INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT (IJSREM)
VOLUME: 08 ISSUE: 11 | NOV - 2024 SJIF RATING: 8.448 ISSN: 2582-3930

HealthWise: AI-Powered Medical Chatbot with Llama 2

Ronak Umesh Bansal
Department of Computer Engineering and Technology
Dr. Vishwanath Karad *MIT World Peace University*
Pune, India
ronakbansal12345@gmail.com

Aditya Krishna Vangipurapu
Department of Computer Engineering and Technology
Dr. Vishwanath Karad *MIT World Peace University*
Pune, India
adikrish75@gmail.com

Swarup Shivaji Satav
Department of Computer Engineering and Technology
Dr. Vishwanath Karad *MIT World Peace University*
Pune, India
swarupsatav6@gmail.com

Ankush Nag
Department of Computer Engineering and Technology
Dr. Vishwanath Karad *MIT World Peace University*
Pune, India
ankushnag03@gmail.com

Rashmi Phalnikar
Department of Computer Engineering and Technology
Dr. Vishwanath Karad *MIT World Peace University*
Pune, India
rashmi.phalnikar@mitwpu.edu.in

***Abstract*— This research paper presents the development of a medical chatbot utilizing cutting-edge technologies, specifically the Langchain framework and the Llama 2 language model, in conjunction with the PineCone database for efficient data storage and retrieval. The primary objective of this project is to create an intelligent conversational agent capable of providing reliable medical information, assisting users in symptom assessment, and guiding them towards appropriate healthcare resources. The integration of Langchain facilitates seamless interaction between the user interface and the underlying LLM, enabling natural language understanding and generation. Leveraging the advanced capabilities of the Llama 2 model ensures that the chatbot can engage in meaningful dialogues while maintaining context and relevance. Additionally, PineCone serves as a robust solution for managing and indexing medical data, enhancing the chatbot's ability to deliver accurate and context-aware responses. This study explores the design, implementation, and potential implications of our medical chatbot, contributing to the emerging field of AI-driven healthcare solutions and underscoring its role in improving patient engagement and access to medical advice.**

***Keywords* – Medical Chatbot, Langchain, Llama-2, Pine Cone, Natural Language Processing, Artificial Intelligence, Machine Learning, Generative Artificial Intelligence.**

I. INTRODUCTION

In recent years, the explosion of artificial intelligence (AI) and machine learning (ML) technologies has profoundly impacted various sectors, including healthcare. At its core, AI encompasses the development of systems capable of performing tasks that typically require human intelligence, such as understanding natural language, recognizing patterns, and making decisions. Machine learning, a subset of AI, employs algorithms that progressively learn from data, improving their performance over time without explicit programming. The advent of generative artificial intelligence (Gen-AI) has further broadened the horizons of AI applications, enabling machines to generate original content, such as text or images, by understanding pre-existing data patterns. This capability is particularly pertinent in healthcare, where chatbots equipped with Gen-AI can assist in providing medical information, enhancing patient engagement, and facilitating timely interactions between patients and healthcare professionals.

In this research, we delve into the utilization of the Langchain framework, which is instrumental in the development of a medical chatbot. Langchain acts as a comprehensive toolset for building applications powered by large language models (LLMs), streamlining the integration of various components essential in the creation of AI-driven applications. The framework simplifies the process of connecting LLMs with external data sources, APIs, and other functionalities, making it easier for developers to design sophisticated chatbot systems that respond to user inquiries with contextual accuracy. The choice of Llama 2 as a specific LLM reinforces this research project. Developed by Meta, Llama 2 excels in natural language understanding and generation, thus enabling the chatbot to engage users effectively while maintaining a high standard of conversational coherence and relevance.

Central to our implementation is the use of Pinecone, a managed vector database that facilitates the storage, retrieval, and searching of embeddings generated by LLMs. By leveraging Pinecone's capabilities, our medical chatbot can efficiently manage large volumes of data, ensuring rapid response times and scalability as more users interact with the system. The integration of these technologies—Langchain, Llama 2, and Pinecone—marks a significant advancement in the development of medical chatbots that can serve as virtual health assistants. Through this research, we aim to demonstrate not only the technical feasibility of such systems but also their potential to transform patient care and information dissemination within the healthcare domain. By providing timely and reliable medical support through an intuitive interface, our project exemplifies how modern AI technologies can be harnessed to address critical challenges faced by healthcare providers and patients alike.

II. LITERATURE REVIEW

In paper [1], Realinho V. and others looked into whether higher education institutions can create knowledge and information based on the data they collect about their students. The study addresses the issue of school dropouts and educational underachievement, which has a negative effect on the economy, employment and competitiveness of a country. The research makes available a comprehensive dataset that incorporates enrollment and academic achievement, among other demographic, socioeconomic, microeconomic and academic variables. Such a dataset was compiled in order to devise an analytical instrument for predicting whether a particular student will succeed academically or not. This dataset can be of use to researchers in comparative analysis of students' academic achievement as well as machine learners working with the same dataset for training purposes.

In paper [2], Oloduowo A. et al. explored the importance of predicting student academic performance and its role in assessing students and helping them enhance their learning strategies. The paper reviews current research on predicting academic success and dropout rates across various fields of study. It provides a classification of the methods and attributes used for Student Academic Prediction and dropout prediction as found in the literature. Additionally, the paper addresses key challenges and limitations in forecasting academic outcomes and dropout rates, highlighting the shortcomings of existing methods. The study concludes with an overview of current research directions in academic and dropout prediction.

In paper [3], J. Berens et al. utilized student data from both public and private institutions to develop an early detection system for predicting student dropout. This Early Detection System (EDS) employs the AdaBoost algorithm to improve prediction accuracy, incorporating methods such as regression analysis, neural networks, and decision trees. By the end of the first semester, the EDS achieved a prediction accuracy of 79% for public institutions and 85% for private colleges of applied sciences. After the fourth academic term, prediction accuracy increased to 90% for public institutions and 95% for private applied sciences institutions. The study underscores the importance of identifying factors contributing to dropout and creating targeted interventions to reduce student attrition.

In paper [4], J. Alvarado-Uribe et al. addressed the issues of high failure and delayed graduation rates in South American higher education institutions, stressing the urgent need to identify students at risk of academic failure. The dataset used includes a range of information on undergraduate students, covering sociodemographic, academic, and aspects of student life. This dataset enables researchers to test various dropout prediction models to provide timely interventions for students who are at greater risk.

In paper [5], B. Kiss et al. focused on using machine learning methods to predict dropout rates and increase graduation rates at a Hungarian technical university. The researchers analyzed dropout predictions based on enrollment data and performance indicators from the first semester. They applied neural networks and boosting ensemble algorithms, examining the added predictive value of early college performance markers compared to pre-enrollment achievement metrics and vice versa. The study aims to identify students at risk and provide timely support to those individuals.

In paper [6], Alyahyan E. et al. offered educators a step-by-step guide for using data mining techniques to predict student achievement. While machine learning tools are increasingly valuable for identifying at-risk students and enhancing their outcomes, they may be challenging for educators without expertise in computer science or artificial intelligence. This work covers various topics, including defining academic success, recognizing key student characteristics, and choosing effective machine learning techniques. The study aims to make data mining tools more accessible to educators, helping them make informed decisions regarding student success predictions.

In paper [7], Zhau Ni. et al. describes on healthcare chatbots suggests their potential to improve access to quality healthcare, but their implementation remains unclear due to limited analysis of their adoption in medical settings. Existing research lacks comprehensive reviews on the use of chatbot technology in healthcare, prompting the need for bibliometric studies to analyze publication trends. This paper outlines a protocol for conducting a bibliometric analysis to systematically evaluate research on health-related chatbots, using databases such as CINAHL, IEEE Xplorer, PubMed, Scopus, and Web of Science. The study will identify patterns in the number of publications, geographic distribution, institutional contributions, and funding trends related to chatbot research. Additionally, it will examine the methodologies and applications of chatbots in healthcare,

providing insights into their features and development. Tools like VOSViewer will be used to visualize bibliometric networks, contributing to a clearer understanding of chatbot adoption in healthcare contexts.

In paper [8], Jingquan Li PhD worked on AI chatbots like ChatGPT emphasizes their potential to revolutionize patient care and public health through advanced natural language processing capabilities. Research highlights how these chatbots facilitate seamless communication by providing fluid, conversational responses, enhancing the user experience in healthcare settings. However, despite their promise, significant concerns exist regarding the data security implications of using such AI systems, as large datasets are required to train and refine them. Existing studies indicate that security risks, particularly concerning personal health information, are often understudied. This growing usage in healthcare necessitates a deeper examination of security vulnerabilities. Current literature calls for developing and implementing robust security safeguards and policies to mitigate risks while leveraging the benefits of AI chatbots like ChatGPT in the medical field.

In paper [9], Sara Hemdi Alqaidi. et al. working on medical chatbots demonstrates a growing trend in leveraging artificial intelligence (AI) and natural language processing (NLP) to enhance patient care. Previous research highlights the use of AI-based systems to improve communication between patients and healthcare providers, with an emphasis on virtual assistants. Chatbots, in particular, have been developed to provide automated medical services such as diagnosing conditions based on symptoms, tracking medications, and offering reminders. Tools like Dialogflow have been instrumental in building responsive, user-friendly chatbot frameworks that assist patients in managing their health and gaining insights into their conditions. This evolving technology addresses the need for accessible and personalized healthcare solutions, improving patient engagement and outcomes.

In paper [10], A mental health chatbot that uses the Structured Association Technique (SAT) for counselling is covered in the paper "A Chatbot System for Mental Healthcare Based on SAT Counselling Method" (2022). It simulates therapeutic talks with the use of natural language processing (NLP). It was effective at offering simple emotional support, but it was unable to handle difficult emotional circumstances or emergencies. The study underlines the need for stronger NLP models to capture subtle emotional language and improve the chatbot's usefulness in addressing important mental health issues.

In paper [11], The effectiveness of a chatbot that offers cognitive behavioural treatment (CBT) and psychoeducation to adults with attention deficit disorder (ADHD) is examined in the study "Mobile app-based chatbot to deliver cognitive behavioural therapy and psychoeducation for adults with attention deficit." The study demonstrated that the chatbot could successfully give individualised CBT techniques and instructional content to assist manage ADHD symptoms. It also focused on user engagement and treatment outcomes. Initial findings revealed positive user experiences and engagement, while the authors underlined the need for more research to analyse long-term efficacy and user retention tactics. In the end, the study indicates that chatbot-based therapies may be a useful tool for helping adults with ADHD, which calls for larger-scale clinical trials to evaluate their effects in practical contexts.

In paper [12], The study "Generative Pre-Trained Transformer-Empowered Healthcare Conversations" (MDPI, 2024) looks into how chatbot interactions can be improved in healthcare contexts by using Large Language Models (LLMs), particularly the GPT-4 architecture. Through the use of sophisticated Natural Language Processing (NLP) methods, the researchers show that these transformer-based models may mimic discussions between patients and providers more accurately, leading to better diagnosis and treatment suggestions. The study demonstrates how LLMs can provide patients with more accurate and beneficial replies by having a deeper understanding of the context and subtleties in medical discussions. But it also brings up ethical questions and the necessity for safe patient data management, underscoring the need of resolving these issues for real-world use in the healthcare industry.

III. SYSTEM ARCHITECTURE

The architecture of the medical chatbot system is designed to provide an efficient, scalable, and modular solution for handling medical queries using state-of-the-art natural language processing (NLP) techniques. A key component of the system is the LLaMA 2 (Large Language Model Meta AI), a powerful open-source model known for its enhanced performance in understanding and generating human language. LLaMA 2, developed by Meta, is fine-tuned specifically for medical applications in this system, enabling it to respond to user queries with a high level of accuracy. The model processes large amounts of medical text and can understand complex medical terminologies, making it well-suited for this domain.

The system's backend is designed to manage and process large medical documents in PDF format. These documents are parsed and divided into smaller, manageable chunks using a text splitter. This preprocessing step ensures that the input data adheres to the model's input token limit. Each chunk is then converted into vector embeddings using sentence transformers, which encode the semantic meaning of the text in a form that the LLaMA 2 model can process effectively. These embeddings allow the system to understand the relationships between words and phrases, facilitating accurate information retrieval.

To manage these vector embeddings, the system utilizes Pinecone, a cloud-based vector database optimized for vector similarity searches. Pinecone allows for fast and efficient

retrieval of relevant medical information from the indexed dataset when a user submits a query. By leveraging these vectorized embeddings, the chatbot can quickly locate the most pertinent data, ensuring that responses are both timely and accurate. This is particularly crucial in medical applications, where the precision of information is of utmost importance.

The user interface is the primary layer through which users interact with the system. It collects user inputs—such as medical questions or concerns—and passes them to the backend for processing. Upon receiving a query, the system employs a retrieval-based question-answering (QA) pipeline, which searches through the vectorized data stored in Pinecone. The relevant data is retrieved and then passed through the fine-tuned LLaMA 2 model, which generates a human-readable response based on the input query. This response is then delivered back to the user through the interface, completing the interaction loop.

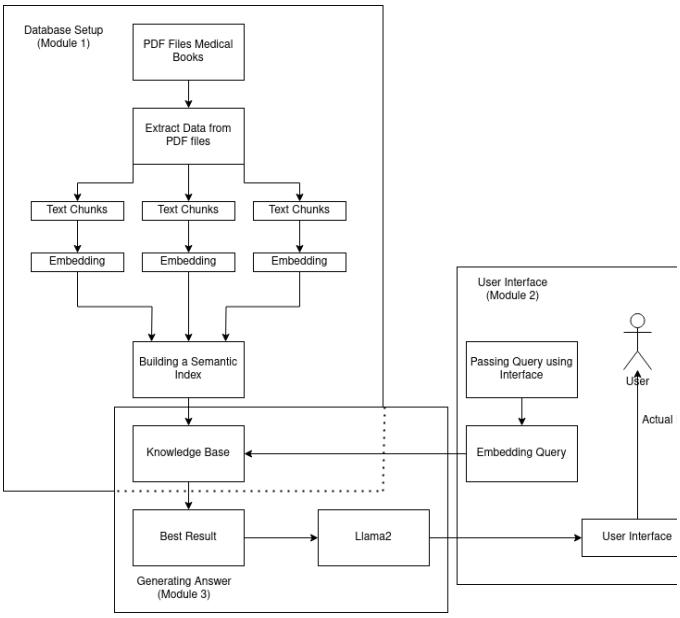

Fig. 1. System Architecture for Medical Chatbot

The architecture's modularity is one of its key strengths. Each component—data ingestion, vectorization, storage, retrieval, and user interaction—is designed to function independently, allowing for easy updates or replacements. This ensures that the system is not only scalable but also adaptable to future technological advancements. The chatbot's design focuses on flexibility, accuracy, and speed, making it a highly effective tool for answering complex medical queries in real-time, while relying on LLaMA 2's advanced language modeling capabilities to ensure high-quality responses.

IV. DATASET PREPARATION

This section provides an overview of the dataset utilized in the research paper, along with the methods used for processing the data.

- Data Collection

  The dataset utilized for this research was extracted from the Gale Encyclopedia of Medicine, Second Edition (Volume 1: A-B). It provides detailed coverage of a wide range of medical topics, including various conditions, diagnostic procedures, treatments, and alternative therapies. With a total of 637 pages and 1,700 in-depth articles, the dataset offers structured medical content on significant diseases, common health issues, and medical procedures. Moreover, the information is presented in a way that simplifies complex medical terminology, making it highly suitable for developing chatbots aimed at engaging with patients in a healthcare environment.

- Dataset Visualization

  1. **Distribution of Medical Topics**

Component	Count
Disorders/Conditions	765
Tests/Procedures	340
Treatments	255
Drugs	170
Alternatives	170

1. Table.1 summarizes the number of entries across the primary categories in the dataset.

2. Pie Chart: Content Distribution

This pie chart illustrates the proportional representation of different medical topics within the dataset:

The chart highlights how Disorders/Conditions make up the largest portion of the dataset, followed by Tests/Procedures, Treatments/Therapies, Drugs, and Alternative Therapies.

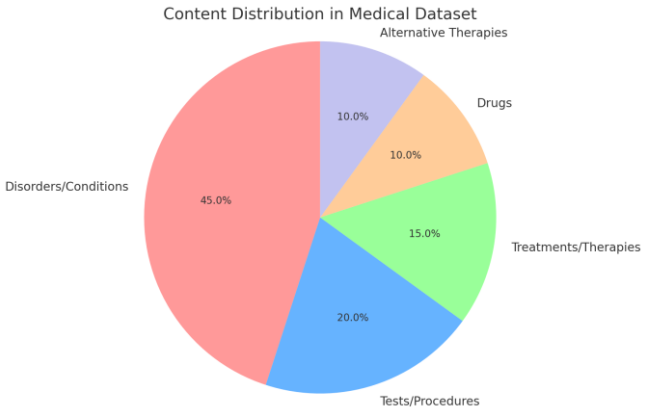

Fig. 2. Representing the distribution of the medical content in the dataset

- Data Cleaning

  The text extracted from the PDF underwent a cleaning process to remove unnecessary elements such as headers, footers, and redundant content, while also addressing any formatting inconsistencies to maintain uniformity. Non-essential information, such as page numbers, was eliminated to avoid introducing irrelevant data into the model.

- Data Structuring

  The extracted content was organized into specific categories, including:
  Disorders/Conditions
  Diagnostic Tests/Procedures
  Treatments/Therapies
  Medications
  Alternative Therapies
  Each entry was further divided into sections such as Definition, Symptoms, Diagnosis, Treatment, and Prognosis. This structure allows the chatbot to refer to relevant parts of a medical entry, improving the accuracy of its responses.

- Data Tokenization and Annotation

  Tokenization and annotation methods were employed to accurately identify and label medical terms. The annotations covered:
  Medical Conditions (e.g., Hypertension, Diabetes)
  Symptoms (e.g., Cough, Fever)
  Procedures (e.g., MRI, Ultrasound)
  These annotations were essential for training the chatbot to recognize medical entities and generate precise responses.

- Data Augmentation

  To enhance diversity and flexibility, data augmentation methods such as synonym substitution, paraphrasing, and context expansion were implemented. This approach allowed the chatbot to be trained on a wide range of inputs while preserving essential medical knowledge.

V. MODEL DEVELOPMENT

The development of the medical chatbot model employs a structured approach that combines data preprocessing, embedding generation, vector storage, and model inference, enhanced by mathematical theories and algorithms. This section outlines the critical components and the underlying mathematical principles involved.

Data Preprocessing and Text Segmentation: The first step in preparing the chatbot is data preprocessing, where medical documents are parsed into meaningful text chunks. Let T represent the total text and C be the length of the text that a transformer model can handle, constrained by the token limit. To divide T into manageable chunks $t_1,t_2,\ldots,t_n$, we use recursive character-based splitting, where each chunk $t_i\_t_i$ maintains a maximum length C while preserving the semantic coherence of the original text. The mathematical formulation can be represented as:

$$ti = T[j:k], where \ | ti | \leq C \ and \ j,k \in N. \ldots.(Eq.1)$$

This ensures that each chunk $t_i\_t_i$ fits within the transformer's token limit, maximizing information retention while minimizing truncation.

Embedding Generation Using Transformer Models: Once the data is segmented, the next step is to transform the text into vector embeddings. A pre-trained transformer model, such as SentenceTransformers, is used for this purpose. For each text chunk $t_i\_t_i$, the model generates an embedding $ei \in Rd$, where d is the dimension of the embedding space. The embedding is obtained through the following equation:

$$ei = f(ti;\theta), \ldots. (Eq.2)$$

where f represents the transformer model and θ denotes the model parameters learned during pre-training. These embeddings capture semantic relationships in the medical text, mapping the input text into a high-dimensional space that reflects the contextual meaning of the medical data.

Vector Storage and Retrieval: The embeddings are stored in a vector database for efficient retrieval during inference. The vector space R_d allows for similarity search using cosine similarity, defined as:

$$cosine\_similarity(ei, ej) = ei \cdot ej, \quad / \quad \| ei \| \| ej \| \ldots.(Eq.3)$$

where e_i and e_j are vector embeddings. The retrieval system ranks the embeddings by their similarity to the user query, retrieving the most relevant vectors for further processing.

Model Inference and Response Generation: For inference, the chatbot follows a Retrieval-Augmented Generation (RAG) framework. Upon receiving a query qqq, the most similar embeddings $e_1,e_2,\ldots,e_k$ are retrieved from the vector database using cosine similarity. These retrieved embeddings are fed into a language model, g, which generates a response r based on the input embeddings:

$$r = g(e1, e2, \ldots, ek; \phi), \ \ldots. (Eq.4)$$

where, ϕ represents the model parameters of the language model. This process ensures that the generated response is contextually accurate and grounded in the medical data, making the chatbot both reliable and informative.

VI. RESULTS

The development of medical chatbots has become increasingly significant, offering users accessible healthcare information through real-time interactions. These chatbots use natural language processing (NLP) and machine learning

models to understand and respond to queries about symptoms, medications, and treatments. This paper discusses the implementation of a medical chatbot using a modular architecture, pre-trained NLP models, and a vector database, highlighting its potential applications in healthcare.

The chatbot's design follows a modular approach, dividing the system into components such as data loading, model handling, and response generation. Medical information is uploaded in PDF format, processed into chunks, and vectorized for efficient querying. The chatbot uses a pre-trained transformer model from Hugging Face, fine-tuned on medical data to ensure accurate responses. The use of a modular structure allows for easy updates and improvements to the system, enabling future adaptability.

For information retrieval, the chatbot uses a vector database to store vectorized text representations. When a user asks a question, the system converts the query into a vector and matches it with relevant data in the database. This method, combined with the chatbot's Flask web application deployment, ensures quick and accurate responses through an intuitive user interface where users can type health-related questions.

VII. CONCLUSION AND FUTURE SCOPE

In conclusion, the development of a medical chatbot utilizing the Llama 2 model has demonstrated a significant advancement in the field of healthcare technology. The integration of Langchain for efficient management of conversational flows, Pine Cone for robust database storage, and various natural language processing (NLP) techniques has provided an effective framework for enhanced user interactivity and engagement. The implementation of machine learning algorithms for predictive analytics has further enriched the chatbot's capabilities, enabling it to offer personalized medical suggestions and facilitate timely interventions. The results obtained from this study illustrate that leveraging Llama 2's capabilities can lead to the creation of a responsive and informative assistant that meets the needs of users seeking medical advice, thereby improving user trust and satisfaction.

Looking towards the future, there are several avenues for further enhancing the functionality and effectiveness of the medical chatbot developed in this study. One potential direction is the incorporation of multimodal capabilities, allowing the chatbot not only to process text but also to interpret images and videos, such as diagnostic imaging or patient-generated content. Additionally, expanding the training data to include diverse patient interactions and medical scenarios can significantly enhance the chatbot's contextual understanding and accuracy in providing medical guidance. There is also potential for integrating speech recognition technologies to enable voice interactions, which could improve accessibility for users who prefer verbal communication.

Moreover, the continuous evolution of AI technologies presents opportunities for more sophisticated and contextually aware medical chatbots. Future research could focus on refining the algorithms for more nuanced understanding of medical jargon, improving response times, and ensuring the chatbots maintain compliance with regulations such as HIPAA (Health Insurance Portability and Accountability Act). Collaborations with healthcare professionals to validate the chatbot responses and keeping the knowledge base regularly updated are essential steps to ensure reliability and trustworthiness in a clinical setting. Overall, the next steps will emphasize enhancing user experience while maintaining ethical considerations and patient safety at the forefront of innovation in medical conversational agents.

2. REFERENCES

[1] S. Chakraborty et al., "An AI-Based Medical Chatbot Model for Infectious Disease Prediction," in IEEE Access, vol. 10, pp. 128469-128483, 2022, doi: 10.1109/ACCESS.2022.3227208.

[2] Li Y, Li Z, Zhang K, Dan R, Jiang S, Zhang Y. ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge. Cureus. 2023 Jun 24;15(6):e40895. doi: 10.7759/cureus.40895. PMID: 37492832; PMCID: PMC10364849.

[3] Srivastava, P., & Singh, N. (2020). Automatized Medical Chatbot (Medibot). 2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and Its Control (PARC). doi:10.1109/parc49193.2020.236624.

[4] A. Babu and S. B. Boddu, "BERT-Based Medical Chatbot: Enhancing Healthcare Communication through Natural Language Understanding," Exploratory Research in Clinical and Social Pharmacy, vol. 13, p. 100419, 2024, doi: 10.1016/j.rcsop.2024.100419.

[5] A. P. S. Sasan, A. Kumar, A. K. Singh, and A. Baruah, "A Research Paper of a Medical Chatbot using Llama 2," International Journal for Research in Applied Science & Engineering Technology (IJRASET), vol. 12, no. 4, pp. 859-865, Apr. 2024, doi: 10.22214/ijraset.2024.59868

[6] A. R. E. Tjiptomongsoguno, A. Chen, H. M. Sanyoto, E. Irwansyah, and B. Kanigoro, "Medical Chatbot Techniques: A Review," in CoMeSySo 2020, R. Silhavy et al., Eds., Springer, 2020, pp. 1-11, doi: 10.1007/978-3-030-63322-6_28.

[7] Zhao Ni, Mary L. Peng, et al. A bibliometric analysis of chatbot technology in healthcare: study protocol. JMIR Research Protocols November 07 (2023). https://preprints.jmir.org/preprint/54349.

[8] Jingquan Li. Security Implications of Artificial Intelligence Chatbots in Healthcare. Journal of Medical Internet Research. March 24 (2023). https://preprints.jmir.org/preprint/47551.

[9] Sara Hemdi Alqaidi, Shahad Mohammed Albugami, et al. Network-integrated medical chatbot for enhanced healthcare services. Telematics and Informatics Reports 15 (2024) 100153.

[10] Longe, J. L., & Blanchfield, D. S. (Eds.). (2002). The Gale encyclopedia of medicine (2nd ed., Vol. 1). Gale Group.

[11] Xu, L., Sanders, L. , Li, K. , Chow, J.C.: Chatbot for health care and oncology applications using artificial intelligence and machine learning: systematic review. JMIR Cancer 7(4), 07850 (2021)

[12] Abd-Alrazaq, A.A., Rababeh, A. , Alajlani, M. , Bewick, B.M., Househ, Me: Effectiveness and safety of using chatbots to improve mental health: systematic review and meta-analysis. J. Med. Internet Res. 22(7), e16021 (2020)

[13] M Roshan and AP Rao"A study on relative contributions of the history, physical examination and investigations in making medical diagnosis."In The Journal of the Association of Physicians of India 48.8, 2000, pp. 771–775