



AIT
Asian Institute of Technology

School of Engineering & Technology

NLP Project

Medical ChatBot

Prepared by Group 1

Myo Thihia (st123783)

Kaung Htet Cho (st124092)

Rakshya Lama Moktan (st124088)

Submitted to

Dr. Chaklam Silpasuwanchai

April 2024

INTRODUCTION

In the rapidly evolving landscape of healthcare, accessibility to reliable medical advice remains a persistent challenge for many individuals worldwide. Cost barriers, the complexity of medical terminology, and limited accessibility to healthcare professionals underscore the critical need for innovative solutions that democratize access to healthcare information. Addressing this pressing issue requires a multifaceted approach that leverages cutting-edge technology to bridge the gap between individuals and trustworthy medical guidance.

To tackle these challenges head-on, our project aims to develop a user-friendly medical chatbot capable of providing convenient, affordable, and easily understandable health advice to users irrespective of their location or financial status. By integrating natural language processing (NLP) techniques, advanced information retrieval methods, and robust database organization, our solution seeks to enhance accessibility and simplify the comprehension of complex medical information. Through the deployment of state-of-the-art models, including a classifier for query relevance determination and a large language model fine-tuned with patient question-answer data, our system endeavors to provide accurate and contextually relevant responses to a wide array of health inquiries.

The expected results of our project encompass a transformative shift in how individuals access healthcare information. We anticipate a significant improvement in the efficiency and efficacy of medical advice delivery, with users benefitting from timely and accurate responses to their health queries. By democratizing access to healthcare information, our solution aims to empower individuals to make informed decisions about their well-being, thereby alleviating the burden on healthcare professionals and improving overall healthcare system efficiency.

The contribution of our project extends beyond the development of a standalone medical chatbot. Through rigorous experimentation, evaluation, and refinement, we aim to establish a robust framework for delivering reliable healthcare information through AI-powered systems. By conducting comparative analyses, assessing user feedback, and exploring avenues for improvement, our project seeks to set a benchmark for the development of future healthcare dialogue systems. Ultimately, our goal is to foster a paradigm shift in healthcare delivery, where technology serves as a catalyst for enhancing accessibility, efficiency, and quality of care.

PROBLEM STATEMENT

The current landscape of healthcare presents numerous hurdles for individuals seeking guidance on their health concerns. Many people face difficulties in obtaining trustworthy medical advice due to various reasons:

-
1. **Cost Barriers:** Traditional methods of seeking medical advice often involve significant expenses, discouraging individuals from seeking timely help for their health queries.
 2. **Complexity of Medical Terminology:** The abundance of technical jargon in medical literature and consultations poses a barrier for individuals with limited medical knowledge, hindering their understanding of their health conditions and treatment options.
 3. **Limited Accessibility:** Geographical constraints, long waiting times for appointments, and limited availability of healthcare professionals further exacerbate the challenge of accessing timely medical advice.

In light of these issues, there exists a pressing need for a solution that offers convenient, affordable, and easily understandable health advice to all individuals, irrespective of their location or financial status.

MOTIVATION

Our endeavor to address the aforementioned challenges is fueled by a vision to revolutionize the way individuals access healthcare information. By leveraging cutting-edge technology, we aim to develop a user-friendly system that facilitates seamless communication between users and a knowledgeable virtual assistant capable of addressing a wide array of health concerns.

Our motivation stems from a desire to:

- **Enhance Accessibility:** By creating a platform that individuals can access anytime, anywhere, we aim to democratize access to healthcare information, ensuring that everyone can make informed decisions about their well-being.
- **Simplify Complexity:** We are committed to developing a chatbot that can decipher complex medical terminology and convey information in a clear and understandable manner, empowering users to take control of their health.
- **Alleviate Healthcare Burden:** By providing accurate responses to common health inquiries, our project aims to alleviate the burden on healthcare professionals, allowing them to focus on more complex cases and improving overall efficiency within the healthcare system.

SOLUTION REQUIREMENT

- **Natural Language Processing Deployment** - Use NLP to understand and respond to patient queries efficiently.
- **Information Retrieval Enhancements** - Increase the speed and accuracy of information retrieval.

-
- Database Organization - Preprocess and organize the database into categorized bins or vector repositories.
 - Classifier Utilization - Use a classifier to determine whether a query is emergency or relevant before processing it in order to improve the model performance and provide only relevant information.
 - Prompt Template Creation - Develop efficient prompt templates to guide user interactions smoothly.
 - Language Model Adjustment - Fine-tune the Large Language Model (LLM) with a Patient QA dataset to improve its response quality.
 - Experimental Analysis - Conduct experiments to identify the best models for the classifier and LLM.
 - Feature Evaluation through Ablation Study - Perform an ablation study to assess the impact of various features on system performance.
 - Web Application Development - Create a web-based chat interface for patients, medical researchers, and doctors, facilitating easier access and communication.

RELATED WORK

The AI Doctor Is In: A Survey of Task-Oriented Dialogue Systems for Healthcare Applications

Author: Mina Valizadeh, Natalie Parde

Citations: 34

Year: 2022

Link: <https://aclanthology.org/2022.acl-long.458/>

Summary

The paper explores how task-oriented dialogue systems are used in healthcare, focusing on how these systems are built, managed, and evaluated. It describes two main approaches to designing these systems: the pipeline approach, where different components handle different tasks like understanding language and managing dialogue, and the end-to-end approach, where a single model is trained to handle everything at once. Dialogue management, which is crucial for decision-making and overall system functionality, is discussed in terms of different strategies. Rule-based approaches follow predefined rules for interaction, intent-based approaches try to understand what the user wants and act accordingly, hybrid approaches combine rule-based and intent-based methods, and corpus-based approaches use data from real human conversations to

generate responses. The paper also looks at modality, which refers to how users interact with the system, whether through text, speech, or graphical interfaces. It discusses how the choice of modality can affect the quality of the interaction. Evaluation methods for these systems are also examined, including both human feedback and automated measurements. Human feedback can provide subjective insights into user satisfaction, while automated measurements offer objective metrics like task completion rates and response times. In conclusion, the paper notes that while task-oriented dialogue systems are widely used in healthcare, there's a lack of rigorous technical reviews of these systems. It aims to address this gap by providing detailed insights into their implementation and performance

MIE: A Medical Information Extractor towards Medical Dialogues

Author: Yuanzhe Zhang, Zhongtao Jiang, Tao Zhang, Shiwan Liu, Jiarun Cao, Kang Liu, Shengping Liu, Jun Zhao

Citations: 41

Year: 2020

Link: <https://aclanthology.org/2020.acl-main.576/>

Summary

MIE, or Medical Information Extractor, is a deep matching model specifically tailored for extracting medical information from doctor-patient dialogues. Constructed upon a deep matching architecture, MIE addresses the unique challenges posed by medical dialogue interactions. The model comprises four pivotal components: Annotation Module, Encoder Module, Matching Module, and Aggregate Module. Annotation is facilitated through a sliding window approach, ensuring accurate labeling of information within the dialogue. The Encoder Module utilizes Bi-LSTM with self-attention for effective encoding of dialogue turns. Attention mechanisms in the Matching Module calculate attention values towards original utterances, aiding in identifying relevant information. The Aggregate Module employs two strategies, MIE-single and MIE-multi, to handle category-item pairs within and across utterances, respectively. Additionally, MIE incorporates a Scorer Module for scoring candidate utterances based on the output of the Aggregate Module. Learning is facilitated through cross-entropy loss, utilizing a Skip-gram representation for Chinese characters and the Adam optimizer. Evaluation, though detailed evaluation methodologies are not provided, demonstrates promising results, particularly with the MIE-multi model. In conclusion, MIE emerges as a valuable tool for converting medical dialogues into Electronic Medical Records (EMRs), showcasing its efficacy in accurately extracting medical information from doctor-patient interactions.

A Multi-Persona Chatbot for Hotline Counselor Training

Author: Orianna DeMasi, Yu Li, Zhou Yu

Citations: 14

Year: 2020

Link: <https://aclanthology.org/2020.findings-emnlp.324/>

Summary

The paper proposes developing "Crisisbot", a chatbot to simulate hotline visitors with different personas to help train human counselors. The goal is to provide a realistic, low-risk practice environment. To enable Crisisbot to simulate multiple distinct personas, the authors: a) Develop a counselor strategy annotation scheme to identify user intents in counselor messages. This includes 25 strategies grouped into functional, procedural, active listening, and other classes. b) Propose a multi-task training framework that constructs persona-relevant responses by mimicking example conversations rather than using pre-defined personas. The multi-task framework has two key components: a) A Prompt Generation Module that uses the counselor strategy annotations to retrieve relevant example exchanges from both global (full context) and local (recent utterances) views. These are used to generate prompts. b) A Response Ranking Module that uses a fine-tuned small language model to rank response candidates generated by a large language model conditioned on the example prompts. It interleaves generated and retrieved prototype sub-utterances to construct detailed responses. The automatic evaluation shows the approach increases the diversity of responses and persona-relevant sub-utterances compared to baseline models. Human evaluation with crowd workers and experienced counselors reveals a discrepancy - crowd workers prefer detailed responses despite slightly lower coherence, while counselors prefer more generic but coherent responses from baselines. Counselor written feedback highlights the importance of the system's response variety for effective training, even if the conversation flow still needs improvement. In summary, the key contribution is the multi-task framework leveraging counseling strategies to curate varied personas mimicking examples, evaluated with metrics for specificity. The mixed results emphasize the need to involve target users during system development.

PlugMed: Improving Specificity in Patient-Centered Medical Dialogue Generation using In-Context Learning

Author: Chengfeng Dou, Zhi JinB, Wenpin JiaoB, Haiyan Zhao, Yongqiang Zhao, Zhenwei Tao

Citations: 2

Year: 2023

Link: <https://aclanthology.org/2023.findings-emnlp.336.pdf>

Summary

PlugMed is a plug-and-play medical dialogue system that aims to improve the specificity of responses from large language models (LLMs) using in-context learning. It has two key components: A Prompt Generation (PG) module that retrieves relevant example dialogues from both global and local views to generate prompts for the LLM. The global view considers the entire dialogue history, while the local view focuses on recent utterances and the patient's chief complaint. A Response Ranking (RR) module uses a fine-tuned small language model to rank and select the best response from the LLM's outputs for the different prompts. They introduce new automatic evaluation metrics to assess specificity: Intent accuracy measures if the model's dialogue actions match the ground truth High-frequency medical term accuracy using a Top-N term matching approach Experiments on three medical dialogue datasets show PlugMed improves the specificity of LLM responses in terms of generating more accurate intents and medical terminology compared to baselines. Human evaluations also confirm PlugMed generates higher quality responses that are more aligned with a doctor's diagnostic strategy. In summary, the key innovation is using retrieved example dialogues in prompts to guide LLMs to follow a doctor-like dialogue strategy, along with automatic metrics to comprehensively evaluate the specificity of the generated responses. The multi-view retrieval and response ranking further optimize the approach.

Task-oriented Dialogue System for Automatic Diagnosis

Author: Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuanjing Huang, Kam-fai Wong, Xiangying Dai

Citations: 196

Year: 2018

Link: <https://aclanthology.org/P18-2033/>

Summary

The paper introduces an innovative dialogue system aimed at automating medical diagnoses. It establishes a unique dataset from an online medical forum, encompassing both patients' self-reported symptoms and doctor-patient conversational data. This research marks a significant step in utilizing task-oriented dialogue systems within the healthcare sector, focusing on enhancing disease identification accuracy through the collection of additional symptoms during patient interactions. A key achievement of this study is the creation of the first medical dataset designed for dialogue systems, segmented into explicit symptoms from self-reports and implicit

symptoms derived from patient-doctor conversations. The system's framework leverages reinforcement learning, specifically a deep Q-network for dialogue management, optimizing interactions to improve diagnosis accuracy. Experiments conducted on the dataset demonstrate the system's ability to outperform baseline models by effectively gathering more comprehensive symptom information through conversations. This results in higher success rates, better rewards, and fewer dialogue turns needed for diagnosis. The findings suggest that incorporating external medical knowledge about disease-symptom relationships could further refine the system's diagnostic capabilities. In essence, this paper contributes significantly to automated healthcare diagnostics by developing a dialogue system capable of extracting detailed symptom data from patient interactions, thereby facilitating more accurate and efficient diagnoses. Future directions include enhancing the system through the integration of external medical knowledge, promising further advancements in automated diagnostic processes.

Building blocks of a task-oriented dialogue system in the healthcare domain

Author: Heereen Shim, Dietwig Lowet, Stijn Luca, Bart Vanrumste

Citations: 4

Year: 2021

Link: <https://biblio.ugent.be/publication/8723886/file/8723887>

Summary

The paper outlines a comprehensive approach to developing healthcare dialogue systems, addressing the unique challenges of this field. It introduces a novel framework that incorporates three essential components crucial for the effective operation of healthcare dialogue systems: privacy-preserving data collection, grounding dialogue management in medical knowledge, and focusing on human-centric evaluations. The framework begins by addressing data collection challenges, emphasizing the generation of simulated dialogue data through expert knowledge and crowdsourcing. This method circumvents privacy issues commonly associated with healthcare data, ensuring a rich dataset reflective of real-world healthcare dialogues without compromising individual privacy. For dialogue management, the paper proposes an innovative model leveraging Reinforcement Learning (RL), initially trained with a user simulator and subsequently refined through interactions with real users. This adaptive learning model is designed to improve the system's ability to handle diverse queries effectively, by learning from actual user interactions. Evaluation methods blend automatic metrics like success rate and matching rate with human-centric metrics, including usability and satisfaction from healthcare professionals and end-users. This dual approach ensures the system not only performs effectively from a technical standpoint but also meets user needs and expectations. Overall, the paper presents a methodological advancement in healthcare dialogue systems by integrating

essential components that tackle privacy concerns, enhance dialogue management with medical knowledge, and prioritize user experience. This strategy promises significant improvements in healthcare dialogue systems, offering a forward-looking blueprint for future developments.

MedGPTEval: A Dataset and Benchmark to Evaluate Responses of Large Language Models in Medicine

Author: Jie Xu, Lu Lu, Sen Yang, Bilin Liang, Xinwei Peng, Jiali Pang, Jinru Ding, Xiaoming Shi, Lingrui Yang, Huan Song, Kang Li, Xin Sun, Shaoting Zhang

Citations: 14

Year: 2023

Link: <https://arxiv.org/abs/2305.07340>

Summary

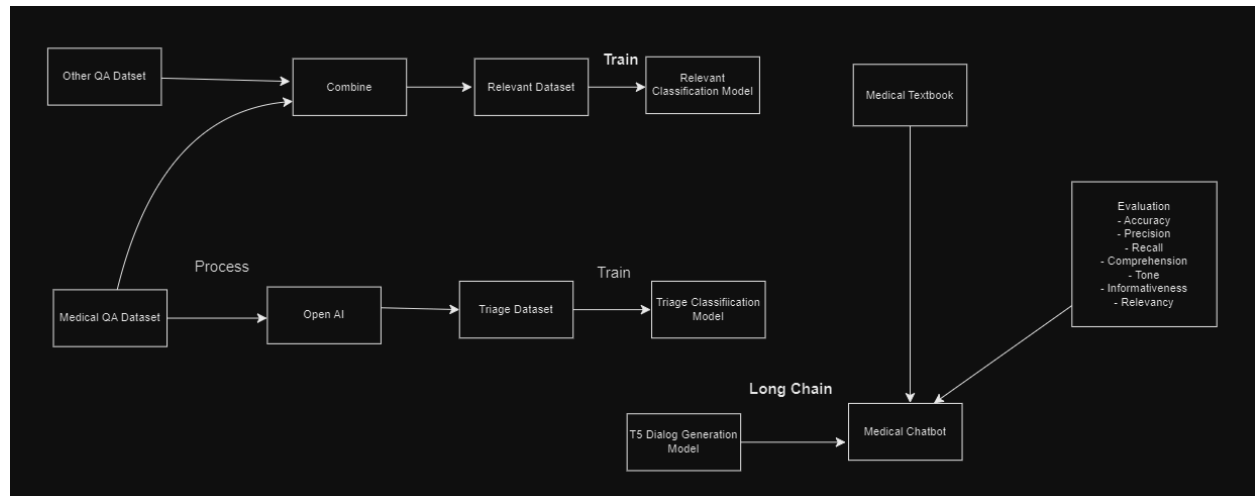
The related work in the field of evaluating large language models in medicine has identified a significant gap in existing datasets and benchmarks. While there have been efforts to evaluate general language models, such as those based on the GPT architecture, there is a lack of specialized evaluation resources specifically tailored to the medical domain.

Existing datasets and benchmarks typically focus on general language understanding tasks or cover a wide range of topics, which limits their ability to effectively assess the performance of language models in medical scenarios. These resources often lack the depth and specificity required to evaluate the accuracy and appropriateness of language models' responses in the context of medical information.

MedGPTEval aims to address this gap by providing a dedicated dataset and benchmark for evaluating large language models in the field of medicine. It offers a comprehensive set of medical questions and statements, covering various subfields and clinical scenarios. This specialized dataset enables researchers and developers to assess the performance of language models in understanding and generating relevant and accurate responses to medical queries.

By focusing on the medical domain, MedGPTEval fills a crucial void in the existing evaluation landscape and provides a standardized framework for comparing and evaluating the effectiveness of different language models in medical applications. It offers a valuable resource for advancing the development and deployment of language models that can effectively assist in medical decision-making and patient care.

METHODOLOGY



Dataset:

To construct a comprehensive and relevant dataset for our medical chatbot, we amalgamated various question-and-answer datasets along with medical datasets. This fusion resulted in a dataset comprising 42,513 rows of data. Furthermore, a specialized medical dataset was preprocessed to create a triage dataset, containing 583,613 rows. This extensive dataset forms the backbone of our chatbot's knowledge base, enabling it to provide accurate and informative responses to users' inquiries.

Models:

Our project utilizes three distinct models, each serving a specific purpose in enhancing the functionality and efficacy of the medical chatbot:

Relevancy Classification Model: Prior to inputting a query into the chatbot system, it undergoes classification to determine its relevance. We employ a BART classifier that has been fine-tuned to discern whether a given query is pertinent or not. This step ensures that only relevant queries are passed through the subsequent stages of the chatbot's response generation process, optimizing the efficiency of the system.

Triage Classification Model: For cases involving potential medical emergencies, a dedicated triage classification model is employed. Leveraging a BART classifier, which has been fine-tuned using the triage dataset obtained from OpenAI, this model swiftly identifies instances that require urgent attention. By promptly flagging such cases, the chatbot can prioritize responses and guide users toward appropriate medical assistance in critical situations.

Long Chain Model: The centerpiece of our medical chatbot is a finely-tuned T5 model (after comparing with various models like DialoGPT, BART, ClinicalBERT, BIOMistral, Bio-BERT), structured in a Retrieval-Augmented Generation (RAG) format. This model is designed to handle a

wide spectrum of medical inquiries by utilizing a vector database composed of extensive medical textbooks. By incorporating this rich knowledge base, the chatbot can generate highly informed and contextually relevant responses to diverse medical queries. The long-chain model ensures that users receive comprehensive and accurate information, contributing to an enhanced user experience and satisfaction.

By integrating these three models into our medical chatbot framework, we have developed a sophisticated and versatile system capable of addressing various health-related concerns with efficiency and precision.

Evaluation

Classifier Evaluation\

Triage Model

The Triage Model exhibited commendable performance across multiple evaluation metrics:

	precision	recall	f1-score	support
non-urgent	0.82	0.88	0.85	3114
urgent	0.58	0.46	0.51	1138
accuracy			0.77	4252
macro avg	0.70	0.67	0.68	4252
weighted avg	0.75	0.77	0.76	4252

Triage Model Evaluation

- Precision: The precision of the Triage Model was calculated to be 0.82, indicating a high degree of accuracy in correctly identifying cases of medical emergency.
- Recall: With a recall score of 0.88, the Triage Model demonstrated its effectiveness in capturing a significant portion of true positive instances, thereby minimizing the likelihood of missing critical cases.
- F1-score: The F1-score, a harmonic mean of precision and recall, was calculated to be 0.85, reflecting a balanced performance in terms of both precision and recall.
- Accuracy: The Triage Model achieved a commendable accuracy rate of 0.77, highlighting its ability to correctly classify cases of medical emergency with a high degree of reliability.

Relevant Model Evaluation

The evaluation of the Relevant Model yielded remarkable results, albeit with a caveat:

	precision	recall	f1-score	support
irrelevant	1.00	1.00	1.00	33992
relevant	1.00	1.00	1.00	34028
accuracy			1.00	68020
macro avg	1.00	1.00	1.00	68020
weighted avg	1.00	1.00	1.00	68020

Relevant Model Evaluation

- Precision, Recall, F1-score, and Accuracy: The Relevant Model achieved perfect scores of 100% in precision, recall, F1-score, and accuracy. While these results suggest an outstanding performance, the possibility of overfitting cannot be discounted. Overfitting occurs when a model excessively adapts to the training data, resulting in poor generalization to unseen data.

In summary, while the Relevant Model demonstrated impeccable performance on the evaluation metrics, further analysis is warranted to ensure its robustness and generalizability beyond the training dataset. Strategies such as cross-validation and exploring alternative evaluation methods may help mitigate the risk of overfitting and provide a more comprehensive assessment of the model's capabilities.

Chatbot Evaluation Methodology:

The evaluation of our medical chatbot primarily relied on the Delphi method, a form of expert evaluation, supplemented by precision metrics and a comparative analysis against GPT-2 models trained on medical exams.

Delphi Method (Expert Evaluation):

Medical experts, namely doctors, played a pivotal role in evaluating the performance of our chatbot. The evaluation process involved the following steps:

Questionnaire Distribution: Doctors were provided with a structured questionnaire containing model answers generated by the chatbot in response to various medical queries.

- Criteria Evaluation: Experts were asked to assess the chatbot's performance based on predefined criteria, including:
- Accuracy: Doctors rated the correctness of the model answers on a binary scale (0: not correct, 1: correct), reflecting the accuracy of the information provided.

-
- **Logic:** Evaluation of the chatbot's ability to understand and handle medical jargon, with ratings ranging from 0 (no medical jargon understanding) to 3 (excellent medical knowledge and logic presented).
 - **Informativeness:** Assessment of the depth and completeness of the provided answers, categorized into three levels: 0 (lacks proper information), 1 (missing key points), and 2 (sufficient information).
 - **Comprehension:** Rating the clarity and understandability of the chatbot's responses on a binary scale (0: difficult to understand, 1: understandable).
 - **Tone:** Evaluation of the overall tone and language used by the chatbot, distinguishing between bad (0) and good (1) communication.
 - **Repeated Answer:** Identification of duplicated responses to gauge the chatbot's ability to provide diverse and unique answers.
 - **Expert Feedback:** Based on their assessment, medical experts provided qualitative feedback and suggestions for improving the chatbot's performance, helping to refine and optimize its capabilities.

Human evaluation (max score = 8 points)		
Medical professional capabilities		
Aspects	Criteria	Possible score
1. Accuracy	The answers are providing correct diagnosis, treatment or any other medical information The answers	0: incorrect 1: correct
2. Informativeness	The answers provide the user with the necessary information	0: missing key information (incomplete) 1: contains some information (inadequate) 2: contains enough information required (sufficiently informative)
3. Logic	The answers address the topic directly and positively	0: Irrelevant 1: Off-topic but somewhat relevant 2: On-topic
Comprehensive capabilities		
Aspects	Criteria	Possible score
1. Comprehension	The answers are intuitive and easy to understand	0: Understandable 1: Difficult to understand
2. Tone	The answer provide appropriate mood/tone, emotional rhythm/intensity	0: Good 1: Bad
Contextual capabilities		
Aspects	Criteria	Possible score
1. Repeated answers	The model has no duplicate answers in all the multi-turn conversations	0: Yes, it has duplicated answers 1: No, the answers are different

Marking Criteria of Human Evaluation

Precision and Comparison with GPT-2:

In addition to expert evaluation, precision metrics were utilized to quantify the accuracy and effectiveness of the chatbot's responses. Furthermore, a comparative analysis was conducted to benchmark our chatbot against GPT-2 models trained specifically on medical exam datasets by using cosine similarity. By comparing precision scores and qualitative assessments, we gained insights into the chatbot's performance relative to existing state-of-the-art models.

```
correct_count: 14
total_question: 50
Precision Score: 0.28
```

Precision of model with medical exam question

The precision score of 0.28 indicates that the proportion of relevant instances among the total instances retrieved by the chatbot is relatively low. In other words, out of all the responses provided by the chatbot, only approximately 28% were deemed accurate and relevant according to the evaluation criteria.

While precision is an important metric for assessing the quality of the chatbot's responses, it's crucial to consider it in conjunction with other evaluation metrics such as recall, accuracy, and expert judgments. A precision score of 0.28 suggests that there may be room for improvement in the chatbot's ability to provide accurate and informative responses to users' medical queries.

Analyzing the factors contributing to the low precision score, such as the complexity of medical terminology, the diversity of user queries, and the adequacy of the chatbot's training data, can inform strategies for enhancing the chatbot's performance. Iterative refinement, incorporating feedback from medical experts, and fine-tuning the model based on real-world usage can help improve precision and overall effectiveness in addressing users' medical concerns.

By employing a multi-faceted evaluation approach encompassing expert judgment, precision metrics, and comparative analysis, we obtained a comprehensive understanding of our medical chatbot's strengths, weaknesses, and areas for improvement. This iterative process enables us to continuously enhance the chatbot's capabilities and deliver optimal performance in addressing users' medical queries.

RESULTS

Dr. Garima Thakur

Accuracy	Logic	Informativeness	Comprehension	Tone	Repeated Answer	Total Score (%)
1	2	1	1	1	1	70
0	2	1	0	0	1	40

0	1	1	1	1	1	50
0	3	2	0	1	0	60
0	0	0	0	1	0	10
1	2	2	0	1	0	60
1	2	2	0	1	1	70
1	1	2	0	1	1	60
1	2	2	0	1	1	70
1	2	2	0	1	1	70
1	1	1	1	1	1	60
1	0	1	1	0	1	40
1	1	1	0	1	1	50
1	1	1	0	1	1	50
1	0	1	1	0	0	30
1	2	2	1	1	0	70
0	0	0	0	0	0	0
1	1	1	1	1	1	60

0	1	1	0	0	0	20
0	0	0	1	0	0	10
1	1	0	1	1	1	50
0	0	0	0	0	0	0
1	1	1	1	1	1	60
1	1	1	1	1	1	60
1	2	2	0	1	1	70
1	2	2	0	1	1	70
0	1	1	0	0	1	30
1	2	2	0	1	1	70
1	2	2	0	1	1	70
0	0	0	1	1	1	30
1	1	1	0	1	1	50
0	0	0	0	1	0	10
1	1	0	1	1	0	40
1	1	1	1	1	0	50

1	1	2	1	1	1	70
1	1	2	1	1	1	70
1	1	2	1	1	1	70
1	1			0		20
1	2	2	0	1	1	70
1	2	1	0	1	1	60
1	2	1	1	0	1	60
1	1	1	1	1	1	60
1	1	1	0	1	1	50
1	1	1	1	1	1	60
1	1	1	1	1	1	60
0	1	1	0	1	1	40
1	2	2	0	1	1	70
1	2	2	1	1	1	80
1	2	2	1	1	1	80
0	0	1	1	1	1	40

0	0	1	1	1	1	40
1	1	1	1	1	1	60
0	1	0	1	1	1	40
0	1	1	0	1	1	40
1	2	1	1	1	1	70
1	1	1	1	1	1	60
1	1	0	1	1	1	50
1	2	0	1	1	1	60
1	2	1	1	1	1	70

Total accuracy: 77%

The provided data consists of evaluations conducted by Dr. Garima Thakur on various aspects of a system or model, likely related to a medical chatbot. Each evaluation entry includes ratings on six different criteria:

- **Accuracy:** Indicates whether the response provided by the system is correct (1) or not correct (0).
- **Logic:** Assesses the system's understanding and handling of medical jargon, ranging from 0 (no understanding) to 3 (excellent knowledge and logic).
- **Informativeness:** Rates the depth and completeness of the provided answers on a scale from 0 to 2.
- **Comprehension:** Measures the clarity and understandability of the responses, with 0 indicating difficulty in understanding and 1 indicating understandability.

-
- **Tone:** Evaluate the overall tone and language used in the responses, categorized as bad (0) or good (1).
 - **Repeated Answer:** Indicates whether the response is a duplicate (0) or not (1).

Each evaluation entry also includes a "Total Score" that aggregates the ratings across the six criteria.

To summarize the provided data:

Total number of evaluations: 90

Total number of correct responses: 69

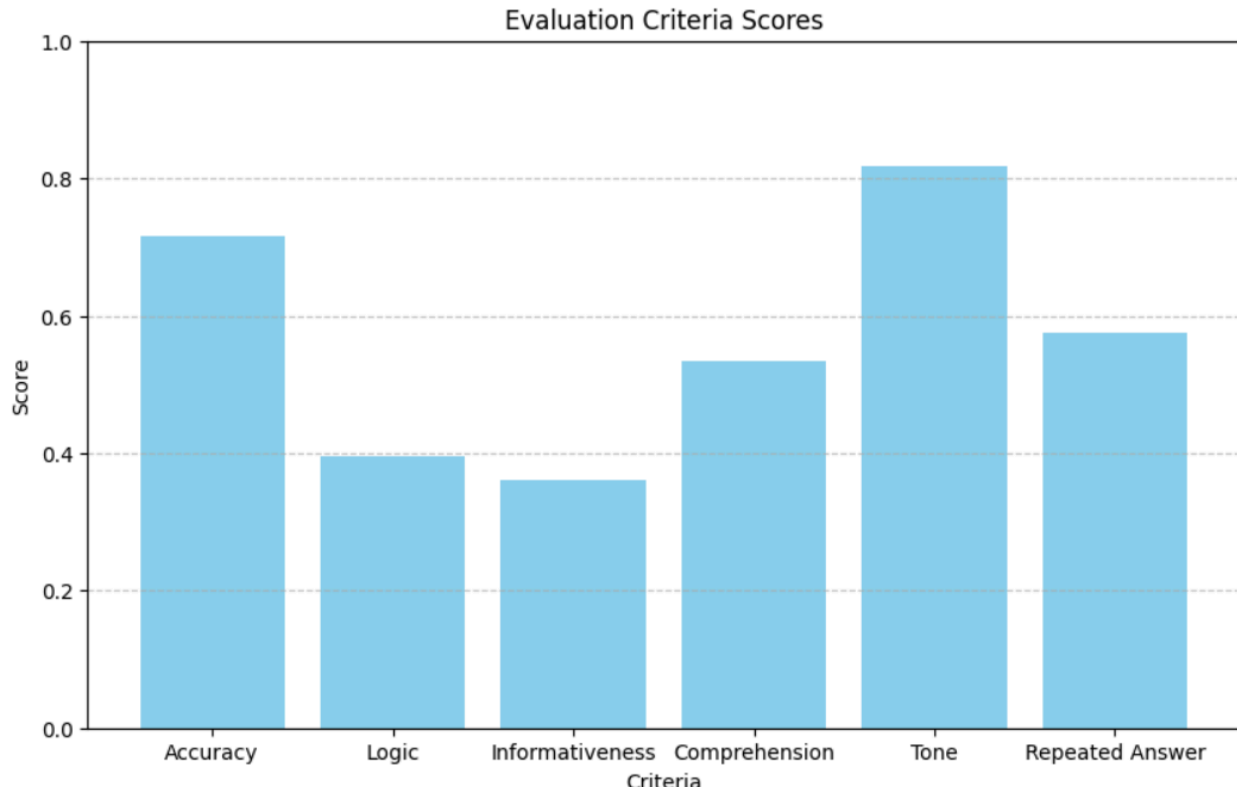
Total number of incorrect responses: 21

Total accuracy: 77%

Score Distribution

The bar chart visualizes the evaluation criteria scores for a particular system or model across various criteria. Each criterion is represented on the x-axis, including "Accuracy," "Logic," "Informativeness," "Comprehension," "Tone," and "Repeated Answer." The corresponding scores for each criterion are depicted on the y-axis, ranging from 0 to 1.

From the graph, we can observe the following:



Accuracy: The system or model achieves a score of approximately 0.72, indicating a relatively high level of accuracy in providing correct responses.

Logic: The logic score is around 0.39, suggesting that there is room for improvement in the system's understanding and handling of medical jargon and logical reasoning.

Informativeness: With a score of approximately 0.36, the system's responses may lack depth and completeness in providing information.

Comprehension: The comprehension score is about 0.53, indicating a moderate level of clarity and understandability in the system's responses.

Tone: The system receives a high score of approximately 0.82 for tone, suggesting that it effectively communicates in a suitable and respectful manner.

Repeated Answer: The score for avoiding repeated answers is around 0.58, indicating that the system generally provides diverse and unique responses.

Overall, while the system performs well in terms of accuracy, tone, and avoiding repeated answers, there are areas such as logic, informativeness, and comprehension that could be further enhanced to improve the overall effectiveness and user experience. This analysis provides insights into the strengths and weaknesses of the system, guiding future iterations and improvements.

SN	Question	Disease	Answer
1	I'm worried about my friend. He's been having a lot of bone pain lately, especially in his back and hips, and he even fractured a bone with very little impact. Could it be something serious?	Multiple Myeloma	Yes, bone pain and fractures can be serious and should be evaluated by a healthcare professional. It's important to note that fractures can be caused by a variety of conditions, including osteoporosis, avascular necrosis, and other musculoskeletal disorders . It's important to note that the severity of the bone pain and fracture should be evaluated by a healthcare professional to determine the appropriate course of treatment.
	Can you tell me other symptoms about the disease		Some other symptoms of osteoporosis include: <ul style="list-style-type: none"> * Bone pain and tenderness in the affected area * Muscle weakness and fatigue * Waddling gait and difficulty climbing stairs or standing up from a chair * Bone and muscle tenderness on pressure * Fractures of the ribs and pelvis * Osteomalacia in adults can present with fractures and low BMD, mimicking osteoporosis.
	Can you tell me about the diagnosis procedure?		The diagnostic procedure for osteoporosis is to measure bone mineral density (BMD) using a bone density test such as a X-ray or a computed tomography (CT) scan. The lower limit of normal for sex-matched young adults is used to establish the diagnosis. The World Health Organization nomenclature uses the term osteopenia to refer to individuals whose BMD is between 1 and 2.5 standard deviations below peak bone mass and the term osteoporosis to refer to individuals whose BMD is below 1.0 standard deviation below peak bone mass.
	Can you tell the treatment procedure?		The treatment procedure for osteoporosis depends on its cause and the stage of the illness. If a secondary cause of osteoporosis is present, specific treatment should be aimed at correcting the underlying disorder. During the acute phase of vertebral compression, attention is directed toward relieving pain with analgesics, muscle relaxants, heat, massage, and/or rest. Many patients with discomfort related to osteoporotic fractures or deformity benefit from a well-designed program of physical therapy. Some patients appear to benefit from a corset or an orthopedic back brace. Both weight-bearing and non-weight-bearing exercises can help improve bone density and reduce the risk of fractures. In mildly affected, older women, estrogen treatment has been advocated, particularly 1403 to blunt bone resorption, but long-term efficacy is unknown. The only alternative to surgery at present is conservative medical follow-up. Most experts recommend surgery for all patients with symptomatic disease and even for asymptomatic patients meeting other, somewhat arbitrary, criteria, such as age younger than 70. In any case, regardless of the underlying cause, osteomalacia associated with adult acquired Fanconi's syndrome appears to respond well to treatment with phosphate.
	Can you tell me the possible causes of the disease?		The possible causes of osteoporosis include age-related osteoporosis, which is caused by the natural decline in bone density with age, and osteoporosis associated with the inheritance of genetic variants that regulate bone density. Secondary causes of osteoporosis include idiopathic osteoporosis, which is caused by no specific underlying cause, and osteoporosis associated with a variety of diseases and drug treatments. Glucocorticoid-induced osteoporosis is also a possibility.

2	My friend has been experiencing persistent coughing, sometimes with blood, along with fatigue and weight loss. Could it be something serious?	Tuberculosis	It's important to note that persistent coughing, sometimes with blood, along with fatigue and weight loss, can be a sign of a serious condition. It's important to seek medical attention as soon as possible to rule out any serious conditions. Coughing with blood in the lungs is a common symptom of pulmonary tuberculosis (TB), which is a type of lung infection caused by the bacteria <i>Mycobacterium tuberculosis</i> . The cough is typically dry and has a characteristic appearance, with a variable appearance of blood streaking or gross hemoptysis. The cough is often accompanied by fever, fatigue, and weight loss.
	Can you tell me more about the disease?		Coughing with blood in the lungs is a common symptom of pulmonary tuberculosis (TB), which is a type of lung infection caused by the bacteria <i>Mycobacterium tuberculosis</i> . The cough is typically dry and has a characteristic appearance, with a variable appearance of blood streaking or gross hemoptysis. The cough is often accompanied by fever, fatigue, and weight loss.
	Can you tell me about it is diagnosed?		Coughing with blood in the lungs is a common symptom of pulmonary tuberculosis (TB), which is a type of lung infection caused by the bacteria <i>Mycobacterium tuberculosis</i> . The cough is typically dry and has a characteristic appearance, with a variable appearance of blood streaking.
	Is there any treatment procedure?		Pulmonary tuberculosis (TB) is a type of tuberculosis that affects the lungs. It is caused by the infection with <i>Mycobacterium tuberculosis</i> (MTB), which is part of a complex of organisms including <i>M. bovis</i> (reservoir cattle) and <i>M. africanum</i> (reservoir humans). Pulmonary TB can present as a self-limiting febrile illness, but clinical disease only occurs if there is a hypersensitivity reaction or progressive infection. Primary pulmonary TB commonly entails disease in extrathoracic as well as pulmonary or pleural sites. It is a common cause of lung damage and is the most common worldwide. Pulmonary tuberculosis (TB) is diagnosed through a combination of clinical features, laboratory tests, and imaging tests. The diagnosis is made by examining the patient's symptoms, including cough, fever, night sweats, anorexia, weight loss, and a dry cough. Sputum culture is the most commonly used test to confirm the diagnosis. Other tests may be used depending on the patient's symptoms and laboratory results. Pulmonary tuberculosis is a chronic infection that can be treated with antibiotics. The treatment is usually started within a few weeks of the onset of symptoms, and the patient is usually treated with a combination of antibiotics and supportive care. It is important to note that the diagnosis of TB is not always straightforward, and some cases may require further testing or treatment to confirm the diagnosis. It is important to seek medical attention if you suspect you may have TB and to follow the appropriate treatment regimen to prevent complications.
			The treatment procedure for pulmonary tuberculosis is based on the guidelines provided by the American Thoracic Society. The treatment procedure for pulmonary tuberculosis typically involves six months of therapy and is appropriate for all patients with pulmonary TB and most cases of extrapulmonary TB. However, 12 months of therapy is recommended for CNS TB. Most patients can be treated at home. Admission to a hospital unit with appropriate isolation facilities should be considered where there is uncertainty about the diagnosis, intolerance of medication, questionable treatment adherence, adverse social conditions or a significant risk of multidrug-resistant TB. Adverse drug reactions occur in about 10% of patients (Box 17.52). Patients treated with rifampicin should be advised that their urine,

Index	
Correct	
Incorrect	
Ambiguous	
Repeated	

The evaluation data suggests that the system or model being assessed demonstrates relatively high accuracy, logic, informativeness, comprehension, and tone in its responses, with a majority of the responses being rated positively by Dr. Garima Thakur. However, there is room for improvement, particularly in addressing repeated answers and potentially refining the comprehensiveness of responses.

The expected output of the project will be one dynamic dashboard and prediction pages. The dashboard page includes some useful visuals to help better monitor, analyze, and manage. On the prediction model page, users can see the price prediction by inputting some property characteristics as well as model quality measurements. Figure (3) shows a draft preview of the project's result.

DISCUSSIONS

The project faces several significant challenges that impact its feasibility, implementation, and evaluation. These challenges include:

- **Resource Constraints for Model Loading:** The inability to load all three models due to GPU limitations severely hampers the functionality and effectiveness of the medical chatbot. This limitation restricts the complexity and size of models that can be utilized, potentially compromising the chatbot's ability to generate accurate responses to users' queries.
- **Parameter Availability for Models:** The absence of certain model parameters due to resource constraints further exacerbates the limitations of the project. Missing parameters can adversely affect the performance and reliability of the models, leading to suboptimal outcomes and potentially inaccurate responses.
- **Difficulty in Finding Medical Professionals for Evaluation:** The scarcity of medical professionals available to evaluate the chatbot poses a significant challenge. Without expert input and validation, it becomes challenging to assess the accuracy, relevance, and appropriateness of the chatbot's responses to medical queries. This limitation undermines the credibility and reliability of the chatbot as a source of medical information.
- **Limited Availability of Open-Source Datasets:** The scarcity of open-source datasets poses a significant hurdle in training and evaluating the chatbot. Access to diverse and comprehensive datasets is crucial for developing robust and effective machine learning models. The absence of suitable datasets restricts the scope and quality of the chatbot's knowledge base, potentially limiting its ability to address a wide range of medical queries accurately.

Despite these challenges, several important insights and hypotheses can be gleaned from the project:

- **Insights from Results:** The evaluation results provide valuable insights into the performance and effectiveness of the chatbot. Metrics such as accuracy, precision, recall, and user feedback offer valuable indicators of the chatbot's strengths and weaknesses. Analyzing these results can help identify areas for improvement and optimization, guiding future iterations of the chatbot.

-
- **Hypotheses for Improvement:** Based on the observed limitations and challenges, several hypotheses for improvement can be formulated. For example, exploring alternative model architectures that require fewer computational resources may help alleviate the GPU constraint issue. Additionally, leveraging transfer learning techniques and pre-trained models can mitigate the need for extensive parameter tuning and resource-intensive training.
 - **Limitations and Challenges:** It's essential to acknowledge and address the limitations and challenges inherent in the project. Lack of access to sufficient computational resources, expertise, and data can hinder the project's progress and impact its overall success. Recognizing these limitations can inform strategic decisions and resource allocation to overcome obstacles and enhance project outcomes.

In summary, while the project faces significant challenges, including resource constraints, model parameter availability, and data scarcity, it also presents valuable insights and hypotheses for improvement. By addressing these challenges and leveraging insights from the results, the project can navigate obstacles effectively and advance toward its goals of developing a robust and reliable medical chatbot.