Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/55/20241433

The investigation of symptom to disease chatbot based on LLAMA-2

Naijie Xu

College of Information Science and Engineering, Northwest University, Xi'an, 710127, China

xunaijie@stumail.nwu.edu.cn

Abstract. In the swiftly advancing realm of Artificial Intelligence (AI), there has been a remarkable evolution in Natural Language Processing (NLP). Language models like ChatGPT have ushered in a new era of human-computer interaction, bestowing upon people an unparalleled capacity to comprehend and generate text that mimics human language. When these models are fine-tuned for specific domains, such as medicine, they become endowed with domain-specific knowledge, bridging the chasm between language comprehension and specialized expertise. However, a noteworthy research gap exists when it comes to harnessing AI-powered NLP for medical diagnosis and symptom analysis, particularly in the development of a Symptom to Disease Chatbot. This innovative approach aims to provide accurate and accessible healthcare information, enhancing the initial steps of medical consultation. This research proposes and develops a pioneering Symptom to Disease Chatbot, powered by the sophisticated Llama 2 13b pre-trained language model. By integrating extensive medical knowledge and AI capabilities, this chatbot seeks to empower individuals to make informed decisions about their health, potentially transforming personal health management. The research includes data preparation involving curated medical datasets, transforming them into a conversational format suitable for training. The base model, Llama 2, is fine-tuned for this medical context. While preliminary results are promising, further exploration with larger datasets is essential to enhance performance. Additionally, an open-source Gradio interface enhances user interaction. This research addresses the critical need for accessible healthcare information and demonstrates the potential of AI-powered language models in the healthcare sector.

Keywords: Artificial Intelligence, Natural Language Processing, Medical Diagnosis, Llama 2, Healthcare Accessibility.

1. Introduction

In the swiftly evolving landscape of Artificial Intelligence (AI), one of the most revolutionary advancements has materialized in the domain of Natural Language Processing (NLP). Central to this progression are intricate language models, exemplified by ChatGPT, which have fundamentally reshaped human-computer interaction. These models, fortified by extensive datasets and state-of-the-art algorithms, exhibit an unprecedented capacity to comprehend and generate human-like text [1]. Of particular significance is their adaptability to specific domains through a process known as fine-tuning. By training these models on domain-specific data, they become finely attuned to distinct areas of expertise, be it law, medicine etc. This expertise not only equips them with the ability to comprehend

DOI: 10.54254/2755-2721/55/20241433

the nuances of language within these specific domains but also enables them to generate responses that are contextually relevant, thereby providing insights and information that align with the subject matter. It is at this intersection of general language proficiency and domain-specific mastery that a realm of innovative possibilities is realized. The landscape of medical diagnostics, in particular, stands to benefit profoundly from the potential harbored by AI-powered NLP models. The lexicon of medical symptoms, conditions, and treatments necessitates an acumen that transcends conventional language processing capabilities. The amalgamation of extensive pre-trained language models, typified by Llama 2 [2], and their subsequent fine-tuning within the medical sphere, presents an unparalleled opportunity to bridge the chasm between linguistic comprehension and medical expertise. This convergence holds applications, notably the conception of a Symptom to Disease Chatbot, poised to revolutionize the initial stages of medical consultations.

Amid the rapid evolution of AI-powered NLP, a notable research gap beckons for exploration. While significant progress has been made in integrating AI technologies into the healthcare sector, a promising avenue remains largely unexplored-leveraging these advanced language models for medical diagnosis and symptom analysis, particularly through the development of a Symptom to Disease Chatbot. Despite the remarkable strides witnessed in AI-assisted medical diagnostics, there exists an untapped potential in harnessing the capabilities of extensively pre-trained language models to create an intelligent interface that can bridge the gap between individuals experiencing symptoms and the initial steps of identifying potential health conditions. This uncharted territory presents an intriguing research opportunity, as it involves not only enhancing the accuracy of medical inquiries but also making healthcare information more accessible and comprehensible to a broader spectrum of users. Through the identification and resolution of this research gap, new avenues for the utilization of AI-powered language models in the healthcare sector can be found, potentially transforming how individuals access initial medical advice.

This research aims to bridge the identified gap by proposing and developing a pioneering Symptom to Disease Chatbot, empowered by the sophisticated Llama 2 13b pre-trained language model. This chatbot is intended to function as a valuable tool for individuals seeking preliminary medical insights. Through the utilization of the extensive knowledge stored within LLAMA2 and its fine-tuning with specialized medical datasets, the chatbot aspires to offer users informed responses regarding potential health conditions based on reported symptoms. The integration of AI and medical expertise holds the potential to address the critical need for accurate and accessible healthcare information, thus enabling users to make well-informed decisions about their health. This research seeks to harness the synergy between AI technology and medical knowledge, contributing to improved healthcare accessibility and augmenting the initial steps of medical diagnosis. Ultimately, the development and evaluation of this Symptom to Disease Chatbot may pave the way for a more proactive and empowered approach to personal health management, while also shedding light on the broader implications of AI in healthcare settings.

2. Method

2.1. Data preparation

The data shown in Figure 1 utilized in this project was sourced from Kaggle [3], a prominent open-source platform for machine learning enthusiasts and practitioners. The primary dataset, Symptom2Disease, was meticulously curated by the diligent user Niyar R Barman. This dataset is a treasure trove of medical information, featuring a comprehensive compilation of 24 distinct diseases, each accompanied by detailed descriptions of 50 associated symptoms. In total, Symptom2Disease comprises a dataset of 1200 datapoints, elegantly structured with two pivotal columns: "label" and "text." The "label" column designates the disease's name, while the "text" column houses an intricate description of the respective ailment [4].

DOI: 10.54254/2755-2721/55/20241433

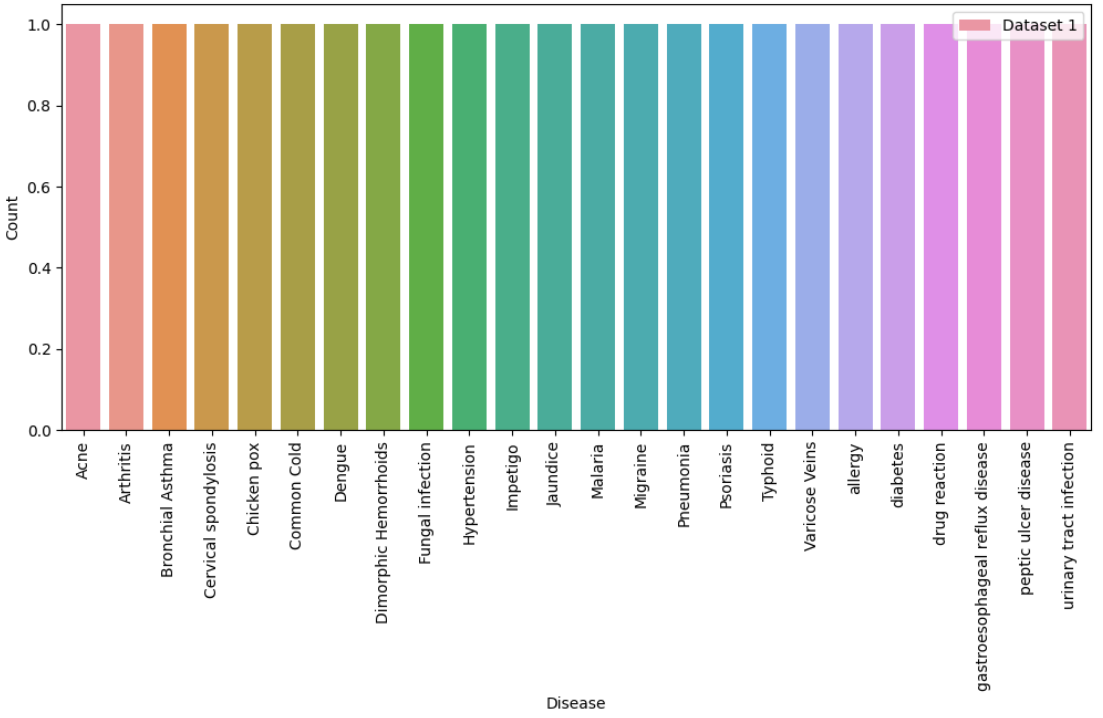

Figure 1. Distribution of Diseases in Dataset 1 (Photo/Picture credit: Original)

In the pre-processing phase, this project ingeniously amalgamated the disease names with their corresponding descriptions, uniting them within a single column. This strategic fusion facilitated the creation of a streamlined and coherent data frame, an essential precursor for the upcoming training endeavors.

However, it's important to acknowledge that this dataset, despite its value, comes with certain constraints. It is limited to encompassing only 24 distinct diseases, making it inadequate to comprehensively tackle a wider array of medical cases and provide informed solutions. Therefore, there is a pressing requirement for additional data to enhance the knowledge repository. The optimal solution involves augmenting the current dataset by seamlessly integrating it with supplementary datasets. This collaborative approach will not only expand the dataset's volume but also diversify its coverage, equipping people to effectively address a broader spectrum of medical scenarios. This project also leveraged another invaluable dataset, specifically the "diseases" dataset curated by Ujjwal Jindal and obtained from Kaggle [3]. Differing notably from the previous dataset, this one boasts a more extensive collection, encompassing a staggering 679 distinct disease types. Each disease in this dataset is accompanied by ten comprehensive descriptions. Notably, the dataset is structured into three CSV files, with the "training.csv" file being particularly well-suited for pre-processing to achieve a format closely aligned with that of the first dataset.

In its initial format, the primary column of this dataset represents the disease name, with the following columns containing symptom names indicated by binary values (0 for "absent" and 1 for "present"). The pre-processing methodology applied to this dataset entails the assembly of symptoms marked as "1" into coherent sentences and the consolidation of the ten descriptions into a single line. This harmonizing transformation brings the dataset's structure into closer alignment with the original dataset, simplifying the subsequent analysis and modeling endeavors.

Moving on to the next phase of data pre-processing, the subsequent step involves combining two separate datasets into a unified format suitable for training. The training data requirement is a JSON file containing rounds of conversation. This step entails transforming the two datasets into conversational formats and merging them into a single JSON file that fulfills the prerequisites for training[5].

DOI: 10.54254/2755-2721/55/20241433

This step's significance should not be underestimated, as the accurate preparation and organization of data directly impact the model's performance and training outcomes. Ensuring data accuracy and consistency is, therefore, a crucial task. Upon completing this stage, a comprehensive JSON file containing the necessary conversation data for training will be at the disposal.

2.2. The base model Llama 2

Llama is a large language model optimized for dialogue use cases. Llama 2 is an updated version of Llama 1, trained on a new mix of publicly available data. The models range in scale from 7 billion to 70 billion parameters, and have released variants of Llama 2 with 7B, 13B, and 70B parameters. This study has also trained 34B variants but are not releasing. The official paper of Llama 2 provides a detailed description of the technical aspects of Llama 2, including the new mix of publicly available data used to train the models, the increased size of the pretraining corpus by 40%, the doubled context length of the model, and the adoption of grouped-query attention [2].

Llama also allow programmers and users to use it as a pre-trained model and use data to fine-tuning it for specific task, such as law and paper summarization. QLoRa is a common method of fine-tuning the Llama 2 model [5]. In this project, QLora fine-tuning way is also used to link the dataset with the LLM model's output [6, 7].

3. Results and discussion

The model training process has progressed smoothly, and the data appears to be functioning optimally, leading to successful fine-tuning of the Llama 13b model. The training process is reflected in the step-loss graph displayed below. However, it's important to note that this graph has some limitations due to the relatively small number of training steps, which may not fully capture the nuances of the training process.

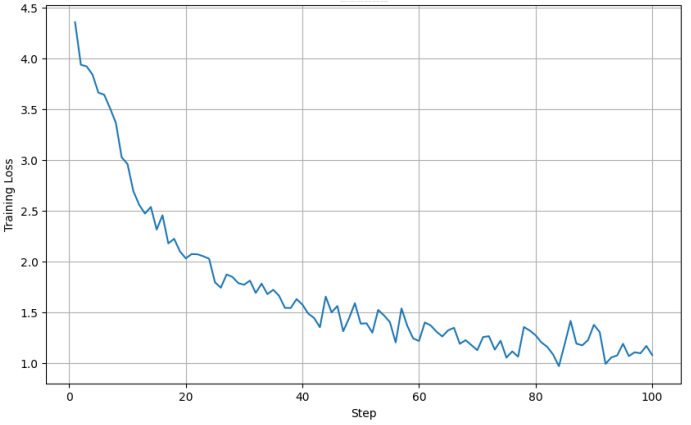

Figure 2. Training step-loss figure (Photo/Picture credit: Original)

The final loss shown in Figure 2 is around 1.1, as the training environment limitation, this project only uses one A100 device for model training, which doesn't have the capability to deal more steps and bigger dataset [8].

As depicted in the graph, the training process appears to exhibit a good fit, indicating that the model is learning effectively. Nevertheless, the limited number of training steps, only 100 in this case, means that the graph provides only a partial view of the training process.

To enhance user interaction and accessibility, this project has integrated an open-source Gradio interface [9] into this project's system. This user-friendly interface serves as a bridge between the model and the end-users, facilitating a seamless and intuitive interaction experience. Overall, while the initial training results are promising, further exploration and experimentation with a larger dataset and more training steps may yield even more robust performance. In addition, more advanced modules e.g. attention mechanisms may be considered in the future for further improving the performance [10].

DOI: 10.54254/2755-2721/55/20241433

4. Conclusion

This research embarks on a transformative journey at the intersection of Artificial Intelligence and healthcare, aiming to bridge a critical gap in the field of medical diagnostics. The development of a pioneering Symptom to Disease Chatbot, powered by the sophisticated Llama 2 13b pre-trained language model, holds immense promise. By leveraging extensive medical datasets and fine-tuning AI models, this project aspires to offer users valuable preliminary medical insights based on reported symptoms, thereby enhancing healthcare accessibility and empowering individuals to make informed decisions about their well-being. The amalgamation of AI technology and medical knowledge not only represents a significant leap forward in healthcare but also signifies the potential of AI in addressing real-world challenges. While this project's initial training results are promising, there remains room for further exploration with larger datasets and more training steps to enhance the model's robustness. This research serves as a stepping stone toward a future where AI plays a pivotal role in improving healthcare outcomes, revolutionizing how individuals access initial medical advice, and shedding light on the broader implications of AI in healthcare settings.

References

[1] Floridi L and Chiriatti M 2020 GPT-3: Its nature, scope, limits, and consequences. Minds and Machines, 30, 681-694.
[2] Touvron H et al 2023 Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.
[3] Ujjwal J 2023 diseases https://www.kaggle.com/datasets/jindalujjwal0720/diseases
[4] Niyar R B 2023 Symptom2Disease, https://www.kaggle.com/datasets/niyarrbarman/symptom2disease
[5] Dettmers T 2023 Qlora: Efficient finetuning of quantized llms. arXiv preprint arXiv:2305.14314.
[6] UD 2023 Instruction fine-tuning Llama 2 with PEFT's QLoRa method, https://ukey.co/blog/fine tune-llama-2-peft-qlora-huggingface/
[7] philschmid 2023 Efficient Large Language Model training with LoRA and Hugging Face, https://www.philschmid.de/fine-tune-flan-t5-peft
[8] baeldung 2023 Training and Validation Loss in Deep Learning, https://www.baeldung.com/cs/training-validation-loss-deep-learning
[9] gradio 2023 Quickstart, https://www.gradio.app/guides/quickstart
[10] Qiu Y et al 2022 Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training Biomedical Signal Processing and Control 72 103323