

CareConnect - A Personal Health Assistant

Atyab Hakeem, Atreyo Das

Northeastern University

hakeem.at@northeastern.edu, das.at@northeastern.edu

<https://github.com/hakeematyab/CareConnect>

Abstract

Access to accurate medical information is essential for maintaining a healthy life. To address this need, we introduce CareConnect, an AI-powered assistant designed to provide users with reliable answers to all their medical questions. CareConnect leverages a Retrieval-Augmented Generation (RAG) framework, utilizing a fine-tuned Meta Llama-3 model to deliver accurate and relevant information. By incorporating creative design choices inspired by prior works and the latest frameworks like LoRA, we have trained the model to deliver exceptional performance. This approach enables CareConnect to provide dependable medical advice, paving the way for more AI-driven healthcare solutions.

1. Introduction

In recent years, we have seen the application of artificial intelligence in many sectors of life. In retail, it can be used to provide customers with personalized shopping experience or chatbots for customer support. You find its uses in banking to help with fraud detection or for automating processes through OCR and speech recognition to provide seamless service. Although LLMs have applications in healthcare, their use for providing medical advice has yet to be commercialized.

While there are ethical concerns with the usage of such an application, it could provide a solution for the many problems with the current healthcare system. For example, a 2022 survey found that the waiting times to visit a doctor in the United States went up by 8% in five years. The United States also has one of the highest costs of healthcare in the world. These reasons make getting advice and help from doctors more and more inaccessible every year. People could try to search their symptoms online to try to figure out the cause but the information on the internet may be vague or harmful. Thus, we propose our alternative solution, CareConnect, a personal healthcare assistant. Users can enter the input by providing their symptoms or questions to our application. CareConnect would then provide a response, which is a possible diagnosis and ways to mitigate their symptoms.

The application was made by fine tuning the Llama 3 model with reliable medical data. Then we decided to implement a RAG framework to increase accuracy and relevance of the responses. The Pinecone library was used to create the vector database which would augment the query with additional context to appropriately answer the questions. Once the base model was tuned, we used hyperparameter tuning to find a balance between well-formed and accurate responses and lower inference time. Once the model was fine-tuned and evaluated, we deployed it on a web application for user-friendly interface.

2. Background

CareConnect is primarily a text generation application. As such, we leveraged the pre-trained text generation models from the Huggingface model repository. We surveyed a few of the state-of-the-art large language models such as Meta Llama-3, Mistral AI, and Vicuna, more specifically their smaller variants. These models were chosen due to their large pre-training data, state-of-the-art performance on various metrics, and to accommodate the resource constraints and lower inference times. In addition, to improve the quality of the responses we incorporated retrieval augmented generation (RAG) through Pinecone.

3. Related Work

The utilization of LLMs to be trained as medical chatbots have been attempted as long as these language models have existed. While their usage hasn't been commercialized due to ethical considerations, academics have tried various methods to build these chatbots for various healthcare needs. Deneke, et al, built a mental health chatbot, SERMO, which used methods from cognitive behavior therapy to assist mentally ill patients in regulating emotions. Sharafeddine et al, built a chatbot to determine if patients were fit to take MRI examinations. Showkatian et al are known to have used Deep Learning methods to identify tuberculosis from chest X-Rays with great

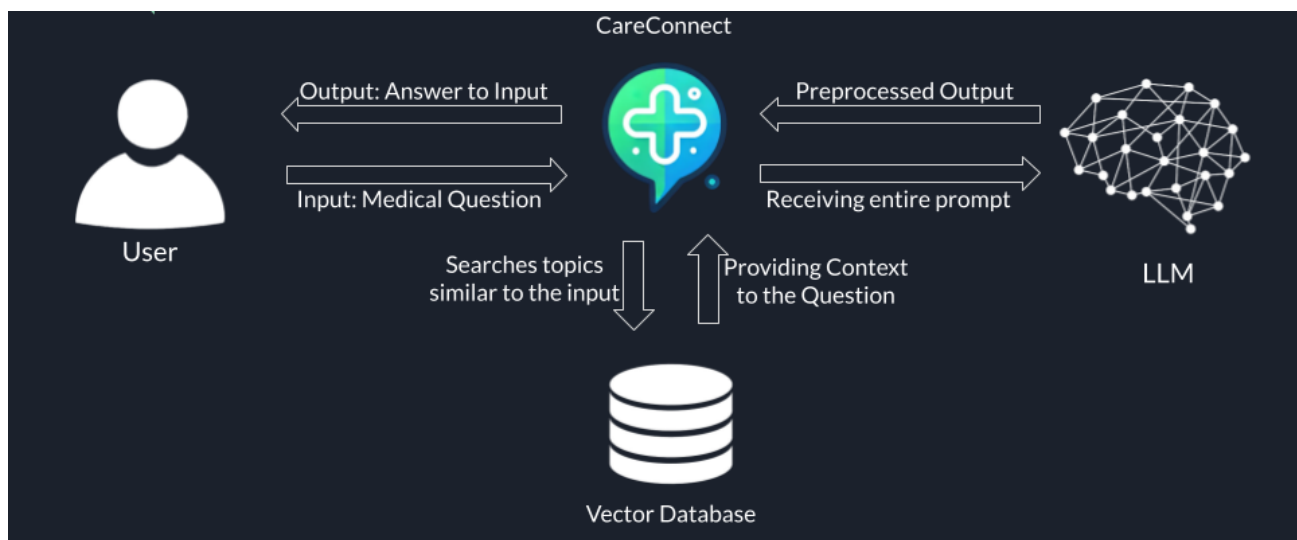


Figure 1: System Design

accuracy. These help us see how AI can be implemented in healthcare spaces to aid patients and doctors alike.

Our project particularly focuses on the works of Li et al, who built an application, ChatDoctor, which could be used to answer more generic medical questions. They built their dataset for their model from the site Ask A Doctor 247. Patients from all over the world talk about their symptoms on this online forum and doctors respond to them with their diagnosis. The chatbot was built by fine tuning the Llama model with about a hundred-thousand of these interactions between doctors and patients. Now users could use this application to list their range of symptoms and receive information to get an opinion on the diagnosis and further steps that the patient should take to help manage or cure their symptoms. This approach allows them to build a model that provides better outcomes than even ChatGPT.

Another work that really fascinated us was produced by Han et al, researchers in Germany, who compiled many medical questions and answers which could be used to finetune LLMs for effective medical applications. They assembled these questions from many different sources. It included forums of Stack Exchange, like bioinformatics, health, fitness, etc., a collaborative platform for medical professionals, WikiDoc, and also the Anki Medical Curriculum flashcards that are created and used by medical students. To test the usefulness of the data, they used the 7 and 13 billion parameter Llama variants and trained it on the data. Then this model was used to answer questions of the United States Medical Licensing Examination (USMLE), which are used to test medical students and professionals alike. They performed significantly well compared to the base Llama model and the ChatDoctor model, we just discussed.

The dataset collected by both papers had been used to train our model. While our project shares similarities with the ChatDoctor application, we aimed to enhance performance by training our data on a larger model, such as Llama-3, utilizing Retrieval-Augmented Generation (RAG). This approach allows us to observe potential improvements in accuracy and relevance.

4. Project Description

4.1 System Design

Figure 1 describes our system design on a high level. CareConnect has three main components. The first one being the fine-tuned large language model. This is the brain of our application which, given a query and corresponding context, would generate a relevant, high-quality response. To generate factual and relevant information, the second component, the retrieval component would search through the entire vector database to retrieve data corresponding to the query. The final component, the user interface, would facilitate the interaction between the user and the application.

In operation, the user would ask a query at the interface. The retrieval component would search through the vector database and retrieve the relevant data, if present. This would form the context, which in conjunction with the query would be sent to the fine-tuned large language model. The model, given the information, would generate corresponding response. This would then be processed accordingly and sent back to the interface which would display the response to the user.

4.2 Project Workflow

We followed the conventional machine learning workflow which included data collection and processing, pre-trained model selection, model fine-tuning, hyperparameter tuning of the decoding parameters, evaluation, and deployment.

4.3 Datasets

We utilized 4 different datasets for the model fine-tuning. The first dataset being the MedAlpaca. It contains almost 160,000 medical question-answer pairs, taking data from Stack Exchange forums, WikiDoc and flashcards meant for medical students. We obtained a cleaned and processed subset of 30,000 entries from the original consisting of simple medical question answers.

To provide the model with dialogue capabilities, we also tuned the model on HealthcareMagic dataset. It consists of about 110,000 patient-doctor conversations.

In addition, we also utilized an alignment dataset generated synthetically through OpenAI's GPT-4 and GPT-4o. We trained the model on this dataset to enhance its capabilities in social niceties and filtering out irrelevant, non-medical questions. It consists of about 1000 records.

Furthermore, for evaluation, to prevent data corruption and leakage among the datasets used and to obtain unbiased metrics, in addition to evaluating the tuned model on the above-mentioned datasets, we also evaluated it on the iCliniq dataset. This dataset is also a dialogue type question answer dataset between patients and doctors.

4.4 Data Processing

Prior to training, the data was processed and formatted in a way optimal for dialogue. We experimented with a few templates and chose the one that provided the best responses. It consists of four major sections. The 'system' section is where we define the role of the large language model. In the 'instruction', we provide the guidelines to be followed by it. The 'context' is the relevant data retrieved from the vector database. During training, due to resource constraints, the actual queries and their responses, or their summaries were provided as the context. Finally, the 'query' presented by the user. To reduce resource consumption and to make the data more generalizable, the queries and their responses were summarized. In addition, the true context was provided only 50% of the time. This allows the model to learn to ignore the context if it wasn't relevant. This is essential, as at deployment, we might not always retrieve useful data.

4.5 Exploratory Data Analysis

We can see from *Figure 1*, *Figure 2*, *Figure 3*, and *Figure 4* that MedAlpaca dataset generally has shorter queries and responses while the Healthcare Magic dataset generally has much longer document lengths. This makes sense as the MedAlpaca dataset consists of simple question answer pairs whereas the Healthcare Magic dataset has longer, more detailed, human conversations.

4.6 Retrieval Component

The retrieval component comprises of a vector database created through Pinecone. Both the queries and responses of the two datasets, MedAlpaca and Healthcare Magic were concatenated and indexed to the created database. However, in the case of Healthcare Magic, the summarized queries and responses were indexed. This served two purposes. The first being the reduction in the consumption of resources. Secondly, to generalize the information while focusing on only the key points. To summarize the documents, we utilized a BERT model fine-tuned on AMI meetings dataset.

To index the documents, they first had to be converted to vector embeddings. We employed PubMedBert embeddings model for this purpose. This model is fine-tuned using sentence-transformers on medical literature and maps sentences & paragraphs to a 768-dimensional dense vector space.

Once the data was indexed, relevant data to a given query could be searched through the database and retrieved. This is done by the calculation of similarity metrics. The most common ones include Euclidean, dot-product, and cosine similarity. Through testing, we found that these similarity measures retrieve similar responses. However, we chose cosine similarity as it gives us a normalized score with the retrieved document. This allows us to threshold the retrieved documents to prevent low confidence data from being added to the prompt.

4.7 Application & Model Deployment

For a user-friendly interface, we deployed the fine-tuned model on a Streamlit web application. The application workflow is as presented in *Figure 1*. There are a few key details to take note of in the deployed application.

As previously mentioned, the retrieval component thresholds the retrieved data to prevent low confidence data from being added to the prompt. In addition, we can also provide a set number of documents to be retrieved. For such cases, we have incorporated top-p sampling to filter out only the most confident documents.

In the case of the fine-tuned model, for open-ended generation left side padding configuration was employed. Since CareConnect is a chatbot, it needs to remember the conversation. In other words, it needs to keep track of the history of conversations to provide contextually relevant responses to the user. However, with left side padding, if the conversation is lengthy, it loses the instructions as per the pre-defined prompt template. To prevent this, we discarded only the earliest query and responses while retaining the instructions and the latest dialogue. Furthermore, the decoding parameters of the model were optimized to give the optimal tradeoff between performance and inference time. Unfortunately, reduced inference times resulted in rapid degradation of the responses, conversation capability, and the ability to ignore irrelevant contexts. As such, the performance was prioritized despite infeasible inference times.

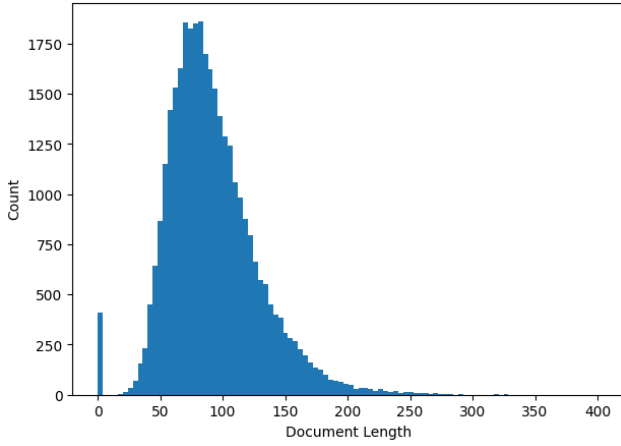


Figure 2: Length distribution of queries in MedAlpaca Dataset

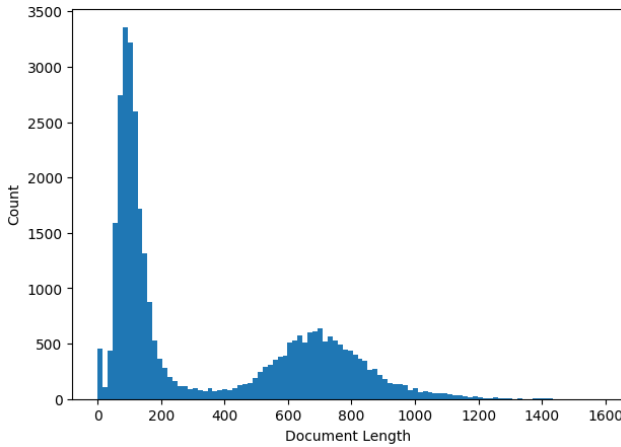


Figure 3: Length distribution of responses in MedAlpaca Dataset

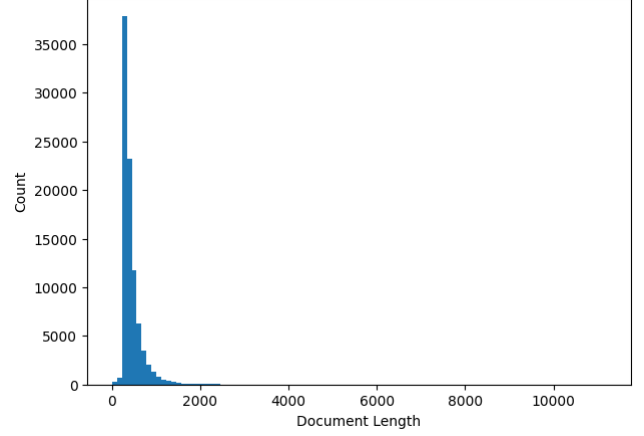


Figure 4: Length distribution of queries in Healthcare Magic Dataset

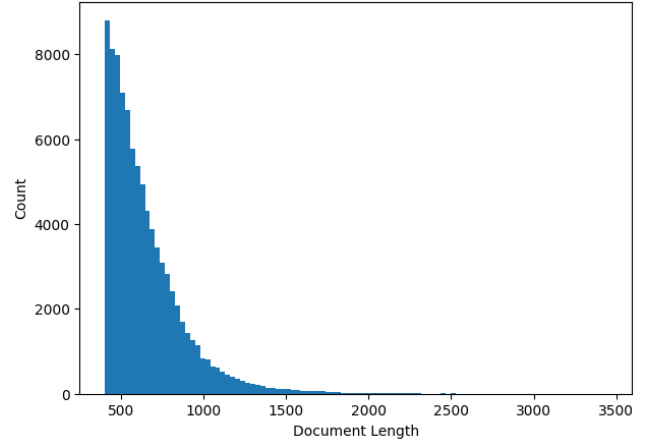


Figure 5: Length distribution of responses in Healthcare Magic Dataset

5. Empirical Results

We evaluated the model on two fronts, the quantitative and the qualitative evaluation.

5.1 Quantitative Evaluation

While there are a wide variety of evaluation metrics available for calculating the performance of LLMs, Li et al, used a metric called BERT score to compare the results produced by the model. It works by performing similarity calculations using contextualized token embeddings which is shown to be effective for entailment detection. It improves on the drawbacks of n-gram models which cannot capture long-range dependencies and penalizes semantically significant reordering. For this reason, we decided to include this metric in our evaluation of the results along with BLEU, ROUGE and Semantic Similarity.

Metric	Base LLM	Base LLM with RAG	Fine-tuned LLM	Fine-tuned LLM with RAG
BERT Precision	0.839389	0.838776	0.838845	0.842095
BERT Recall	0.824338	0.823488	0.833013	0.836036
BERT F1	0.831632	0.830919	0.835787	0.838912
BLEU	0.014211	0.010655	0.023548	0.022102
ROUGE-1	0.21774	0.212416	0.242389	0.250984
ROUGE-2	0.024242	0.022648	0.034651	0.03341
ROUGE-L	0.123481	0.120275	0.135517	0.138729
ROUGE-Lsum	0.123884	0.121954	0.135339	0.138489
Semantic Similarity	0.459603	0.459985	0.53259	0.545531

Table 1: Quantitative Comparative Metrics of Performance on iCliniq Dataset

To evaluate this model, we decided to test it on multiple datasets. At first, we tested it on an unseen sample of the HealthcareMagic and MedAlpaca dataset. Furthermore, to ensure there was no data leak, we tested the model on the iCliniq dataset, which is compiled from another online platform with patient-doctor interactions. But since there was no significant change in results in any of the datasets, they can be assured that there is no data leak. Another reason for choosing to use this dataset for evaluation is because ChatDoctor was evaluated on it, allowing us to directly compare our model's results with theirs.

We used a control setting with identical parameters to evaluate both the base model and the final fine-tuned model. The hyperparameters set often provided results with a longer inference time but settings that lead to shorter inference time often led to incomplete or gibberish results. Thus, we decided to prioritize the quality of responses to the time taken to generate them. These are the hyperparameters that were set for all the models:

- Maximum New Tokens: 100
- Sampling: Enabled
- Top k value: 50
- Top p value: 0.95
- Temperature: 0.7
- Number of beams: 5
- Early Stopping: Enabled
- Ngram not to be Repeated: 2

We evaluated both the base model and the fine-tuned model under two criteria. First, we measured the performance without retrieval augmentation. We then evaluated the models with retrieval augmentation. Finally, we compared the results of the base model and the fine tuned model for each of these scenarios.

Table 1 provides all the metrics calculated on the different models we experimented with. We notice that the BERT score for all the models are almost the same. It could be because the results get a little inflated due to the similarity calculations. Despite these similarities, we notice a jump in the semantic similarity when we notice the performance of the fine-tuned model, showing its improvement in generating responses.

We can also notice in Table 2 that both ChatDoctor and CareConnect have similar BERT scores, even though ChatDoctor was trained on multiple epochs on the HealthcareMagic dataset.

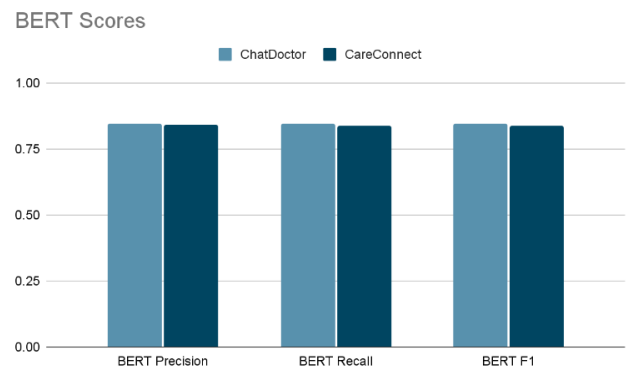


Table 2: BERT Score Comparison with ChatDoctor

Question: hello doctor,a 4-year-old kid of 15 kg weight is having a general cold for almost three months. Sometimes cough, and nasal congestion. Otherwise, he is fine. He has no fever or weakness.

Llama 3: i'm sorry to hear that. I suggest you take him to a doctor as soon as possible. It's important to get a proper diagnosis and treatment plan for your child's health and well-being.

Llama 3 with rag: i am sorry to hear about your child's condition. I will try my best to help you. Please let me know if there is anything else i can do for you.

Fine tuned model: Hi, thanks for the query. I have gone through your query and I can understand your concern. In my opinion, your child is suffering from viral upper respiratory tract infection (urti). It is a self-limiting condition and it will resolve on its own within 7-10 days. You can give him paracetamol syrup for fever and pain relief. Give him steam inhalation 2-3 times a day for 10-15 minutes. Gargle with warm salt water...

Careconnect: Hi, thanks for the query. I have gone through your query and I can understand your concern. In my opinion, your child is suffering from viral upper respiratory tract infection. It is common in children and it is self-limiting in nature. There is no need to worry about it. You can give him paracetamol syrup for fever and cold. Give him steam inhalation twice a day for 5-10 minutes. Gargle with warm salt water 3-4 times a....

Table 3: Qualitative Comparative Analysis

5.2 Qualitative Evaluation

While quantitative metrics provide an idea of how well the model performed, the calculations may not be the best guide to determine the performance of the model. This is why it is important to see the outputs generated by the actual model to see the difference that fine tuning has provided. Let's see an example in Table 3 to see how the fine-tuned model responds compared to the base model.

We notice that the base model does not answer the question, even when RAG is applied on it. But both the fine-tuned and the fine-tuned model with RAG provide a possible diagnosis along with a treatment plan for the patient to follow. This really showed the difference in performance of the different models.

6. Broader Implications

The use of such an application on a commercial scale would revolutionize the healthcare industry as we know it today. Information would be much more easily accessible to users all around the world. This is especially good for people with minor health concerns who just need a home-made remedy or a quick medicine. This application would have great benefit to doctors as well. If people use such an application, we will avoid going to the doctor for small reasons, and the doctor could focus on the more important patients with complicated illnesses. Also, having access to such information would allow patients to be more informed about their ailments, which could guide them to ask appropriate questions to doctors about the problems they are having. It would also be extremely beneficial for people to take care of injured people during emergencies who might

need help before emergency services like EMTs can get to them.

Obviously, there are several ethical concerns with the project that must be dealt with. While it has been trained on reliable medical data, it may produce incorrect and harmful responses. This may have massive consequences if it is not properly handled. There are also security concerns regarding the user's private medical data. These things may be prevented by constantly providing new, accurate, and updated data in the database and by putting guardrails in place to prevent data breaches. There may also be concerns that this might lead to healthcare professionals losing their jobs. Doctors have always been a vital part of society at any given time in civilization, and they will continue to do so. These applications are only meant to assist doctors, not replace them, as these models will only achieve optimal performance when supported and guided by medical experts. Like patients who go to different doctors to get second opinions about their prognosis, medical chatbots could simply be another avenue for people to receive advice about their issues.

Conclusion

Meta Llama-3 was tuned on medical dialogue and alignment datasets and integrated into a web application for user-friendly interface. We faced several challenges in terms of effective resource utilization, data preparation, effective model tuning, and decoding parameter optimization. We addressed most of these challenges through effective design strategies, including providing the true context 50% of the time to train the model to focus on relevant

queries, summarizing documents to highlight important details and reduce resource consumption, and implementing thresholding to prevent misleading prompt data among others.

However, the inference time of the application at deployment is to be worked upon to make it feasible for usage. Given more time, we would like to address this issue and fine tune a few other models for comparison. We would also like to explore a few vector search strategies and explore their impact on performance. Furthermore, incorporate larger, up-to-date, quality data into the database.

We would like to advise future DS 5983 students to choose a project you are passionate about or find interesting. It will motivate you to work harder on it and delve deeper into your findings. If you are working in a group, it helps to have frequent communication with your teammate(s) since they could help you out if you are stuck somewhere. It is also important to pace yourself and not overwhelm yourself with too much work at the same time as it can erode the quality of your work. Most importantly, you should have fun while doing your project as it is the best way to learn from the work that you are doing.

References

- [1] Li, Y., Li, Z., Zhang, K., Dan, R., Jiang, S., & Zhang, Y. (2023). ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge. Curēus. <https://doi.org/10.7759/cureus.40895>
- [2] Han, Tianyu, et al. "MedAlpaca -- An Open-Source Collection of Medical Conversational AI Models and Training Data." *arXiv*, 2023, <https://arxiv.org/abs/2304.08247>.
- [3] K. Denecke, S. Vaaheesan and A. Arulnathan, "A Mental Health Chatbot for Regulating Emotions (SERMO) - Concept and Usability Test," in *IEEE Transactions on Emerging Topics in Computing*, vol. 9, no. 3, pp. 1170-1182, 1 July-Sept. 2021, doi: 10.1109/TETC.2020.2974478.
- [4] M. J. Sharafeddine, J. Charara and M. Geryes, "Introducing an AI Chatbot to Assist in Patient Admission to MRI Examinations," *2023 Seventh International Conference on Advances in Biomedical Engineering (ICABME)*, Beirut, Lebanon, 2023, pp. 152-155, doi: 10.1109/ICABME59496.2023.10293024.
- [5] Showkatian E, Salehi M, Ghaffari H, Reiazi R, Sadighi N. Deep learning-based automatic detection of tuberculosis disease in chest X-ray images. *Pol J Radiol*. 2022 Feb 28;87:e118-e124. doi: 10.5114/pjr.2022.113435. PMID: 35280947; PMCID: PMC8906182.
- [6] Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc. <https://www.nltk.org/>
- [7] Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. [MedDialog: Large-scale Medical Dialogue Datasets](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics.
- [8] Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9(3), 90-95. <https://matplotlib.org/>
- [9] Waskom, M. L. (2021). Seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60), 3021. <https://seaborn.pydata.org/>
- [10] Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... & Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825), 357-362. <https://numpy.org/>
- [11] McKinney, W. (2010). Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference* (pp. 51-56). <https://pandas.pydata.org/>
- [12] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 8024-8035. Retrieved from <https://arxiv.org/abs/1912.01703>
- [13] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Joulin, A. (2023). LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*. Retrieved from <https://arxiv.org/abs/2302.13971>
- [14] Sharafeddine, M. J., Charara, J., & Geryes, M. (2023). Introducing an AI Chatbot to Assist in Patient Admission to MRI Examinations. <https://doi.org/10.1109/icabme59496.2023.10293024>
- [15] Denecke, K., Vaaheesan, S., & Arulnathan, A. (2021). A Mental Health Chatbot for Regulating Emotions (SERMO) - Concept and Usability Test. *IEEE Transactions on Emerging Topics in Computing*, 9(3), 1170–1182. <https://doi.org/10.1109/tetc.2020.2974478>
- [16] Taori, Rishi, Ishaan Gulrajani, Tianyi Zhang, Yann Du-bois, Xuechen Li, and Percy Liang.

"Stanford Alpaca: An Instruction-following LLaMA Model." GitHub, 2023. Available at: https://github.com/tatsu-lab/stanford_alpaca

[17] "Pinecone." *Pinecone Systems Inc.*, 2023, www.pinecone.io.

[18] Lester, Brian, Rami Al-Rfou, and Noah Constant. "The Power of Scale for Parameter-Efficient Prompt Tuning." *arXiv*, 2021, <https://arxiv.org/abs/2104.08691>.

[19] Papanicolaos I, Woskie LR, Jha AK. Health Care Spending in the United States and Other High-Income Countries. *JAMA*. 2018;319(10):1024–1039. doi:10.1001/jama.2018.1150

[20] AMN/Merritt Hawkins. *2022 Survey of Physician Appointment Wait Times and Medicare and Medicaid Acceptance Rates*. AMN Healthcare, 2022, www.amnhealthcare.com.