2024 International Conference on Emerging Technologies and Innovation for Sustainability (EmergIN)

Healthcare Diagnostic RAG-Based Chatbot Triage Enabled by BioMistral-7B

Kushagra Sinha
Department of Computer Science and Engineering Graphic Era (Deemed To be University)
Dehradun, India
kushagrasinha123ks@gmail.com

Vaibhav Singh
Department of Computer Science and Engineering Graphic Era (Deemed To be University) Dehradun, India
vaibhav27003@gmail.com

Ankit Vishnoi
Department of Computer Science and Engineering Graphic Era (Deemed To be University) Dehradun, India
ankitvishnoi.cse@geu.ac.in

Parul Madan
Department of Computer Science and Engineering Graphic Era (Deemed To be University) Dehradun, India
parulmadan.cse@geu.ac.in

Yadvendra Shukla
Department of Computer Science and Engineering Graphic Era (Deemed To be University)
Dehradun, India
yadvendrashukla919@gmail.com

2024 International Conference on Emerging Technologies and Innovation for Sustainability (EmergIN) | 979-8-3503-9126-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/EMERGIN63207.2024.10961136

***Abstract*—Retrieval-augmented generation (RAG) approaches represent a significant step forward in patient triage and assistance, effectively enhancing user engagement and diagnostic accuracy when combined with the development of chatbots. This paper proposes an advanced design of a chatbot system aimed at enhancing support for the triage and diagnosis of patients through the utilization of sophisticated models like BioMistral7B and PubMedBert. It employs these models along with the Conversational Retrieval Chain (CRC) in effectively handling user queries that fall under core medical and triage-based questions. Accordingly, the core medical reference books—among which are included The Merck Manual of Diagnosis and Therapy, Harrison's Principles of Internal Medicine, The Gale Encyclopedia of Medicine, and The Oxford Handbook of Clinical Medicine—are converted into the form of vector embeddings and stored in the Qdrant vector database for subsequent quick retrievals. The performance evaluations further proved that the chatbot was able to attain responses within the range of 22 to 35 seconds, with an average response time of 28 seconds, when tested on modest hardware setups, while the BLEU & ROUGE scores reflect its outputs to be of decent quality. The findings reveal that AI-driven chatbots hold promising possibilities in enhancing diagnostic accuracy as well as improving patient engagement in various healthcare settings.**

***Keywords*—bioMistral-7B, large language models, patient triage, rag, vector embeddings, conversational retrieval chain, vector databases, patient triage**

I. INTRODUCTION

The domains of patient communication and medical diagnosis are changing as a result of the application of artificial intelligence in healthcare services. Current advancements show how AI may significantly increase the accuracy and efficacy of medical chatbots, particularly when combined with models such as BERT. These advancements have improved chatbots' ability to understand and respond to medical queries with high precision, which has addressed the drawbacks of traditional systems [1]. This study investigates the integration of ChatGPT and GPT4 technologies into eHealth and mHealth systems, demonstrating their capacity to enhance patient care and medical data transmission [2]. The integration of AI models into healthcare is aimed at improving access and education.

According to Wang et al. (2024) [3] study, ChatGPT excels in real-time patient feedback analysis in hospitals compared to CNN and LSTM, indicating its ability to improve patient experience and engagement.

Holderried et al. (2023) [4] found GPT-3.5's chatbots to offer satisfactory experiences and believable responses when collecting medical history. The study indicates that the chatbot may need to enhance its medical accuracy by minimizing responses influenced by social desirability.

AI and ML are essential to healthcare - particularly in diagnoses, patient management, and triage. Feasibility study results show that besides accuracy, AI systems need to be examined for their scalability and the feasibility of their potential applications in real-world scenarios. Recent research has proven that the AI models deployed in emergency departments drive better patient outcomes through better resource distribution and give increased prediction accuracy over the classic triage methods [5]. Studies like this highlight the need to assess the feasibility of the functionality of the AI systems in real-life clinical setups; it is crucial to effectively implement AI.

This paper developed an advanced chatbot system for enhancing patient triage and assistance based on sophisticated models such as BioMistral-7B, PubMedBert, and Qdrant, in an effort to achieve improved diagnostic accuracy and faster data retrieval. The advancement of a Conversational Retrieval Chain (CRC) to better conversational context and the relevance of responses, together with a strong backend and an accessible frontend, underlines the promise of the system not just with respect to performance but also to scalability and applicability in real-world settings. This study provides important perspectives on the use of AI technologies for practical implementation in healthcare, addresses the challenges mentioned in the feasibility analyses performed earlier, and encourages personalized patient participation.

II. LITERARY REVIEW

In the editorial "AI's Influence on Medicine: Perspectives from a Chatbot," King discusses the significant impact of AI in healthcare, emphasizing its role in areas like personalized

treatments, surgical support, and diagnostic imaging. The article reviews AI advancements in medicine and projects future applications that could improve diagnosis and treatment options for specialists in radiology, oncology, and primary care [6].

The chatbot for Multiple Myeloma using Retrieval Augmented Generation is introduced, which utilized state-of the-art NLP methods to analyze and curate patient genomic data. The platform uses BioMed-RoBERTa-base for embedding and the Mistral-7B model for question-answering. It uses Amazon OpenSearch Service and Amazon Kendra for scalable information retrieval, personalizing treatment plans, and advancing research [7].

To address the complexities of queries, this study introduces RQ-RAG, an enhanced Retrieval-Augmented Generation model that incorporates relevant information using tailored search queries. It outperforms other models in multi-hop tasks and improves by 1.9% on single-hop QA tests [8].

Prompt-RAG is introduced as a new RAG method in which, through the substitution of natural language prompts instead of traditional ones, performance is improved in generating specialized topics by LLMs. Our method surpasses both ChatGPT and competitive models in relevance and informativeness by retrieval-augmented generation, LLM-guided heading selection, and preprocessing by a table of contents [9].

Another experiment evaluates pretraining strategies to improve the model's performance in classification tasks across domains. It was discovered that a combination of domain adaptive and task-adaptive techniques improves the results, especially in low-resource settings where multi-phase pretraining further improves the outcome [10].

A study on 10 mental health apps found that chatbots, while providing personalized support, often struggle with crisis identification and excessive reliance. Moreover, the research discovered that though they present an impartial atmosphere for disclosing personal information, they might promote overdependence thus diminishing actual human contact. Hence, it is very important to carefully come up with and modify chatbots for them to give effective mental support [11].

In the paper, a different AI model is proposed that can help with changing behavior by designing chatbot traits based on user comprehension, relational dexterity, persuasive dialogue skills, means and ends evaluation. The objective of this study was to advocate for collaboration among various fields and embrace ethical issues when applying AI as an effective tool in behavior change. The framework seeks to direct AI chatbots' design and assessment aimed at enhancing physical exercise along with dietary habits [12].

Numerous restrictions remain in place, even in light of the scientific literature that emphasizes significant advancements in artificial intelligence and chatbots within the healthcare sector.

Many models show decent performance in specific medical specialties, such as myeloma or pregnancy, but also face challenges when attempting to generalize across diverse medical areas, mainly due to inconsistencies in data and issues related to unstructured data. In particular to agents, challenging issues apply, both when creating datasets and in terms of accuracy. Specific to mental health, chatbots cannot help to manage a crisis and exacerbate over-reliance. Healthcare systems require strong and scalable AI solutions, which is noted by ethical considerations including averting bias and transdisciplinary collaboration.

III. DATASET AND DATA PREPARATION

A. Biomedical *Text Corpora*

Benchmarking is very important in the health industry as far as the effectiveness and dependability of biomedical language models are concerned. Such an ability enables the comparison of several different models in an organized manner thereby making their benefits and drawbacks clear. This review is fundamental to establishing what precise models can be used in treatment recommendations or for research activities on the advancement of medical LLM [13].

The Open Medical-LLM Leaderboard measures different large language models against multiple medical question and answer tasks. Models such as GPT-4-base, Med-PaLM2, Starling-LM-7B, gemma-7b, and Mistral-7B-v0.1 have recorded high levels of accuracy on diverse medical data sets hence they are robust across different medical fields.

These models' performances vary when it comes to tasks like understanding biomedical literature and making clinical decisions. One model stands out most distinctly among others; BioMistral-7B which has been trained using PubMed Central resources. This particular LLM model has excelled in multilingual medical evaluation benchmarks making it much better than other open-source models and also competes effectively with proprietary ones. The relevance of this Biomedical model lays in its focus on biomedical texts hence it is especially important for improving diagnostic accuracy and personalized treatment.

B. Vector Databases and Text Embeddings

Embeddings: Embeddings are used to capture the underlying meanings and interrelations of words, phrases, and texts in a manner that allows computer systems to understand language. The textual data is transformed into numerical vectors that would unveil semantic patterns through embeddings during our project.

Task	BioMistral-7B	Mistral 7B Instruct	BioMistral-7B En-semble	BioMistral-7B DARE	BioMistral-7B TIES	BioMistral-7B SLERP	MedAlpaca 7B	PMCLLaMA 7B	MediTron-7B	BioMed GPT-LM-7B	GPT-3.5 Turbo
Clinical KG	59.9	62.9	62.8	62.3	60.1	62.5	53.1	24.5	41.6	51.4	74.71
Medical Genetics	64.0	57.0	62.7	67.0	65.0	64.7	58.0	27.7	50.3	52.0	74.00
Anatomy	56.5	55.6	57.5	55.8	58.5	55.8	54.1	35.3	46.4	49.4	65.92
Pro Medicine	60.4	59.4	63.5	61.4	60.5	62.7	58.8	17.4	27.9	53.3	72.79
College Biology	59.0	62.5	64.3	66.9	60.4	64.8	58.1	30.3	44.4	50.7	72.91
College Medicine	54.7	57.2	55.7	58.0	56.5	56.3	48.6	23.3	30.8	49.1	64.73
MedQA	50.6	42.0	50.6	51.1	49.5	50.8	40.1	25.5	41.6	42.5	57.71
MedQA 5 opts	42.8	40.9	43.6	45.2	43.2	44.3	33.7	20.2	28.1	33.9	50.82
PubMedQA	77.5	75.7	77.5	77.7	77.5	77.8	73.6	72.9	74.9	76.8	72.66
MedMCQA	48.1	46.1	48.8	48.7	48.1	48.6	37.0	26.6	41.3	37.6	53.79
Average	57.3	55.9	58.7	59.4	57.9	58.8	51.5	30.4	42.7	49.7	66.0

TABLE I BENCHMARKING RESULTS *Note. Adapted from BioMistral-7B: Mistral 7B based LLM for medical domains, by Anakin (2023). https://anakin.ai/blog/biomistral-7b/. Copyright 2023 by Anakin AI, Inc.*

PubMedBert, a biomedical language model that translates medical text into 768-dimensional vector space, accomplishes the above. Its strength in handling complex medical terminology makes it an extremely effective tool to capture healthcare data.

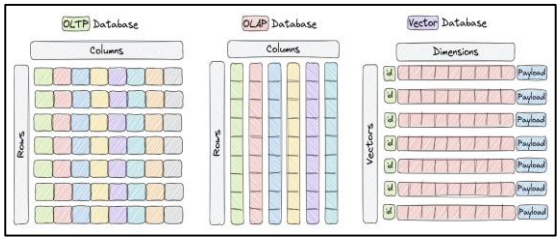

Figure 1. OLTP v/s OLAP v/s Vector Database https://qdrant.tech/documentation/overview/

1) Vector Databases: Vector databases efficiently handle high-dimensional vectors, unlike traditional relational databases as shown in Fig.1. They support complex data representations essential for recommendation systems, natural language processing, and image recognition.

2) Qdrant: Qdrant is a vector similarity search engine designed for managing and querying high-dimensional vectors. It utilizes advanced indexing techniques such as HNSW and Product Quantization for Approximate Neighbors, along with distance measures like Euclidean Distance, Cosine Similarity, and Dot Product. Qdrant enhances search functionality by managing vectors and additional data, and offers a userfriendly API for comprehensive results.

IV. METHODOLOGY

A. BioMistral-7B: Architecture, Fine-tuning, and Evaluation

1) Mistral-7B-v0.1 Architecture: The Mistral-7B model exemplifies a cutting-edge transformer design strategy that includes several advanced techniques to optimize sequence processing and efficiency. A key innovation is the Sliding Window Attention mechanism, which restricts each token's attention to a fixed-size window of preceding tokens. This approach drastically reduces computational complexity and memory usage, making it highly effective for managing extensive sequences. Complementing this mechanism is a rolling buffer cache, which dynamically updates and manages memory by overwriting outdated data, thus maintaining efficiency during processing. With a substantial window size of 4096 tokens, 32 layers, and 32 attention heads, Mistral7B can theoretically attend to an impressive 131,000 tokens, greatly enhancing its capability to handle and process long and complex sequences. These architectural choices not only improve the model's speed and scalability but also its overall performance, enabling it to efficiently process intricate and lengthy data with high precision and responsiveness [14].

Algorithm 1 Sliding Window Attention with Rolling Buffer Cache

Require: L (sequence length), W (window size), k (number of layers)
Ensure: Attention spans up to $k \times W$ tokens
Initialize *cache* ← empty array of size W
for $i \leftarrow 1$ to L **do**
  current token ← token at position i
  attention _window ← tokens from position $\max(1, i - W)$
  $h_i \leftarrow$ compute attention for *current _token* over *attention window*
  $cache[i \bmod W] \leftarrow h_i$
  for *layer* ← 1 to k **do**
    $h_i^{(layer)} \leftarrow$ attend to previous layer's tokens
    $h_i^{(layer)} \leftarrow$ shift window forward by W
  end for
end for

Algorithm 2 Pre-fill and Chunking for Sequence Generation

Require: *prompt*, W (window size)
Ensure: Efficient memory usage during sequence generation
Initialize *cache* ← empty array of size W
chunks ← split *prompt* into chunks of size W
for each *chunk* in *chunks* **do**
  cache ← pre-fill with current *chunk*
  Compute attention over *cache* and *chunk*
  cache ← update with new keys and values
end for
while generating next tokens
  current token ← next token to predict
  attention window ← tokens from cache and chunk
  $h_i \leftarrow$ compute attention for *current _token* over *attention window*
  $cache[i \bmod W] \leftarrow h_i$
end while

2) Fine-tuning: Mistral 7B Instruct v0.1 was transformed into BioMistral-7B using the AdamW optimizer on the Jean Zay HPC with specialized techniques like Grouped-Query Attention and Sliding Window Attention to increase training efficiency. The investigation focused on improving out-of-domain generalization and performance through model merging techniques such as SLERP and TIES, which mix generic and domain-specific models. Bits and Bytes (BnB) and Activation-aware Weight Quantization (AWQ) are two examples of quantization algorithms that were used to reduce memory consumption and enable deployment on smaller devices with minimal performance loss.

Fig.2 shows training loss during the further pre-training of Mistral 7B Instruct v0.1 on PubMed Central.

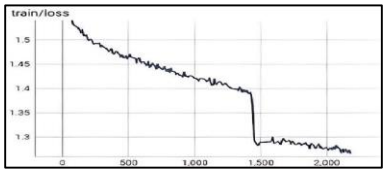

Figure 2. BioMistral 7B loss. Reprinted from "BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains" by Labrak, Y., Bazoge, A., Morin, E., Gourraud, P-A., Rouvier, M., & Dufour, R., 2024, arXiv preprint arXiv:2402.10373v2.

3) Evaluation: The official documentation states that a strong test approach was undertaken on the performance of the BioMistral 7B model [15]. For this purpose, a variety of known downstream medical reasoning tasks in English were obtained by employing four large-scale medical corpora, namely: MMLU, MedQA, MedMCQA, and PubMedQA. These resources represent real-life situations that occur in the clinic and contain a wide variety of common medical domains ranging from anatomy to clinical cases, and genetics. While MedQA tests wide-ranging medical knowledge, such as medication dosages and the symptoms of diseases, it is based on the US Medical Licensing Examination (USMLE). Differently, PubMedQA tests medical reasoning by using PubMed and MedMCQA, which bases its approach on Indian medical entrance examinations (AIIMS/NEET) but spans a broad range of health care.

This assessment was based on a GPT-4 medical evaluation template and utilized the instruction-prompting approach. Every task was a multiple choice and the model predicted the answer based on the choices made. SFT was coupled with supervised fine-tuning, and the QLoRa approach was primarily used for the BioMistral 7B model mainly for resource efficiency through 8-bit quantization.

Calibration ensures that the predicted probabilities of the model match up with the actual outcomes. By employing the Expected Calibration Error (ECE) metric, we can measure how confident the model was with its predictions and how it compared the achieved confidence level with actual performance.

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} . |acc(B_m) - conf(B_m)| \quad (1)$$

B. Modeling Bio-med Application Architecture

To customize our LLM for diagnosis, the foundation of medical knowledge, we examined classic medical literature. These texts were transformed into semantic vector embeddings, which were stored in Qdrant as shown in Fig.2. The texts we studied cover Harrison's Principles of Internal Medicine, First Edition, The Merck Manual of Diagnosis and

Therapy, The Oxford Handbook of Clinical Medicine, and The Medical Dictionary by Gale.

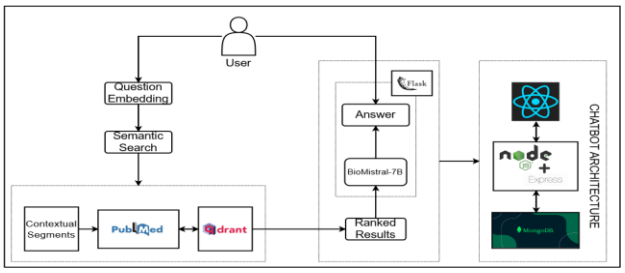

Figure 3. Architecture Overview of the Diagnostic Chatbot

The implementation of the Conversational Retrieval Chain enhances the contextual correctness and performance of the LLM. The system fetches the document system search from the vector database near the embeddings of the user query. To provide the most accurate and relevant response, it fetches documents with relevant information. The LLM can concentrate on the most relevant facts as it only gives context in the form of top-matching document embeddings which exclude unnecessary processing. Receiving user queries and conversation history helps the LLM to maintain the flow of the conversation and make more intelligent responses by retaining context. The framework leverages the use of LangChain and Llama CPP to integrate the LLM, vector database, and CRC. Llama CPP performs all operations on the components and LangChain controls the overall conversation process that maximizes response quality as well as retrieval efficiency.

C. Application Workflow and System Integration

The application workflow is engineered for the seamless integration of multiple technologies to deliver a productive and intuitive chatbot experience for medical diagnostics. The architecture uses Flask for creating and managing endpoints to facilitate interaction between the backend and the LLM. Flask receives and relays API requests to the LLM for processing.

Compared to traditional chatbot approaches, our proposed system has several advantages; the traditional methods usually work based on keyword-based processing, which gives many scalability issues and needs manual updates; meanwhile, our approach uses NLP with Retrieval-Augmented Generation (RAG) for giving context-aware responses while offering unlimited scalability through dynamic integration into BioMistral-7B advanced architectures such as Flask and Express.js [16].

As shown in Table II, whereas traditional systems have to be borrowed from predefined datasets, our chatbot learns with every interaction and automatically improves on feedback and contextual retrieval.

Aspect	Traditional Approach	Proposed Approach
How do they work?	Keyword-based processing	Leverages Natural Language Processing (NLP) with RAG (Retrieval-Augmented Generation) for context-aware responses
How scalable are they?	Limited scalability, often requiring manual updates	Unlimited scalability through dynamic integration of BioMistral-7B and advanced architectures like Flask and Express.js
How are they updated?	Must be trained explicitly on predefined datasets	Learns from each interaction, continuously improving through feedback and contextual retrieval
How do customers interact with them?	Button-focused interaction, primarily through text	Supports both text and voice interactions via React Speech Recognition, enhancing user engagement

Tech Stack	Basic chatbot frameworks (e.g., AI/ML, rulebased systems)	Advanced tech stack including Flask, Express.js, ReactJS, Tailwind CSS, and BioMistral-7B for superior performance
Context Management	Minimal to no context retention, often leading to disjointed conversations	Utilizes Conversational Retrieval Chain (CRC) to maintain context and provide coherent responses

TABLE II COMPARISON OF TRADITIONAL APPROACH vs PROPOSED APPROACH

As shown in Table II, whereas traditional systems have to be borrowed from predefined datasets, our chatbot learns with every interaction and automatically improves on feedback and contextual retrieval.

Third, customer interactions are upgraded; whereas normally, the traditional chatbots take input only by button clicks or only texts, our system allows for both text and voice interactions with React Speech Recognition. This highly increases user engagement. Our tech stack consists of Flask, Express.js, ReactJS, Tailwind CSS, and BioMistral-7B, ensuring outstanding performance. However, while traditional approaches retain very little context resulting in very disjointed conversations, our implementation uses a CRC to preserve context, thus resulting in coherent answers.

A responsive user interface is developed using ReactJS for frontend. The UI boasts a contemporary and attractive appearance, thanks to Tailwind CSS styling. With voice prompts, React Speech Recognition improves user interaction, allowing users to easily give queries and receive responses. User authentication functions, such as sign-in and Login, are driven by MongoDB. Effective management of user data ensures secure interactions and personalized chatbot experiences.

V. RESULT ANALYSIS AND DISCUSSIONS

This research was able to establish the chatbot architecture of appropriate design showing good execution intended for effective patient triage and diagnostic support. The conclusion drawn from this study supports the possibility of practical applicability of AI models in the most critical healthcare contexts, especially concerning ER triage procedures. These results are in accordance with previous studies. Such adaptability of this chatbot has been reinforced by its acceptance to handle a wide array of questions, ranging from general queries to core medical and triage-based inquiries. It has tremendous potential as a useful triage tool for patients and to support healthcare workers.

As illustrated in Fig.4, the responsive interaction of the chatbot shows how its architecture effectively runs and presents the information with minimum distortion. As illustrated in TABLE IV, the evaluation results consisting of a response time between 22 and 35 seconds, with an average response time of 28 seconds, and BLEU and ROUGE scores representing decent responses, authenticate the good performance of the chatbot, when tested on low-end hardware configurations.

These metrics indicate the possibility of an AI-based chatbot in improving diagnosis accuracy and patient engagement in healthcare settings. The ability to deploy and test the chatbot will highlight its promise in being a valuable resource in this field of medical assistance and patient triaging.

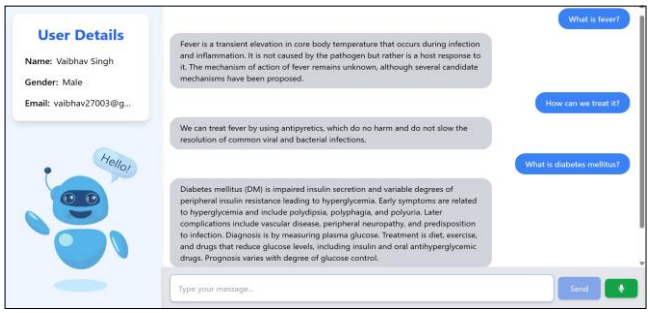

Figure 4. Chatbot-driven patient assistance

VI. CONCLUSIONS AND FUTURE WORKS

In conclusion, our work was able to show the potentiality of using state-of-the-art biological language models, as in the case of BioMistral-7B, along with an AI-based system, not only for diagnostic purposes but also for triaging. Qdrant and CRC's mechanism came very close to improving the task's context awareness and increased the prediction accuracy as well within the chatbot. This makes it an easy-to-use, real-time medical support application in healthcare settings, as shown through its demonstrable proof in low-end hardware configurations. This is consistent with earlier studies demonstrating the advantages of AI-powered systems in critical healthcare applications, such as ER triage. The study therefore demonstrates the usefulness of AI in the advancement of medical diagnosis, with substantial benefits toward improved healthcare procedures and outcomes.

In the future, techniques like query expansion and filtered vector searches will be implemented so that retrieval will become more refined, allowing the system to pull up the most contextually relevant documents for each query. This will lead to better alignment between user questions and the most accurate medical information, ultimately improving diagnostic accuracy. Additionally, the system's ability to handle complex queries and multi-turn conversations will be enhanced with adaptive query management, ensuring that it stays on track and provides up-to-date information throughout longer interactions. Dynamic context updates will help keep the information relevant in real time, especially for medical diagnostics. Moreover, interactive features like personalized doctor recommendations, tailored to the patient's location and preferences, will further enhance the user experience. This customization will make the system even more patient-centric, offering practical, personalized healthcare support.

Table	Description	Formula
BLEU	Measures the precision of n-grams between the generated text and reference text. Focuses on exact matches.	BLEU = $\sqrt[4]{p1.p2.p3.p4}$, where p1, p2, p3, p4 represent unigram, bigram, trigram, and 4-gram precision scores respectively.
ROGUE	Evaluates recall by comparing the overlap between n-grams in the reference and generated text.	ROUGE-1 Recall = $\frac{count\ of\ word\ matches}{count\ of\ words\ in\ referen}$

	Primarily used to assess recall-oriented tasks.	ROUGE-1 Precision = $\frac{count\ of\ word\ matches}{count\ of\ words\ in\ generat}$ ROUGE-1 F1 = 2. $\frac{precision.recall}{precision+recall}$
RESPONSE TIME	Measures how long the chatbot takes to respond to a query. It assesses the system's efficiency.	Response Time = $T_{end} - T_{start}$

TABLE III EVALUATION METRICS

Query	Bot Response	BLEU Score	ROUGE Score	Response Time
What are the common symptoms of the flu?	Fever, cough, sore throat, runny nose, body aches, headache, chills, fatigue.	0.33	0.55	35 seconds
What are the treatment options for Type 2 Diabetes?	There are several treatment options for type 2 diabetes including lifestyle changes, oral medications, insulin therapy, and a combination of these therapies.	0.46	0.41	25 seconds
have a severe headache, nausea, andblurred vision. Should I seek immediate medical attention?	Yes, you should seek immediate medical attention because these are symptoms of a migraine or other neurological disorders that require prompt treatment.	0.36	0.60	32 seconds
Can you provide guidance for managing a rare allergic reaction to a medication I took last week?	The most important thing you can do is stop taking the medication immediately. If you're still having an allergic reaction, you should go to the emergency room right away.	0.06	0.15	28 seconds
What lifestyle changes can help improve cardiovascular health?	Lifestyle changes that can help improve cardiovascular health include weight loss, exercise, smoking cessation, and a healthy diet.	0.18	0.12	22 seconds

TABLE IV AI CHATBOT EVALUATION RESULT SYSTEM CONFIG - AMD RYZEN 5 (6 CORES, 2.1GHZ), CPU - 8GB RAM

REFERENCES

[1] 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.

[2] M. Gams, M. Smerkol, P. Kocuvan, and M. Zadobovsek, "Developing a˙ medical chatbot: Integrating medical knowledge into gpt for healthcare applications," in *Intelligent Environments 2024: Combined Proceedings of Workshops and Demos & Videos Session*, pp. 88–97, IOS Press, 2024.

[3] X. Wang, S. M. Abubaker, G. T. Babalola, and S. Tulk Jesso, "Codesigning an ai chatbot to improve patient experience in the hospital: A human-centered design case study of a collaboration between a hospital, a university, and chatgpt," in *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp. 1–10, 2024.

[4] F. Holderried, C. Stegemann-Philipps, L. Herschbach, J.-A. Moldt, A. Nevins, J. Griewatz, M. Holderried, A. Herrmann-Werner, T. FestlWietek, M. Mahling, *et al.*, "A generative pretrained transformer (gpt)– powered chatbot as a simulated patient to practice history taking: Prospective, mixed methods study," *JMIR medical education*, vol. 10, no. 1, p. e53961, 2024.

[5] S. Tyler, M. Olis, N. Aust, L. Patel, L. Simon, C. Triantafyllidis, V. Patel, D. W. Lee, B. Ginsberg, H. Ahmad, and R. J. Jacobs, "Use of artificial intelligence in triage in hospital emergency departments: A scoping review," *Cureus*, vol. 16, no. 5, p. e59906, 2024.

[6] M. R. King, "The future of ai in medicine: a perspective from a chatbot," *Annals of Biomedical Engineering*, vol. 51, no. 2, pp. 291–295, 2023.

[7] M. A. Quidwai and A. Lagana, "A rag chatbot for precision medicine of multiple myeloma," *medRxiv*, pp. 2024–03, 2024.

[8] C.-M. Chan, C. Xu, R. Yuan, H. Luo, W. Xue, Y. Guo, and J. Fu, "Rq-rag: Learning to refine queries for retrieval augmented generation," *arXiv preprint arXiv:2404.00610*, 2024.

[9] B. Kang, J. Kim, T.-R. Yun, and C.-E. Kim, "Prompt-rag: Pioneering vector embedding-free retrieval-augmented generation in niche domains, exemplified by korean medicine," *arXiv preprint arXiv:2401.11246*, 2024.

[10] S. Gururangan, A. Marasovic, S. Swayamdipta, K. Lo, I. Beltagy,´ D. Downey, and N. A. Smith, "Don't stop pretraining: Adapt language models to domains and tasks," *arXiv preprint arXiv:2004.10964*, 2020.

[11] M. R. Haque and S. Rubya, "An overview of chatbot-based mobile mental health apps: insights from app description and user reviews," *JMIR mHealth and uHealth*, vol. 11, no. 1, p. e44838, 2023.

[12] J. Zhang, Y. J. Oh, P. Lange, Z. Yu, and Y. Fukuoka, "Artificial intelligence chatbot behavior change model for designing artificial intelligence chatbots to promote physical activity and a healthy diet," *Journal of medical Internet research*, vol. 22, no. 9, p. e22845, 2020.

[13] Y. Perlitz, A. Gera, O. Arviv, A. Yehudai, E. Bandel, E. Shnarch, M. Shmueli-Scheuer, and L. Choshen, "Benchmark agreement testing done right: A guide for llm benchmark evaluation," *arXiv preprint arXiv:2407.13696*, 2024.

[14] A. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. Chaplot, D. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, *et al.*, "Mistral 7b. arxiv 2023," *arXiv preprint arXiv:2310.06825*.

[15] Y. Labrak, A. Bazoge, E. Morin, P.-A. Gourraud, M. Rouvier, and R. Dufour, "Biomistral: A collection of open-source pretrained large language models for medical domains," *arXiv preprint arXiv:2402.10373*, 2024.

[16] J. Praveen Gujjar, H. R. Prasanna Kumar, and M. S. Guru Prasad, "Advanced nlp framework for text processing," in *2023 6th International Conference on Information Systems and Computer Networks (ISCON)*, pp. 1–3, IEEE, 2023.