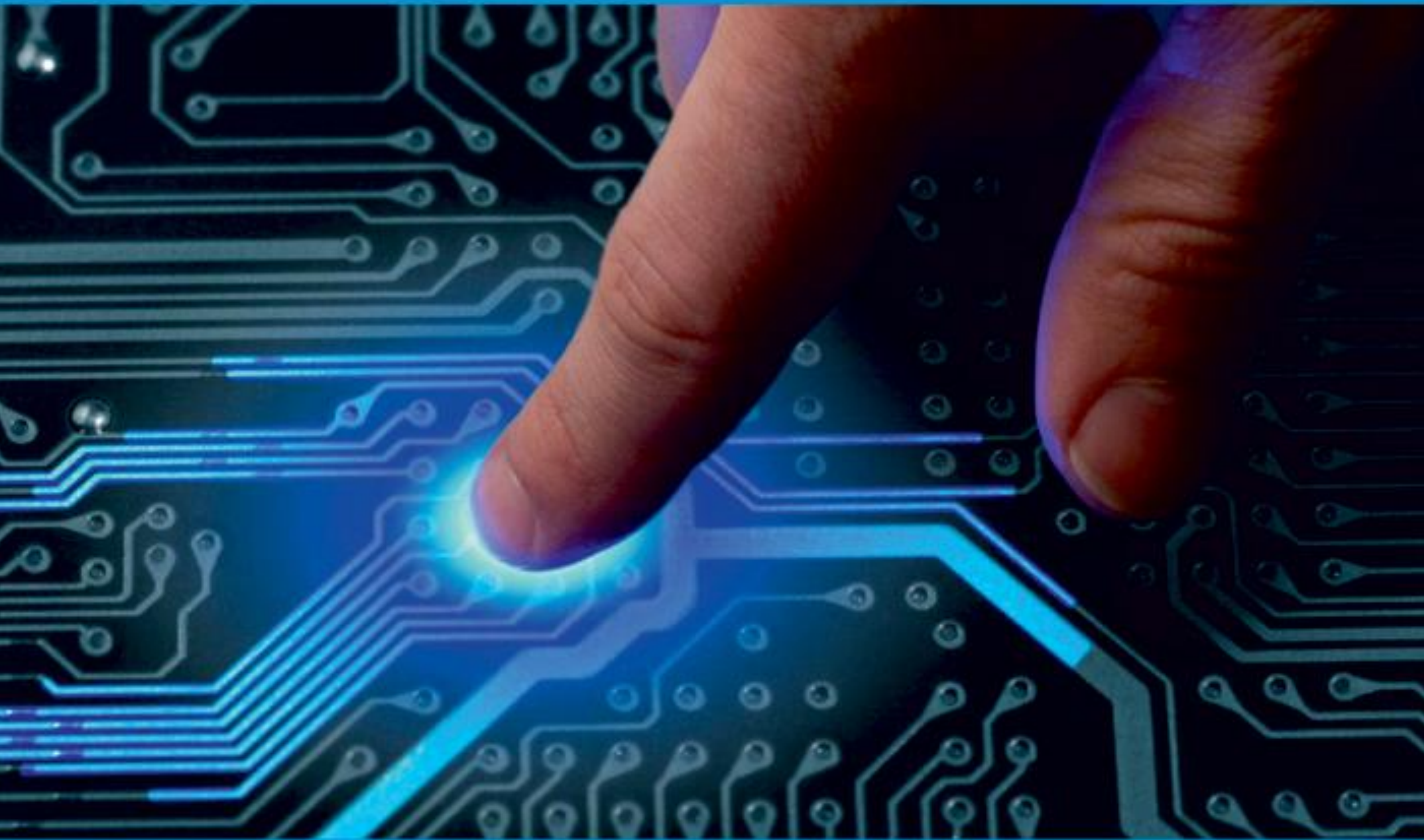
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 5, May 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205258 |

ChatBot For Healthcare System Using Artificial Intelligence

Siva E, Thanigaivelan A, Mohammed Ismail A and Selvaganesh N

UG Student, Department of IT, Sri Venkateswara College of Engineering, Pennalur, Sriperumbudur, India

UG Student, Department of IT, Sri Venkateswara College of Engineering, Pennalur, Sriperumbudur, India

UG Student, Department of IT, Sri Venkateswara College of Engineering, Pennalur, Sriperumbudur, India

Assistant Professor, Department of IT Sri Venkateswara College of Engineering, Pennalur, Sriperumbudur, India

ABSTRACT: Applying Large Language Models (LLMs) to the field of medicine has been a growing trend, particularly with the advent of advanced language models such as ChatGPT, LLaMa etc. The system works by adapting and refining the large language model Meta-AI (LLaMA) using a large dataset of 100k patient-doctor dialogues sourced from a widely used online medical consultation platform, HealthcareMagic. In addition to the model refinement an external knowledge base with the name, symptoms, reason and medications were collected from iCliniq medical platform. Low Rank Adaptation fine tuning method is used for training the LLM. Training the model with real-world patient-doctor interactions significantly improved the model's ability to understand patient needs and provide informed advice and an accuracy of 96% is achieved.

KEYWORDS: Large Language Model Meta AI (LLaMa), Low Rank Adaptation (LoRa).

I. INTRODUCTION

The development of instruction-following Large Language Models (LLMs), such as ChatGPT, has gained significant attention due to their remarkable success in instruction understanding and human-like response generation. These auto-regressive LLMs are pre-trained on web-scale natural language by predicting the next token and then fine-tuned to follow large-scale human instructions. These models show robust performance on a wide range of Natural Language Processing (NLP) tasks and can generalize to unseen tasks, demonstrating their potential as unified solutions to various problems in natural language understanding, text generation, and conversational artificial intelligence. However, the exploration of such general-domain LLMs in the medical domain remains relatively scarce, despite their great potential in revolutionizing medical communication and decision-making. In general, these common-domain models were not trained to capture the medical-domain knowledge specifically or in detail, resulting in models that often provide incorrect medical responses. By fine-tuning large linguistic dialogue models on data from real-world patient-physician conversations, these models' ability in understanding patients' inquiries and needs can be significantly improved. In addition, to further enhance the models' credibility, offline sources like medical-domain databases can be incorporated into the models to retrieve real-time information to facilitate answering medical questions.

II. RELATED WORK

In [3] Authors demonstrated a comprehensive text-based Bangla healthcare chatbot named "Disha". Six different machine learning algorithms have been applied to classify diseases: Decision Tree (DT), Random Forest (RF), Multinomial Naive Bayes (MNB), Support Vector Machine (SVM), AdaBoost, and K Nearest Neighbor (KNN). For vectorization of the Bangla text, Term Frequency-Inverse Document Frequency (TF-IDF) is used, while the Cosine Similarity Measure is employed to determine the similarity between texts.

In [5] Authors introduced a chatbot model developed through ensemble learning techniques. Leveraging Naïve Bayes (NB), Decision Tree (DT), K Nearest Neighbor (KNN), Random Forest (RF), Logistic Regression, and Gradient Booster (GB) machine learning algorithms, an ensemble model is constructed. These models are trained on respective datasets general health and the Pima Indian diabetes dataset to furnish diagnostic decisions to the NLU engine. The developed ensemble model achieves an accuracy of 91% which is partially greater than the accuracies of the other individual algorithms.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205258 |

In [6] Authors proposed the methods of teaching chatbots to process natural language text. It discusses about processing natural language using Recurrent Neural Network (RNN). The sequence to sequence long short-term memory cell neural network (LSTM) is used to train the model. The work proposed here discusses the need of a chatbot on ecommerce site to enhance user experience with 0.87 accuracy. The chatbot gives information about the products on an e-commerce site which is useful for a customer to buy exactly what they want.

In [7] Authors designed an automated self-learning system to provide conversational healthcare for a personalized proactive experience. This system is developed in conjunction with a contactless monitoring device, which uses vision-based real-time monitoring to track vital signs such as oxygen levels, heart rate, and respiration rate. The integration of this technology allows patients to monitor these vital signs without physical contact, enhancing convenience and comfort.

III. PROPOSED ALGORITHM

The proposed system addresses the limitations observed in the medical knowledge of prevalent large language models by creating a specialized language model with enhanced accuracy in medical advice. Here’s an outline of the system:

Data Collection and Preprocessing: Authentic patient-doctor conversations were gathered, collecting around 100k such interactions from the online medical consultation website, HealthCare Magic. Specifically, the conversations that were too short were filtered out automatically, most of which did not answer anything of practical significance. A Comprehensive medical knowledge base was built offline, acting as an external brain for our system. This database contains information on various diseases, including their associated symptoms.

Text Pre-Processing: Llama Tokenizer module is used for input text processing. It is a class from the Transformers library used for preparing text data specifically for Meta AI’s LLaMA model. LlamaTokenizer utilizes a technique called Byte-Pair Encoding (BPE). BPE is a method for segmenting text into subword units, which are more efficient for large language models to process compared to whole words. The tokenizer starts by splitting the input text into individual bytes. Next, the tokenizer analyzes this byte sequence and identifies frequently occurring pairs of bytes.

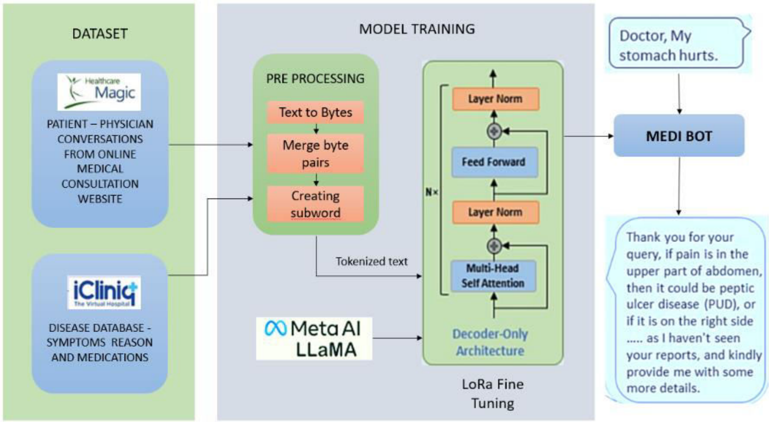
ChatBot development using LoRa: Low Rank Adaptation (LoRa) method is used for training the chatbot model in the proposed system. LoRa is a technique that significantly speeds up the fine-tuning process of large language models while consuming less memory. LoRa reduces the number of trainable parameters by decomposing weight matrices into low-rank approximations. In conventional LLM fine-tuning, all the weight values are updated and this process is not always feasible since the number of trainable parameters can reach up to hundreds of billions. Instead, with LoRA, the initial weights are frozen, and only the weight update ΔW receives gradient updates.

Model Training: The proposed Chatbot model was developed using Meta’s publicly accessible LLaMA-7B model, which uses Transformers with the structure of the decoder only. Despite its relatively modest 7 billion parameters, the LLaMA model exhibits comparable performance to the much larger GPT-3 model (with 175 billion parameters) across several NLP benchmarks. This performance enhancement was achieved by diversifying the training data rather than increasing network parameters. The model was first fine-tuned with Alpaca’s data to acquire basic conversation skills, followed by further refinement on HealthCareMagic-100k using 6 * A100 GPUs for three hours. The training process followed these hyperparameters: total batch size of 192, learning rate of 2 x 10-5, 3 epochs, maximum sequence length of 512 tokens, and a warmup ratio of 0.03, with no weight decay.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205258 |



IV. PSEUDO CODE

```
Text Preprocessing
def tokenize(prompt, add_eos_token=True):
    result = tokenizer(
        prompt,
        truncation=True,
        max_length=cutoff_len,
        padding=False,
        return_tensors=None,
    )
    if (
        result["input_ids"][-1] != tokenizer.eos_token_id
        and len(result["input_ids"]) < cutoff_len
        and add_eos_token
    ):
        result["input_ids"].append(tokenizer.eos_token_id)
        result["attention_mask"].append(1)

    result["labels"] = result["input_ids"].copy()
    return result

LoRa Configuration
config = LoraConfig(
    r=lora_r,
    lora_alpha=lora_alpha,
    target_modules=lora_target_modules,
    lora_dropout=lora_dropout,
    bias="none",
    task_type="CAUSAL_LM",
  )
model = get_peft_model(model, config)

Trainer Configuration
trainer = transformers.Trainer(
    model=model,
    train_dataset=train_data,
    eval_dataset=val_data,
    args=transformers.TrainingArguments(
        per_device_train_batch_size=micro_batch_size,
        gradient_accumulation_steps=gradient_accumulation_steps,
        num_train_epochs=num_epochs,
        learning_rate=learning_rate,
```

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

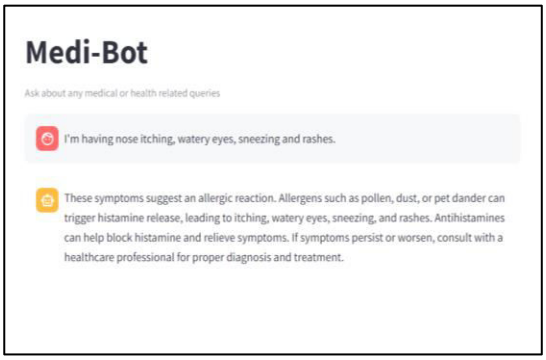
| DOI: 10.15680/IJIRCCE.2024.1205258 |

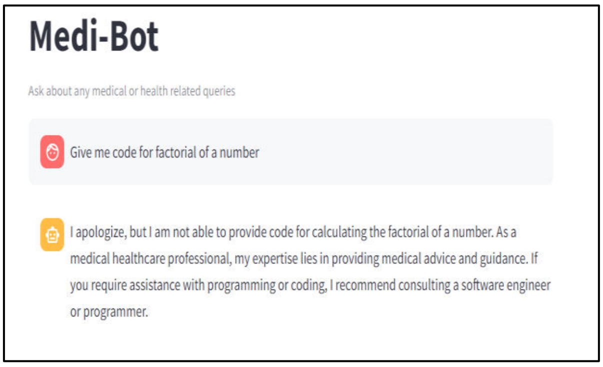
```
        fp16=True
    ),
    data_collator=transformers.DataCollatorForSeq2Seq(
        tokenizer, pad_to_multiple_of=8, return_tensors="pt", padding=True
    ),
)
```

V. SIMULATION RESULTS

The Medi bot is trained to specialize exclusively in addressing medical and health-related inquiries. Its responses are trained to provide fundamental medical guidance, primarily recommending basic-level medications. However, should symptoms escalate or pose a more serious concern, the bot prioritizes user safety and advises prompt consultation with a qualified medical professional. This approach ensures users receive reliable support for their health inquiries while emphasizing the importance of seeking professional medical advice when necessary.

The trained model is stored in a pretrained folder. It is then retrieved in the UI configuration file, where the user queries are taken as input. Using the trained model, the answer text is generated and displayed as output using streamlit framework.





VI. CONCLUSION AND FUTURE WORK

With adequate training and online/offline supervision, Medi-Bot can potentially improve accuracy and efficiency in medical diagnosis and reduce the workload for medical professionals. It may also increase access to high-quality medical consultations, especially for patients in underserved regions with limited medical resources. The further developments and applications of Medi-bot may eventually help to improve patient outcomes and advance medical research. To ensure that Medi-Bot remains up-to-date with the latest medical terms and advancements, which may not be included in the initial training dataset, the model is equipped with the capability to autonomously retrieve information from external knowledge bases.

In future developments, the chatbots can be programmed to communicate in multiple languages ensuring that everyone has access to healthcare information and support regardless of their language proficiency. It can also be developed to access a patient’s medical history and tailor educational materials accordingly. This personalized approach empowers patients to better understand their conditions, treatment options and preventive measures.

REFERENCES

1. Lal, Mily, and S. Neduncheliyan. "An optimal deep feature–based AI chat conversation system for smart medical application." *Personal and Ubiquitous Computing* 27.4 (2023): 1483-1494.
2. Athota, Lekha, et al. "Chatbot for healthcare system using artificial intelligence." 2020 8th International conference on reliability, infocom technologies and optimization (trends and future directions)(ICRITO). IEEE, 2020.
3. M. M. Rahman, R. Amin, M. N. Khan Liton and N. Hossain, "Disha: An Implementation of Machine Learning Based Bangla Healthcare Chatbot," 2019 22nd International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205258 |

4. Kumar, Mr U. Bhargav, Smt M. Prashanthi, and Smt D. Madhuri. "Health Care Chat Bot By using NLP, Decision Tree and SVM.
5. Sarma, Manash & Chatterjee, Subarna & Mohanty, Samahit & Puravankara, Rajesh & Bali, Manish. (2019). Diabot: A Predictive Medical Chatbot using Ensemble Learning.
6. KP, Roshan & M a, Nahala & Paulose, Christy & Cruz, Fremin & T M, Vivek. (2024). Medibot: A Medical Assistant Chatbot.
7. Wan Zaki, Wan Muhamad Asyraf & Md Shakhih, Muhammad Faiz & Ramlee, Muhammad Hanif & Abdul Wahab, Asnida. (2019). Smart Medical Chatbot with Integrated Contactless Vital Sign Monitor. Journal of Physics: Conference Series. 1372. 012025.
8. Madotto, Andrea & Lin, Zhaojiang & Bang, Yejin & Fung, Pascale. (2020). The Adapter-Bot: All-In-One Controllable Conversational Model.
9. Li, Jiwei & Galley, Michel & Brockett, Chris & Gao, Jianfeng & Dolan, Bill. (2016). A Persona-Based Neural Conversation Model.
10. G. K. Vamsi, A. Rasool and G. Hajela, "Chatbot: A Deep Neural Network Based Human to Machine Conversation Model," 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India.
11. Li Y, Li Z, Zhang K, et al. (June 24, 2023) ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge. Cureus 15(6): e40895.
12. N. Rosruen and T. Samanchuen, "Chatbot Utilization for Medical Consultant System," 2018 3rd Technology Innovation Management and Engineering Science International Conference (TIMES-iCON), Bangkok, Thailand.
13. Kazi, Hameedullah, Bhavani Shankar Chowdhry, and Zeesha Memon. "MedChatBot: an UMLS based chatbot for medical students." (2012).
14. Julius Odede and Ingo Frommholz. 2024. JayBot -- Aiding University Students and Admission with an LLM-based Chatbot. In Proceedings of the 2024 Conference on Human Information Interaction and Retrieval (CHIIR '24). Association for Computing Machinery, New York, NY, USA.
15. Rarhi, Krishnendu and Bhattacharya, Abhisek and Mishra, Abhishek and Mandal, Krishnasis, Automated Medical Chatbot (December 20, 2017).