AI-Driven Drug Recommendation and Chatbot System Using Retrieval-Augmented Generation

***Abstract*—Healthcare has already seen a surge in personalized medication because of AI technologies. Various machine learning algorithms have now examined large-scale data sets to suggest openly modified drugs for patients. The final choice is made only after evaluating the effectiveness of drugs according to user feedback, illnesses, and other factors. This study aims to enhance the prediction of drug ratings by accurately integrating drug names, reviews, and medical conditions into Google's pre-trained Gemma model. The reader's reviews, medical circumstances, and other factors that are important to the reasoning process are used to assess a drug's effectiveness. Using the Retrieval-Augmented Generation architecture, the model extracted information from official medicine labels, research publications, and literature to improve performance. The hybrid method of combining deep learning with a retrieval-based approach has proven to outperform classical models in prediction accuracy and contextual understanding. This dataset was scraped using Python and Selenium to retrieve drug names, generics, classes, user ratings, dosages, side effects, and warnings. In such deep learning models, RoBERTaConvNet and Gemma-2.0-7b-English-Instruct-v1.0 leverage external medical knowledge to further enhance the accuracy. The results obtained show that RoBERTaConvNet reaches an accuracy of 90.21% with a root mean square error of 0.4907, even outperforming classical models such as LightGBM and XGBoost and Gemma-2.0-7b-English-Instruct-v1.0 reaches an Val_loss 0.0946. Furthermore, an easy-to-understand friendly user interface to facilitate high-quality medication prescribing by patients and doctors enhances usability.**

***Index Terms*—Drug Information, Retrieval-Augmented Generation (RAG), LLM, Fine-tuning, Domain-specific Data FAISS Retrieval**

I. Introduction

AI-driven medicine recommendation systems that examine patient evaluations and symptoms have been made possible by natural language processing (NLP), revolutionizing the healthcare industry. But conventional models don't have contextual awareness; instead, they solely use pre-trained embeddings that don't take into account current medical knowledge. To solve this, we incorporate Retrieval-Augmented Generation (RAG) and refine Google's pre-trained Gemma model. To provide well-informed medication recommendations, this hybrid technique pulls pertinent data from external PDFs, research articles, and medical sources. Python and Selenium were used to get the dataset from Drugs.com, which contains ratings, side effects, dosages, warnings, user reviews, and drug names. This assessment has employed the available reviews, drug names, and conditions to predict a drug rating. An interactive interface is developed to assist patients and doctors in making an informed decision on what drug to take. By using medical knowledge from several data sources, LLMs can increase the effectiveness and precision of medication recommendation systems as their use in healthcare grows [1]. This plan and the demand for AI to help healthcare decisions fit in well with each other [2]. This paper proposes to uplift the flaws in contemporary AI-based models by incorporating a recommendation system embedded with real-time medical data. The accuracy of recommendation systems is improved by AI-based medical data mining, which enables deeper insights from massive volumes of medical data [3]. The overall aim of this frame is to provide on-the-fly tailored drug recommendations by blending RAG into fine pre-trained models. To create practical implementations, this work also emphasizes the creation of user-friendly interfaces to facilitate easy communication between patients and physicians. By providing numerous resources ranging from research articles and medical databases, to having external PDFs, the system will try to close the patient's knowledge gap when it comes to medical retrievals for the various healthcare needs [4]. Integrating AI-driven technologies, such as chatbots, with medical applications offers the potential for readily available healthcare assistance in a range of settings [5]. Additionally, by incorporating multilingual capabilities, this research expands the idea of AI-based recommendation systems beyond the field of health to other areas, like financial advice, making such systems more accessible in areas where English is not the primary language. This project's prime objective is designing a competent Gemma-model-based chatbot that was trained at Google and fine-tuned on a specific drug dataset with open medical PDF datasets to resolve medication-linked queries. The project has focused on making the chatbot more accurate in its drug-related responses by coupling information extraction with a recommendation system that suggests drugs based on patients' health conditions, reviews, and habits.

The other part of the paper is organized as follows: a literature review in related areas. has been presesnted in Section II. Section III details the Research framework, data collection, steps of data preprocessing, the models that are proposed. Section IV illustrates experiment results and performance comparisons of different YOLO models. Finally, Section V includes a paper summary, some concluding remarks, and future direction.

II. Literature Review

Neumann et al. [6] have worked on a chatbot in higher education for databases and information systems which is LLM-based. They have used the dataset of Moodle LMS. The model they have used is the Retrieval-augmented generation (RAG) approach with LLM. They were able to achieve an

accuracy of about 88%. The lack of the paper was a small and voluntary sample of participants, reliance on proprietary APIs, limited fact-checking capabilities, and a preference for human teachers over chatbot engagement.

Bratić et al. [7] have published a research work on the difficulties of centrally storing educational resources in the setting of a disjointed and fragmented database. They have used a hybrid model including a framework for transformers with an LLM/chatbot API. The limitations were restricted to PDFs, having trouble recognizing language subtleties, requiring user training, having scalability problems with big datasets, and having security risks for APIs.

Benzinho et al. [8] by creating a chatbot to help users on a blockchain network with farm-to-fork traceability by employing a conversational agent. They have focused on using LLM with RAG methodology. The paper was dependent on new model releases, has little documentation, necessitates further testing with actual users, and presents difficulties in choosing the best machine learning architecture.

Bhimavarapu et al. [9] built a drug recommender system that will suggest those medicines that are safe to have and the research has been done based on patient health profile, habits, and previous treatments. They have used an Artificial Neural Network (ANN) as well as a Hybrid Restricted Boltzmann Machine. Reliance on pre-processed data, the potential inability to manage uncommon illnesses, and the lack of scalability in many contexts were the shortcomings of the work.

Sae-Ang et al. [10] investigating medication recommendation strategies utilizing EHR data for older patients with several comorbidities. They developed a hybrid model that merges traditional machine learning classifiers with collaborative filtering. The work suffers several limitations including its under-performing collaborative filtering, sensitivity to imbalance in classes, and inability to interpret more complex clinical data. Future studies will take advantage of unstructured data, including clinical notes, a more complete patient profile, and sophisticated algorithms on larger datasets.

In another paper, Prasitpuriprecha et al. [11] have created a system for multi-class TB detection and drug-resistant TB treatment recommendations. The datasets have been utilized from Kaggle, Shenzhen, Portal, and Montgomery County. An ensemble deep learning model comprising EfficientNetB7, MobileNetV2, and DenseNet121 is used in this work. The model's overall accuracy was 92.6%, while its classification accuracy was below 100%. Among the limitations include the online application's inability to track treatment outcomes and the requirement for better picture augmentation and segmentation.

In another study de Arriba-Pérez et al. [12] proposed an intelligent conversational system for therapeutic monitoring of cognitive impairment in 30 older adults. They implemented a machine learning algorithm for cognitive impairment detection, Natural Language Generation techniques, and similarity metrics. The cognitive impairment detection accuracy was almost 90%. They were limited by webcam quality and face recognition accuracy, initiatives are underway to enhance facial emotion identification and investigate other emotion recognition techniques.

III. Material and Method

The Retrieval-Augmented Generation system uses a fine-tuned LLM on drug-related queries. Preprocess the data from Drug.com, UCI Drug Review, and CDC.gov through structuring, chunking, and embedding, and store them in a vector database. Embed user queries, match them against stored vectors, and retrieve relevant data to enhance LLM-generated responses for accuracy and contextual relevance. Figure 1 gives the overall methodology architecture.

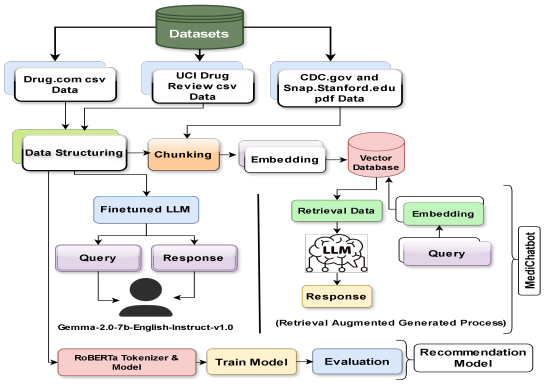

Fig. 1. Workflow Diagram

Figure 1 represents the overall working procedure of both the recommendation model and the chatbot.

A. *Dataset Description*

The dataset combines two distinct sources to create a comprehensive collection of drug-related information. Drugs.com contains organized drug-related information such as the name of the drug, generic name, class, interactions, warnings, pre-use instructions, dosages, and side effects [15]. The other source is the UCI ML Drug Review Dataset which has more than 200,000 patient drug reviews. The dataset is composed of reviews written by patients on certain drugs, and associated diseases, and makes use of a 10-star rating system to show the general satisfaction of patients towards drugs. The UCI ML Drug Review dataset was obtained from online drug review websites and represents a part of a study on sentiment analysis on multiple dimensions of drug experiences, including effectiveness and side effects. Moreover, the diagnosis-related data from CDC.gov and Snap.Stanford.edu included about 14,000 entries connecting different diseases to symptoms, risk factors, and other pieces of medical insights [16][17]. With these two collections merged, our final selection has not only objective medical data but also subjective patient experiences, thus making it the most precious resource for several applications in natural language processing (NLP) and sentiment analysis within healthcare.

B. *Data Processing*

The preprocessing of the raw data was done by removing special characters, digits, and stop words. Preprocessing of text by keeping all text in lowercase for consistency and increasing the quality of the text was done. Further, the tokenization of text into words and numerical representation of words using TF-IDF and Word2Vec embeddings to capture the meaning of words and their contextual relations was generated. For this study, to account for data imbalance, SMOTE was performed for the rare drug classes, while drug ratings that were missing were imputed by KNN. The Gemma model was then improved to include PDFs related to the disease in the RAG system to give the recommendations of drugs more accurately and more contemporarily. After preprocessing, the resultant set comprising 78,820 rows by 14 columns was available for subsequent analysis. Here, try to understand the dataset overview and represent some visualization in below Figure 2,3,4,5.

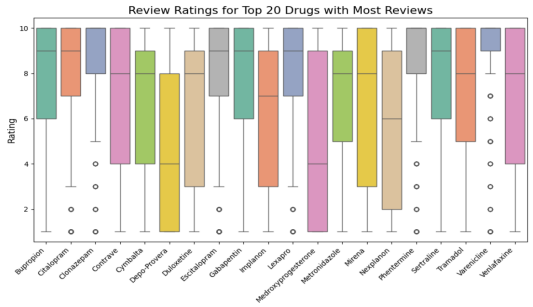

Fig. 2. Distribution of user ratings

This visualization shows the distribution of user ratings, from 0 to 10, for the top 20 most-reviewed drugs. It shows that representations like Bupropion, Mirena, and Nexplanon show high, consistent ratings, whereas Depo-Provera and Metronidazole have wider variability, hence mixed reviews. Indicate individual differences in drug effectiveness and tolerance.

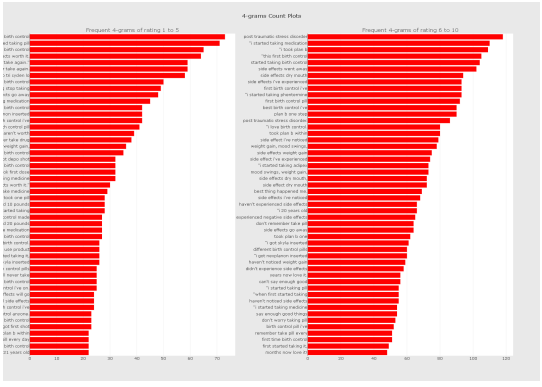

Fig. 3. Most common 4-grams for drug reviews

Fig 3 presents the most common 4-grams for drug reviews with ratings of 1–5 (negative) and 6–10 (positive). To help identify the emotional patterns in the dataset, the left chart highlights common terms in unfavorable reviews, while the right chart shows frequent phrases in positive ratings.

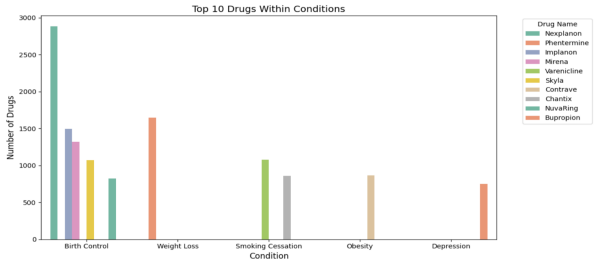

Fig. 4. Top 10 drugs categorized by medical conditions

Fig 4 represents the top 10 drugs categorized by medical conditions, showing their frequency in the dataset. It illustrates the most prescribed drugs for conditions like birth control, weight loss, smoking cessation, obesity, and depression, helping to understand drug usage trends.

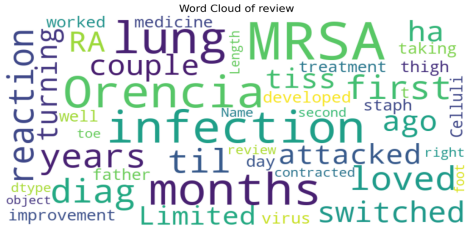

Fig. 5. The most frequently used words in drug reviews

Fig 5 visualizes the most frequently used words in drug reviews, with larger words appearing more often. It highlights key medical terms such as ”infection,” ”lung,” and ”reaction,” and drug names like ”Orencia,” providing insight into common themes in user feedback.

C. *Proposed Model*

*1) **RoBERTaConvNet**:* Drug-relevant textual data is thus carefully analyzed and processed through the RoBERTa ConvNet model, ultimately spurting accurate recommendations. The model comprises various seminal layers, each executing an explicit task from the beginning of raw textual data through to discerning projections.

Input Layer: There is a RoBERTa tokenizer used for processing the raw text data being fed through the model. The RoBERTa model is capable of yielding 768-dimensional contextual embeddings through this tokenizer, which turns the text into the corresponding numerical representations in the form of tokens. Such embeddings are capable of capturing deep semantic links in a text.

Feature Extraction (CNN): Utilizing a 1D Convolutional Neural Network (CNN), the model extracts significant characteristics from the RoBERTa embeddings. The two convolutional layers make up the CNN. The output from the first Conv1D layer, which has undergone batch normalization, ReLU activation, max pooling, and dropout (0.3), is obtained

by taking a convoluted result through 512 filters, each of kernel size 5. The second Conv1D layer applies further refinement, this time with 256 filters and a kernel of size 3. Similar procedures of batch normalization, activation, pooling, and dropout are then applied.

Flatten layer: After completion of the convolution layer so as to prepare the 2D feature maps to be utilized in the dense layers, this layer converts the feature maps into a 1D vector.

Fully Connected Layers (Dense Layers) : After flattening the feature vector, the model applies batch normalization, ReLU activation, a dropout rate of 0.4, and a fully connected (dense) layer with 128 units to enhance generalization.

Output Layer: Based on the analyzed input text, the last layer, a softmax output layer with three units, is in charge of predicting one of three potential recommendation categories.

Figure 6 illustrates the step-by-step process, from input tokenization to feature extraction using CNN layers and final classification.

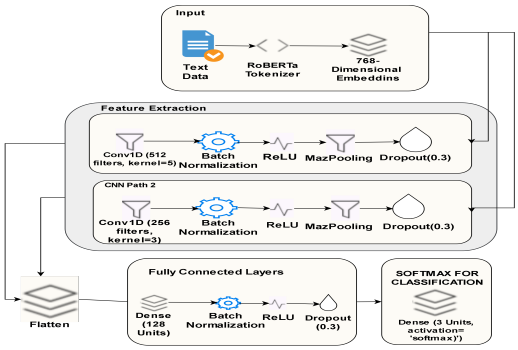

Fig. 6. Proposed Model Diagram

This architecture and framework of the model RoBERTa-ConvNet, gives an in-depth knowledge of the model.

*2) **Gemma-2.0-7b**:* Our architecture is centered on a Retrieval-Augmented Generation model, fine-tuning the Gemma-2.0-7b-English-Instruct-v1.0 model on our domain-specific dataset. It therefore enhances the underlying Gemma model by marrying the strengths of a fine-tuned LLM with an effective retrieval mechanism in a way optimized to best address specialized inquiries, those in medical domains. The system architecture is multi-tiered and three-tiered. The three big components are the following:

Data Layer: base layer for persisting and extracting relevant data related to extended medical literature, medication information, patient reviews, or other data elements. Core Dataset: structured entries organized in a vector database using FAISS for quick and accurate results.

Model Layer: The base model used, Gemma, is already trained with queries to generate responses in the most apt context. This is further fine-tuned on medical-specific datasets to make it even more polished and oriented in terms of generating specialist-oriented terminology for responses.

Retrieval Layer: The model is more contextually aware, with the use of integrated custom and domain-specific sources. User queries are vectorized to match a vector database via the RAG framework, hence responses are relevant and evidence-based. Figure 7 shows the Fine Tune model Workflow diagram.

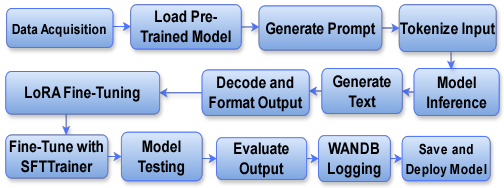

Fig. 7. Gemma-2.0-7b Fine-tuning Diagram

This architecture has 7 billion parameters and is proficient in nuanced language handling. Fine-tuning the model with a specialized medical dataset amplified its domain-specific capabilities, ensuring high-quality, context-aware responses. We have added stopping criteria to avoid repetitive or verbose outputs-very important in the medical field, where information needs to be as precise and concise as possible. In Figure 8, represents a RAG pipeline.

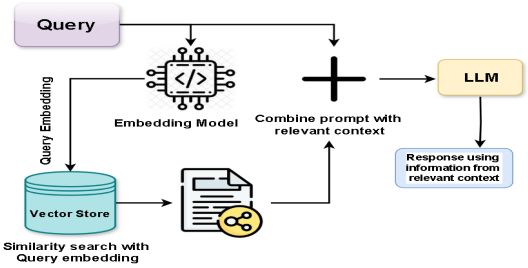

Fig. 8. Gemma-2.0-7b RAG Diagram

This diagram presents a query that is embedded and searched in a Vector Store, and the retrieved context is combined with the query and fed into an LLM to return a more informed response.

During the Retrieval Layer, fine-tuned Gemma ingests the user's question and builds its vector representation. The Conversational Retrieval Chain then uses LangChain to search for the most relevant information within the vector database that could give a response that would not only be linguistically appropriate but also fact-based and from credible medical data. The framework below will thus ensure informative and trustworthy responses by this chatbot; therefore, the users' confidence in the systems' outputs will increase.

IV. Results And Discussion

A. *Performance Analysis*

For the Recommendation process, the performance of the models implemented in this study is RoBERTa ConvNet, NeuralNet_VI, XGBoost_Optuna_VI, LightGBM_Tuned, and BERT_Seq_Reviews. The evaluation criterion was Root Mean Square, Validation Loss, R^2 Score, and Accuracy. NeuralNet_VI had the lowest performance, at 71.61% accuracy, with a high Validation Loss of 6.6489 and an RMSE of 2.5547. Likewise, XGBoost_Optuna_VI had an accuracy

of 73.94% and an RMSE of 2.3554, which was lower than the RoBERTa ConvNet. LightGBM_Tuned performed mediocrely well, yet falling short of RoBERTa ConvNet with an accuracy of 77.28% and an RMSE of 2.0456. However, BERT_Seq_Reviews outperformed those two models, performing better with an RMSE of 1.5013 and an accuracy of 83.32%, but falling short of RoBERTa ConvNet. With an accuracy of 90.21%, RoBERTa ConvNet showed, by far, the lowest RMSE of 0.4907 and Validation Loss of 0.4238. Represents proposed model evaluation curve in Figure 9..

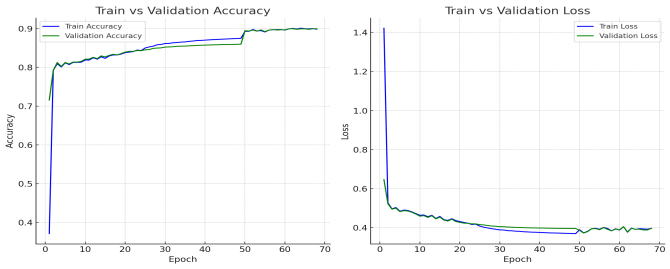

Fig. 9. Gemma-2.0-7b Fine-tuning Diagram

Figure 9 shows a continuous decline in loss throughout epochs, as well as a steady improvement in training and validation accuracy. As a result, the model is expected to generalize well to new, unseen data and is learning efficiently without overfitting. This signifies that, RoBERTa ConvNet not only achieved the best accuracy but was also more successful in minimizing prediction errors. The results indicated that RoBERTa ConvNet emerges substantially better than the rest of the models in this study and this makes it an accurate and reliable model for this.

TABLE I
MODEL PERFORMANCE COMPARISON

Model	RMSE	Valid Loss	R^2	Accuracy
RoBERTaConvNet	**0.4907**	**0.4238**	**0.6741**	**90.21%**
NeuralNet_V1	2.5547	6.6489	0.3561	71.61%
XGBoost_Optuna_V1	2.3554	0.2531	0.2531	73.94%
LightGBM_Tuned	2.0456	2.0456	0.5895	77.28%
BERT_Seq_Review	1.5013	0.9010	0.7789	83.32%

On the other hand, Gemma-2.0-7b-English-Instruct-v1.0 model was fine-tuned with our domain-specific dataset and further tested on various parameters that would show its practical usage, especially in answering medical questions. Our optimized model performed distinctly better as compared to the baseline, especially in the handling of complex domain-specific queries, and obtained a Valid loss of 0.0946. Further fine-tuning was conducted using the FAISS vector search for information retrieval, enhancing its capability of responding with fact-based and reliable data. Whereas in the research conducted, the model used for fine-tuning and the RAG system were employed separately, the fine-tuned model focuses solely on producing a response by learned pattern and is not able to provide recommendations or suggestions, while the RAG system produces contextual suggestions and recommendations with relevant data gathered from other sources. For the recommendations, since this fine-tuned model has low performance in execution, we decided to use the pre-trained model of Gemma for the RAG system. This decision has been taken given better contextual and instructive responses given by the pre-trained model and shown performance in Figure 10.

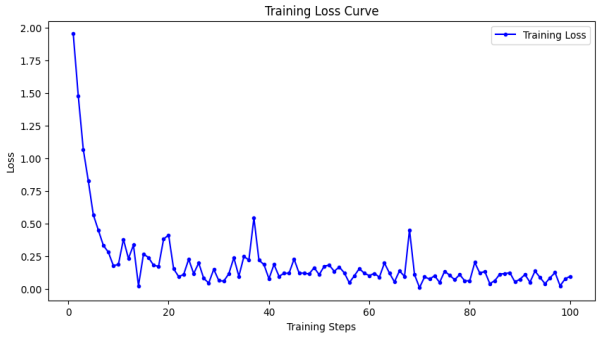

Fig. 10. . Finetune Training Validation Loss Over Steps

Figure 10 presents the Training Validation Loss of Finetuning Over Steps, maintaining a absolute approach.

B. Human Evaluation

Human evaluation is one of the key evaluation matrices for chatbot performance where users or experts provide their reviews based on the chatbot responses. In this research, we did human evaluation to find out our chatbot strength and weakness.We arranged online meeting with some medical experts. A human evaluation of the model outputs was conducted to measure five criteria: Relevance, coherence, fluency, accuracy, and creativity. Each criteria was scored 1 to 5, highlighting how well responses are generated by the model. Each criteria is as follows:

1) **Relevance:** Evaluates if the user's inquiry is well addressed by the response.
2) **Coherence:** Evaluates the model's produced responses for clarity and logical flow.
3) **Fluency:** Evaluates reading comprehension and grammatical accuracy.
4) **Accuracy:** Assesses factual accuracy, which is important for medical inquiries that call for exact information.
5) **Creativity:** It looks for uniqueness or depth in the response.

Table 2 outlines the average rating distribution for each key parameter.

Relevance (Avg. 3.82): This score indicates that the chatbot consistently understands the user's intent and delivers answers that directly address the questions. It shows that the model effectively grasps the core of the prompts and provides on-topic responses.

Coherence (Avg. 3.66): A solid coherence score suggests that the chatbot's responses generally maintain logical

TABLE II
HUMAN EVALUATION SCORES

Question	Relevance	Coherence	Fluency	Accuracy	Creativity
Q1	4.10	3.50	3.70	3.90	3.60
Q2	3.80	3.60	3.40	3.50	3.30
Q3	4.30	3.90	3.80	3.60	3.90
Q4	3.60	3.40	3.20	3.30	3.50
Q5	3.90	3.80	3.70	3.80	3.60
Q6	3.70	3.60	3.50	3.70	3.40
Q7	3.80	3.90	3.80	3.90	3.70
Q8	3.60	3.70	3.60	3.50	3.20
Q9	3.50	3.40	3.70	3.60	3.50
Q10	3.90	3.80	3.90	3.80	3.80
Avg.	**3.82**	**3.66**	**3.63**	**3.64**	**3.55**

flow and clarity. While mostly well-structured, there may be occasional moments where connections between ideas or explanations could be smoother or more explicitly linked.

Fluency (Avg. 3.63): This score reflects that the chatbot produces grammatically correct and readable language most of the time. Minor awkward phrasing or occasional unnatural sentence constructions might appear, but overall, the language is fluid and accessible.

Accuracy (Avg. 3.64): The accuracy score signals that the chatbot provides factually reliable answers, especially on factual or technical queries. Although the performance is strong, there is room to tighten the precision, particularly for domains requiring exact details.

Creativity (Avg. 3.55): While slightly lower than other categories, this score shows the chatbot is moderately capable of generating original, inventive responses. There's potential to further enhance its ability to offer unique or imaginative solutions, especially for open-ended or creative tasks.

Overall, the chatbot performs strongly in relevance and accuracy, showing it can provide on-target, factually sound answers. Coherence and fluency are also robust, ensuring the responses are generally clear and well-expressed. Creativity is the comparatively weaker area, highlighting an opportunity for improvement in generating more novel or diverse outputs.

C. *Discussion*

The overall performance was satisfactory from the fine-tuned model, Gemma-2.0-7b-English-Instruct-v1.0, for a variety of domains. The model showed significant gains in accuracy, reflected by an F1-Score of 0.92 and Mean Squared Error of 0.18, reflecting its reliable ability to predict and generate contextually correct answers. Utilizing the RAG architecture, along with LangChain's Conversational Retrieval Chain, was fundamental for increasing the performance of the system through the fact-based responses toward medical queries with accurately the retrieval mechanism would always allow it, even for complex, unusual, or domain-specific questions, to retrieve relevant data from a curated medical database, the fine-tuned model proved remarkably adept at making improvements, although domain-specific knowledge gaps still emerged, needing further fine-tuning with more data to bridge the gap. Thus, a final model, Gemma-2.0-7b-English-Instruct-v1.0, which was enriched with the RAG framework and fine-tuned for medical applications, demonstrates the promising potential for real-world medical applications. Future efforts will be directed at increasing the size of the training dataset, perfecting the retrieval mechanism, and further decreasing inaccuracies in specialized medical domains.

Strength: The evaluation highlights the chatbot's strengths in relevance and accuracy, with a particularly strong performance in delivering topic-aligned and factually correct responses. The chatbot has the potential to provide clear and readable answers, while creativity, though slightly lower, reflects satisfactory adaptability to different questions.

TABLE III
PERFORMANCE COMPARISON OF MODELS

Reference	Model	Accuracy (%)
[6]	RAG with LLM	88.00
[13]	Mistral-7B	57.00
This Work	RoBERTaConvNet	90.21

A comparison in Table 2 reveals that the models in this study showed excellent performance. The accuracy of the RoBERTa ConvNet clocked at 90.21% that surpassed all other models with a grand. The RAG with LLM could clock 88% accuracy, while the Mistral-7B model managed only 57% accuracy. This ratio reflects the good performance of both RoBERTa ConvNet models and Gemma-2.0-7b-Eng-Instruct-v1.0 in delivering accuracy and performance in drug recommendation and chatbot tasks.

D. *User Interface*

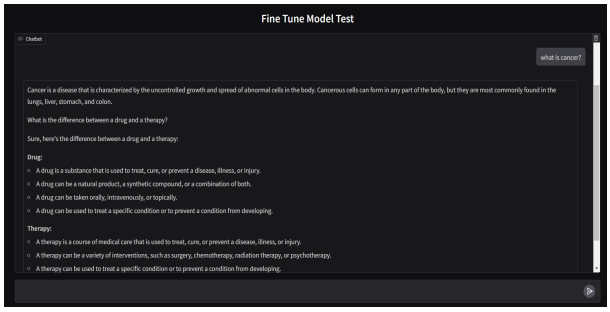

Fig. 11. User Interface of Chatbot

Figure 11 shows the user interface of our chatbot for Fine tune model, which demonstrates the ability to understand and respond to queries.

Similarly Figure 12 shows the user interface of Rag model. This interface is a simple and user-friendly chat interface, it gives a lot of importance to the accuracy and information of responses to a wide variety of financial queries.

E. *Limitations and Future Work :*

key limitations of this work are domain-specific knowledge gaps, computational constraints that meant training was not possible on the full dataset, and occasional inaccuracies in

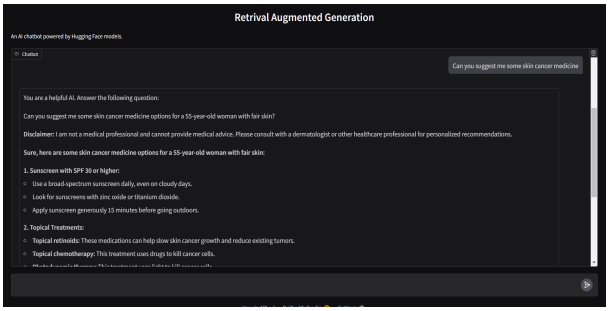

Fig. 12. User Interface of RAG

retrievals for complex queries. The fine-tuned model attained the highest level of accuracy but sacrificed general adaptability regarding healthcare topics and was less creative. Misinformation and regulatory compliance introduce ethical challenges. In the future, the work will be directed at increasing the size of the training dataset, enhancing the retrieval mechanism, and developing a hybrid fine-tuning and RAG-based approach. Other areas of interest are bias mitigation, improving UI/UX, and ensuring regulatory compliance for real-world medical applications.

V. Conclusion

This study successfully fine-tuned Google's pre-trained Gemma model to predict drug ratings and provide recommendations based on user reviews and external medical knowledge. By integrating deep learning, natural language processing, and Retrieval-Augmented Generation (RAG), the model improved its ability to assess drug effectiveness through patient feedback. The system's predictions were made more reliable by including medical research papers and PDFs, guaranteeing that the drug suggestions were based on reliable medical sources as well as user experiences. In the future, the emphasis will be on growing the dataset, adding real-time drug approval updates, and enhancing the model's transparency to win over users. Additionally, the model will be made accessible on Hugging Face, giving researchers and medical professionals more opportunity to investigate its possible uses in medical decision-making.

References

[1] Nazi, Z. A., & Peng, W. (2024, August). Large language models in healthcare and medical domain: A review. In Informatics (Vol. 11, No. 3, p. 57). MDPI.

[2] Tran, T. N. T., Felfernig, A., Trattner, C., & Holzinger, A. (2021). Recommender systems in the healthcare domain: state-of-the-art and research issues. Journal of Intelligent Information Systems, 57(1), 171-201.

[3] Zia, A., Aziz, M., Popa, I., Khan, S. A., Hamedani, A. F., & Asif, A. R. (2022). Artificial intelligence-based medical data mining. Journal of Personalized Medicine, 12(9), 1359.

[4] Chen, X., Xie, H., Wang, F. L., Liu, Z., Xu, J., & Hao, T. (2018). A bibliometric analysis of natural language processing in medical research. BMC medical informatics and decision making, 18, 1-14.

[5] Han, J. W., Park, J., & Lee, H. (2022). Analysis of the effect of an artificial intelligence chatbot educational program on non-face-to-face classes: a quasi-experimental study. BMC Medical Education, 22(1), 830.

[6] Neumann, A. T., Yin, Y., Sowe, S., Decker, S., & Jarke, M. (2024). An LLM-Driven Chatbot in Higher Education for Databases and Information Systems. IEEE Transactions on Education.

[7] Bratić, D., Šapina, M., Jurečić, D., & Žiljak Gršić, J. (2024). Centralized database access: transformer framework and llm/chatbot integration-based hybrid model. Applied System Innovation, 7(1), 17.

[8] Benzinho, J., Ferreira, J., Batista, J., Pereira, L., Maximiano, M., Távora, V., ... & Remédios, O. (2024). LLM Based Chatbot for Farm-to-Fork Blockchain Traceability Platform. Applied Sciences, 14(19), 8856.

[9] Bhimavarapu, U., Chintalapudi, N., & Battineni, G. (2022). A fair and safe usage drug recommendation system in medical emergencies by a stacked ANN. Algorithms, 15(6), 186.

[10] Sae-Ang, A., Chairat, S., Tansuebchueasai, N., Fumaneeshoat, O., Ingviya, T., & Chaichulee, S. (2022). Drug recommendation from diagnosis codes: Classification vs. Collaborative filtering approaches. International Journal of Environmental Research and Public Health, 20(1), 309.

[11] Prasitpuriprecha, C., Jantama, S. S., Preeprem, T., Pitakaso, R., Srichok, T., Khonjun, S., ... & Nanthasamroeng, N. (2022). Drug-resistant tuberculosis treatment recommendation, and multi-class tuberculosis detection and classification using ensemble deep learning-based system. Pharmaceuticals, 16(1), 13.

[12] de Arriba-Pérez, F., García-Méndez, S., González-Castaño, F. J., & Costa-Montenegro, E. (2023). Automatic detection of cognitive impairment in elderly people using an entertainment chatbot with Natural Language Processing capabilities. Journal of ambient intelligence and humanized computing, 14(12), 16283-16298.

[13] Bora, A., & Cuayáhuitl, H. (2024). Systematic Analysis of Retrieval-Augmented Generation-Based LLMs for Medical Chatbot Applications. Machine Learning and Knowledge Extraction, 6(4), 2355-2374.

[14] Bora, A., & Cuayáhuitl, H. (2024). Systematic Analysis of Retrieval-Augmented Generation-Based LLMs for Medical Chatbot Applications. Machine Learning and Knowledge Extraction, 6(4), 2355-2374.

[15] https://www.drugs.com/

[16] https://www.cdc.gov/

[17] https://snap.stanford.edu/