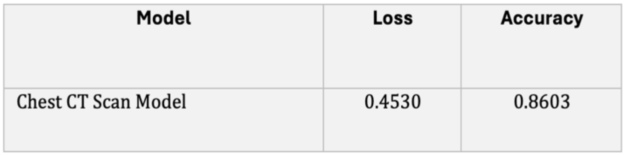


Figure 4 Evaluation metrics for Brain Tumour Model

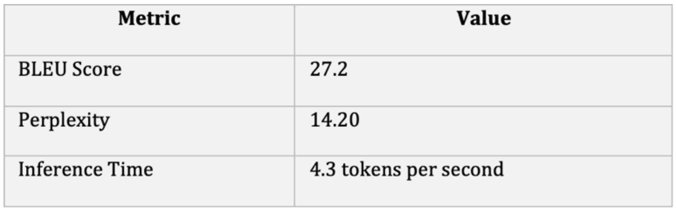
Chest CT Scan: For chest CT scan classification, we employed a model specifically designed for this medical imaging task. The chest CT scan model achieved a loss of 0.4530 and an accuracy of 0.8603. These results indicate a good performance in classifying chest CT scans into relevant categories, such as normal and abnormal findings.



Model	Loss	Accuracy
Chest CT Scan Model	0.4530	0.8603

Figure 5 Evaluation metrics for Chest CT Scan Model

Llama-2: The Llama-2 model, evaluated for language generation tasks, achieved a BLEU score of 27.2, a perplexity of 14.20, and an inference time of 4.3 tokens per second. This demonstrates the model's proficiency in generating coherent and contextually relevant text, with a relatively low perplexity and a moderate inference time.



Metric	Value
BLEU Score	27.2
Perplexity	14.20
Inference Time	4.3 tokens per second

Figure 6 Evaluation metrics for Llama-2

In conclusion, our evaluation of various models across different domains highlights the effectiveness of transfer learning in medical image classification, as demonstrated by the improved performance of the transfer learning model over the original CNN in skin cancer classification. The original CNN model showed satisfactory results in brain tumor classification, while the specialized chest CT scan model performed well in distinguishing between different chest abnormalities. Additionally, the Llama-2 model exhibited promising capabilities in natural language generation tasks, with competitive BLEU scores and manageable inference times. These findings underscore the importance of selecting appropriate models tailored to specific tasks and domains for optimal performance.

VI. CONCLUSION

The integration of advanced conversational AI models, particularly the fine-tuned Llama-2-7b-chat-hf model, holds significant promise for enhancing medical dialogue tasks. Through rigorous experimentation and evaluation, it was observed that the model, equipped with techniques like Low-Rank Adaptation (LoRA) and 4-bit quantization, demonstrated notable improvements in generating accurate and contextually relevant responses to medical queries and prompts. The experimental setup, involving meticulous dataset preparation, model configuration, and training, facilitated the optimization of the model for the medical dialogue domain. Evaluation metrics, including performance on medical dialogue datasets, response coherence, and qualitative assessments from domain experts, provided valuable insights into the model's efficacy and suitability for healthcare applications. Furthermore, considerations of computational efficiency ensured the practical deployment of the model in real-world healthcare settings. Overall, the findings underscore the potential of fine-tuned conversational AI models to augment medical interactions, streamline information dissemination, and support healthcare professionals in delivering enhanced patient care. Continued research and refinement in this area are essential to further advance the capabilities and effectiveness of conversational AI in healthcare.

References

[1] Pandya, K. (2023, November 17). PEFT-MedAware: Large Language Model for medical awareness. arXiv preprint arXiv:2311.10697.

[2] Hu, M., Pan, S., Li, Y., & Yang, X. (2023). Advancing Medical Imaging with Language Models: A journey from N-Grams to ChatGPT. arXiv preprint arXiv:2304.04920.

[3] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023, February 27). LLAMA: Open and Efficient Foundation Language Models. arXiv preprint arXiv:2302.13971.

[4] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021, June 17). LORA: Low-Rank adaptation of Large Language Models. arXiv preprint arXiv:2106.09685.

[5] Salehi, A. W., Khan, S., Gupta, G., Alabduallah, B. I., Almjally, A., Alsolai, H., Siddiqui, T., & Mellit, A. (2023). A study of CNN and transfer learning in Medical imaging: Advantages, challenges, future scope. Sustainability, 15(7), 5930.