IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, November 2024

Health Diagnostic Assistant using LLMs

Laxmikant Malphedwar[1], Anerao Monika[2], Dhole Mangesh[3], Dixit Tanmay[4], Gaikwad Raman[5]
Associate Professor, Dr. D Y Patil College of Engineering & Innovation, Pune, India[1]
UG Students, Dr. D Y Patil College of Engineering & Innovation, Pune, India[2,3,4,5]

Abstract: *The Health Diagnostic Assistant leverages advanced Large Language Models (LLMs) and Natural Language Processing (NLP) techniques to enhance patient diagnosis and healthcare decision-making. This innovative system employs Retrieval-Augmented Generation (RAG) to combine the strengths of pre-trained language models with a dynamic retrieval mechanism, allowing it to access and synthesize real-time medical knowledge from a wide array of databases. By analyzing patient symptoms, medical histories, and contextual data, the assistant generates accurate, context-aware recommendations and insights. The project aims to streamline the diagnostic process, reduce the burden on healthcare professionals, and improve patient outcomes by providing evidence-based suggestions tailored to individual cases. Through continuous learning and integration of user feedback, the Health Diagnostic Assistant aspires to evolve into a reliable tool for both patients and clinicians, fostering informed decision-making in the healthcare landscape.*

Keywords: LLM, NLP, RAG, Medical History, Healthcare Professionals

I. INTRODUCTION

In recent years, the healthcare sector has witnessed a paradigm shift driven by advancements in technology, particularly in the fields of artificial intelligence (AI) and machine learning.[1] As medical knowledge expands exponentially, healthcare professionals face increasing challenges in keeping pace with the latest research, treatment modalities, and diagnostic criteria. To address these challenges, there is a growing need for intelligent systems that can assist in clinical decision-making, enhance patient engagement, and improve diagnostic accuracy. This research paper aims to make the integration of LLMs into medical chat bots and evaluate their effectiveness in improving patient engagement, healthcare accessibility, and overall user satisfaction. [2] By harnessing the vast knowledge encoded within these language models, medical chat bots can offer valuable insights, assist in symptom assessment, provide medication advice, offer lifestyle recommendations, and even facilitate mental health support.

The rapid advancement of artificial intelligence (AI) has covered the way for significant developments in natural language processing (NLP), with large language models (LLMs) standing out as a remarkable achievement. These models leverage deep learning techniques, particularly neural networks with multiple layers, to understand and produce highly proficient human language.[3]

II. MOTIVATION

Using Large Language Models (LLMs) in health diagnostics can be a highly effective way to assist healthcare providers in diagnosing and managing conditions, especially when combined with appropriate medical data and a structured approach. Here's how LLMs can be utilized for health diagnostics, framed in objectives:

Enhance Diagnostic Accuracy Objective: To assist healthcare providers in patient data and identifying potential conditions by processing symptoms, medical history, and lab results.

How: LLMs can help by extracting and organizing symptoms, cross-referencing medical literature, and suggesting possible diagnoses. This process aids clinicians in making more informed decisions and reducing human error.

Intrinsic vs. Extrinsic Motivation

- **Intrinsic Motivation**: This comes from within. It's driven by personal satisfaction or a deep sense of fulfillment. For instance, someone may exercise not to lose weight but because it makes them feel stronger and healthier.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, November 2024

- **Extrinsic Motivation**: This comes from external factors, such as rewards, praise, or avoiding punishment. For example, a person may choose to follow a health regimen to get compliments or avoid the consequences of poor health.
- **In Health Contexts**: Both types of motivation play a role in health and wellness. While intrinsic motivation often leads to more sustainable and long-term health habits (e.g., enjoying physical activity for its own sake), extrinsic motivators (e.g., a health-related goal like weight loss) can provide initial momentum.

III. OBJECTIVE

A Health Diagnostic Assistant using Large Language Models (LLMs) like GPT can be an invaluable tool in supporting healthcare providers and individuals by leveraging AI for various diagnostic and health-related tasks. The objectives for such a system can span multiple areas, including medical assistance, data interpretation, and improving healthcare accessibility.

Medical Knowledge Integration

- **Objective**: Integrate a vast database of medical knowledge (e.g., diseases, symptoms, treatments) to assist in diagnosing conditions.
- **Outcome**: Enable the LLM to interpret symptoms, medical history, and other inputs to offer potential diagnoses or insights.

Symptom Checking and Diagnosis Support

- **Objective**: Provide symptom checking capabilities to help users better understand possible causes for their symptoms.
- **Outcome**: A diagnostic assistant that can ask users relevant questions about their symptoms, past medical history, and lifestyle factors to suggest possible conditions or next steps.

Clinical Decision Support (CDS)

- **Objective**: Assist healthcare professionals by suggesting possible diagnoses and treatments based on patient symptoms, history, and diagnostic results.
- **Outcome**: Improve decision-making efficiency and accuracy for healthcare providers, helping them avoid misdiagnoses.

Patient Education and Health Literacy

- **Objective**: Educate patients about their conditions, treatments, and prevention strategies in a manner that is easy to understand.
- **Outcome**: Empower patients by giving them the knowledge to make informed decisions about their healthcare.

IV. LITERATURE SURVEY

Large Language Models (LLMs):

LLMs are neural networks trained on vast datasets to understand and generate human-like text. They leverage architectures like Transformers, which allow them to learn contextual relationships between words.

Key Models:

- **GPT-3 and GPT-4** (Open AI): Notable for their size and versatility in various NLP tasks. They exhibit impressive coherence and contextual understanding.
- **BERT and its Variants** (Google): Bidirectional models that excel in tasks requiring understanding of context, like question answering and sentiment analysis.

Accuracy and Performance:

- LLMs like GPT-3 and GPT-4 have achieved state-of-the-art results on several benchmarks, including the Super GLUE and SQUAD datasets.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, November 2024

- However, accuracy can vary based on the task; for example, BERT performs exceptionally well in tasks requiring fine-grained understanding of context but may lag in creative text generation compared to GPT models.[11]
- NATURAL LANGUAGE PROCESSING NLP encompasses a range of techniques and methods for processing and analysing human language. Key areas include sentiment analysis, translation, summarization, and named entity recognition.

Advancements:

- Traditional methods were based on rule-based systems or simple statistical models. The advent of neural networks has significantly improved performance.
- Transfer learning with pre-trained models like BERT and LLMs has revolutionized NLP, allowing for rapid advancements across various tasks with fewer resources.

Accuracy:

- As of late 2023, benchmark scores on datasets such as GLUE and SuperGLUE have seen improvements, with many models achieving near-human performance in certain tasks.
- However, challenges remain in domains requiring deep contextual understanding or domain-specific knowledge, where performance can drop.[12]

RETRIEVAL ARGUMENTDED GENERATION

It first retrieves relevant documents or data from a knowledge base and then generates responses based on this information.

Accuracy and Performance:

RAG models, such as those based on T5 and BART, have shown substantial improvements in tasks like open-domain question answering, outperforming traditional generative models that rely solely on their training data.

In benchmark evaluations, RAG systems have been shown to generate more accurate and contextually relevant responses, particularly in dynamic knowledge areas where real-time information is crucial.[13]

V. METHODOLOGY

Decision Trees:

Algorithms like ID3 or C4.5 were used to classify medical data. Decision trees help in making decisions based on a series of questions about patient symptoms or test results.[4]

ID3(Iterative Dichotomiser 3):

- Tree Construction: ID3 constructs a decision tree by recursively splitting the dataset based on feature values.
- Information Gain: The algorithm uses the concept of information gain, derived from entropy, to choose the feature that best separates the classes at each node.
- Stopping Criteria: The recursion stops when all instances in a node belong to the same class or when there are no more features to split on.[5]

C4.5:

- Tree Construction: Similar to ID3, C4.5 builds trees by splitting the dataset based on feature values.
- Gain Ratio: Instead of information gain, C4.5 uses gain ratio to choose the best attribute for splitting. This helps mitigate the bias that information gain has toward attributes with many distinct values.
- Pruning: C4.5 incorporates a pruning step to remove branches that have little importance, reducing the risk of overfitting.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, November 2024

Both algorithms are widely used in various domains, including healthcare, finance, and marketing, for tasks like classification, risk assessment, and decision support. C4.5, in particular, has influenced many modern machine learning tools and algorithms.

Expert Systems:
Systems like MYCIN (for infectious diseases) and DENDRAL (for chemical analysis) utilized knowledge bases and inference engines to assist in diagnosis and treatment recommendations.
Both MYCIN and DENDRAL are significant in the history of artificial intelligence as they demonstrated the practical applications of expert systems. They laid the groundwork for later developments in AI, particularly in the fields of medical diagnosis and scientific research, influencing the design of modern decision support systems and knowledge-based applications.[6]

Bayesian Networks:
These probabilistic graphical models represent a set of variables and their conditional dependencies, allowing for reasoning under uncertainty, which is common in medical diagnoses.[7]
The Health Diagnostic Assistant employs a multi-faceted approach that integrates Large Language Models (LLMs), Natural Language Processing (NLP), and Retrieval-Augmented Generation (RAG) techniques to deliver comprehensive diagnostic support.

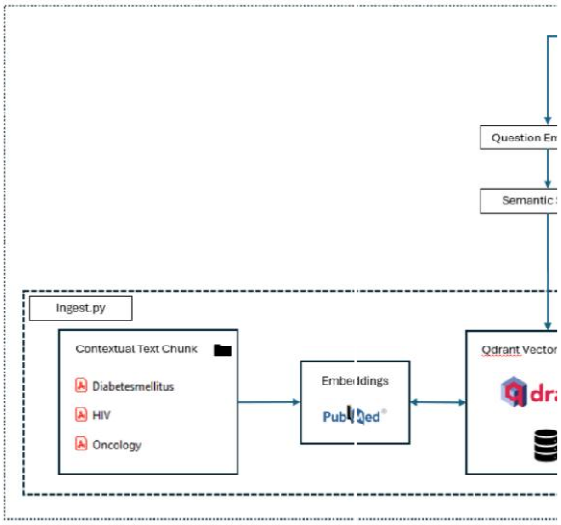
1. Large Language Models (LLMs):
LLMs are capable of understanding and generating human-like text, allowing them to process user inputs effectively and generate coherent, context-aware responses.
A Large Language Model (LLM) is a type of artificial intelligence (AI) designed to understand, generate, and manipulate human language. These models are built on advanced neural network architectures, primarily transformers, which allow them to process and analyze vast amounts of text data.[8]

2. Natural Language Processing (NLP):
Using NLP techniques, the assistant analyzes user inputs to identify key symptoms and relevant medical history. Named Entity Recognition (NER) is employed to extract symptoms, conditions, medications, and other critical information from free-text inputs.
Natural Language Processing (NLP) is a subfield of artificial intelligence focused on the interaction between computers and human language. It involves enabling machines to understand, interpret, and generate natural language text or speech. Key applications include sentiment analysis, language translation, chatbots, and information retrieval. NLP combines linguistics, computer science, and machine learning techniques to analyze and process language data. Challenges include handling ambiguity, context, and the nuances of human communication. As NLP technology advances, it continues to enhance how we interact with machines, making them more intuitive and capable of understanding human language effectively.[9]

3. Retrieval-Augmented Generation (RAG):
RAG enables the assistant to access up-to-date medical databases, clinical guidelines, and research articles in real-time. This ensures that the responses generated are not only accurate but also reflect the latest advancements in medical science.
Retrieval-Augmented Generation (RAG) is a natural language processing framework that integrates information retrieval with generative models. It operates by first retrieving relevant information from a knowledge base based on a user query. This retrieved data is then used to inform the generation of a coherent and contextually relevant response. RAG enhances accuracy and relevance by grounding the generation process in factual content, making it particularly effective for applications like customer support, knowledge-based systems, and conversational AI. By combining these two approaches, RAG provides more informative and nuanced outputs compared to traditional generative models alone.[10]

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, November 2024

VI. ARCHITECTURE



- **PubMed:** PubMed is a free, searchable database of biomedical literature, primarily focused on research articles in life sciences and health. It is hosted by the National Library of Medicine (NLM) at the U.S. National Institutes of Health (NIH). PubMed provides access to millions of references from a wide range of sources, including journal articles, clinical studies, systematic reviews, and more.
- **Searchable database**: You can search for scientific literature using keywords, authors, topics, or PubMed IDs (PMID).
- **Access to abstracts**: For many articles, PubMed provides abstracts (summaries) of the research.
- **Free full-text links**: While PubMed itself doesn't host full-text articles, it often links to publishers' websites where you can access the full text, sometimes for free, sometimes behind a paywall.
- **Filters and advanced search**: PubMed offers various filters to refine searches (e.g., article type, publication date, study species).
- **Citations and references**: PubMed provides citation details and links to other related articles. [14]
- **Qdrant Vector Database:** is an open-source vector database designed for high-performance search and retrieval of vector embeddings. It enables efficient similarity search across large datasets of high-dimensional vectors, which are commonly generated from machine learning models (such as embeddings from natural language processing (NLP) models, image recognition models, etc.) It enables efficient similarity search, allowing users to quickly find and retrieve similar items based on their vector representations.
- **gradio:** is an open-source Python library designed to help developers and data scientists quickly create user-friendly web interfaces for machine learning models and data processing tasks. It allows you to create interactive demos with minimal effort, enabling non-technical users to interact with models and algorithms in a more intuitive way.is often used for prototyping, showcasing models, and collaborating on data science and AI projects.
- It simplifies the process of building graphical interfaces (GUIs), making it possible to quickly demo and interact with models or functions without needing web development skills.

VII. CONCLUSION

This review highlight the promising role of a Health Diagnostic Assistant powered by Large Language Models (LLMs) holds immense potential to revolutionize healthcare by supporting both patients and healthcare professionals. By integrating vast medical knowledge, enabling intuitive interactions, and offering personalized recommendations, such a system can assist in the early detection of health issues, provide decision support for clinicians, and enhance patient

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, November 2024

education and engagement. The key to success lies in ensuring diagnostic accuracy, maintaining privacy and security, and continuously updating the system with the latest medical advancements.

This approach not only has the potential to improve individual patient outcomes, but also to alleviate pressure on healthcare systems by providing efficient, accessible, and scalable tools. However, the implementation must be done with careful attention to issues like data privacy, model bias, and regulatory compliance to build trust and ensure ethical use of the technology

REFERENCES

[1] Jiang, X.; Yan, L.; Vavekanand, R.; Hu, M. Large Language Models in Healthcare Current Development and Future Directions. Preprints 2024, 2024070923

[2]B. Galitsky, "LLM-based Personalized Recommendations in Health," 2024, doi: 10.20944/preprints202402.1709.v1.

[3] Tom, Young., Devamanyu, Hazarika., Soujanya, Poria., Erik, Cambria. (2017). 2. Recent Trends in Deep Learning Based Natural Language Processing. arXiv: Computation and Language,

[4] Ruijuan, Hu. (2011). Medical Data Mining Based on Decision Tree Algorithm. Computer and Information Science, doi: 10.5539/CIS.V4N5P14

[5] Prashant, G., Ahire., Student, Sanket, Kolhe., Kunal, Kirange., Hemant, Karale., Abhilasha, Bhole. (2015). Implementation of Improved ID3 Algorithm to Obtain more Optimal Decision Tree..

[6] Ajith, Abraham. (2005). 4. 130: Rule-based Expert Systems. doi: 10.1002/0471497398.MM422

[7] Wim, Wiegerinck., Bert, Kappen., Willem, Burgers. (2010). Bayesian Networks for Expert Systems: Theory and Practical Applications. Studies in computational intelligence, doi: 10.1007/978-3-642-11688-9_20

[8] S. A. Gebreab, K. Salah, R. Jayaraman, M. Habib ur Rehman and S. Ellaham, "LLM-Based Framework for Administrative Task Automation in Healthcare," 2024 12th International Symposium on Digital Forensics and Security (ISDFS), San Antonio, TX, USA, 2024, pp. 1-7

[9] Irene, Li., Jessica, Pan., Jeremy, Goldwasser., Neha, Verma., Wai, Pan, Wong., Muhammed, Yavuz, Nuzumlali., Benjamin, Rosand., Yixin, Li., Matthew, Zhang., David, Chang., Richard, Andrew, Taylor., Harlan, M., Krumholz., Dragomir, R., Radev. (2021). Neural Natural Language Processing for Unstructured Data in Electronic Health Records: a Review.. arXiv: Computation and Language,

[10] Rama, Akkiraju., Anbang, Xu., D., Bora., Yu, Tan., Lu, An., Vijay, K., Seth., Aaditya, Shukla., Pritam, Gundecha., Harsh, Mehta., Ajay, Kumar, Jha., Prithvi, Raj., Abhinav, Balasubramanian., Murali, Maram., Muthusamy, Gunasegaran., Shivakesh, Reddy, Annepally., S., H., Knowles., Min, Du., N., H., Burnett., Sean, Javiya., Ashok, Marannan., Mamta, Kumari., Sanjeev, Jha., Ethan, Dereszenski., Abhishek, A., Chakraborty., Subhash, Ranjan., Amina, Terfai., Anupma, Surya., T.T., Mercer., Vinodh, Kumar, Thanigachalam., Tamar, Bar., Subhashree, Mallika, Krishnan., Samy, Kilaru., Jasmine, Jaksic., Nave, Algarici., Jacob, Liberman., John, S., Conway., Sachin, Nayyar., Justin, Boitano. "FACTS About Building Retrieval Augmented Generation-based Chatbots." null (2024). doi: 10.48550/arxiv.2407.07858

[11] Jesutofunmi, A., Omiye., Haiwen, Gui., Shawheen, J., Rezaei., James, Zou., Roxana, Daneshjou. (2024). Large Language Models in Medicine: The Potentials and Pitfalls. Annals of Internal Medicine, 177(2):210-220. doi: 10.7326/m23-2772

[12] Roman, Egger., Enes, Gokce. (2022). Natural Language Processing (NLP): An Introduction. Tourism on the Verge, 307-334. doi: 10.1007/978-3-030-88389-8_15

[13] Sudeshna, Das., Yunsheng, Ge., Yuting, Guo., Swati, Rajwal., JaMor, Hairston., Jeanne, M., Powell., Andrew, Walker., Snigdha, Peddireddy., Sahithi, Lakamana., Selen, Bozkurt., Matthew, A., Reyna., Reza, Sameni., Yunyu, Xiao., Sangmi, Kim., Rasheeta, Chandler., Natalie, D., Hernández., Danielle, L., Mowery., Rachel, Wightman., Jennifer, C., Love., Anthony, Spadaro., Jeanmarie, Perrone., Abeed, Sarker. (2024). Two-layer retrieval augmented generation framework for low-resource medical question-answering: proof of concept using Reddit data. doi: 10.48550/arxiv.2405.19519

[14] Jian, Xu., Sunkyu, Kim., Min, Song., Minbyul, Jeong., Donghyeon, Kim., Jaewoo, Kang., Justin, F., Rousseau., Xin, Li., Weijia, Xu., Vetle, I., Torvik., Yi, Bu., Chongyan, Chen., Islam, Akef, Ebeid., Daifeng, Li., Ying, Ding. (2020). Building a PubMed knowledge graph.. Scientific Data, 7(1):205-. doi: 10.1038/S41597-020-0543-2

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, November 2024

[15] (2023). Are Large Language Models Ready for Healthcare? A Comparative Study on Clinical Language Understanding. doi: 10.48550/arxiv.2304.05368
[16] Weiner, S. J., Wang, S., Kelly, B., Sharma, G., & Schwartz, A. (2020). How accurate is the medical record? A comparison of the physician's note with a concealed audio recording in unannounced standardized patient encounters. Journal of the American Medical Informatics Association, 27, 770-775.
[17] R., Kavitha., E., Kannan., S., Kotteswaran. (2016). 5. Implementation of Cloud based Electronic Health Record (EHR) for Indian Healthcare Needs. Indian journal of science and technology, doi: 10.17485/IJST/2016/V9I3/86391
[18] Xiangbin, Meng., Xiangyu, Yan., Kuo, Zhang., Da, Liu., Xiaojuan, Cui., Yaodong, Yang., Muhan, Zhang., Chunxia, Cao., Jingjia, Wang., Xuliang, Wang., Jun, gao., Yuangengshuo, Wang., Zifeng, Qiu., Muzi, Li., Cheng, Qian., Tianze, Guo., Shuangquan, Ma., Zeying, Wang., Zexuan, Guo., Youlan, Lei., Chunli, Shao., Wenyao, Wang., Haojun, Fan., Yifang, Tang. (2024). 1. The Application of Large Language Models in Medicine: A Scoping Review. iScience,
[19] Vavekanand, R., & Kumar, S. (2024). LLMEra: Impact of Large Language Models. Available at SSRN 4857084.
[20] Borude, Shrikant N; Kolhe, Saurabh R; Patil, Harshal R; Malphedwar, Laxmikant; CNN BASED METHOD FOR LUNG DISEASE DETECTION INTERNATIONAL JOURNAL 5/6/2020