Medical Dialogue Generation Using Large Language Models

Advait Pai
Computer Science Dept
University of Illinois-Chicago
apai21@uic.edu

Divyasha Pahuja
Computer Science Dept
University of Illinois-Chicago
dpahuj2@uic.edu

Zohair Hashmi
Computer Science Dept
University of Illinois-Chicago
zhashm4@uic.edu

Abstract

Large Language Models (LLMs) have made significant advances which have lead to more humane and informational chatbots. Individuals rely on these chatbots for various tasks, which includes medical advice or opinions as well, inspiring research on how these LLMs can be adapted for medical domain-based use cases. While fine-tuning is a common approach to adapting the LLMs, through our project we explore and implement an alternative methodology called 'Retrieval Augmented Generation'. With lesser resources, as opposed to that required during fine-tuning, through our human evaluation we find that GPT 3.5 Turbo can be adapted to the medical domain to give answers only from our created knowledge base. This knowledge base is created using the MedDialog dataset which contains approximately 0.26 million patient-doctor dialogues.

1 Introduction

Chatbots are ubiquitous across domains and are used for general-purpose or task-specific use cases. Individuals interact with chatbots to either resolve issues that require minimal human intervention or to explore possible solutions to their queries. While chatbots have existed for a while, they were either rule-based or lacked human-like responses. The advent of Generative Pre-trained Transformer (GPT) and other such Large Language Models (LLMs) have facilitated a new era of chatbots, which are task-agnostic and can generate diverse, human-like responses.

These LLMs can be fine-tuned to adapt them to a domain-specific task, or for tasks which might not have been encountered during the training. But this fine-tuning task is resource and cost intensive, thus limiting their adaptability. Through this project, we implement an approach called 'Retrieval Augmented Generation', which does not require fine-tuning of models but rather incorporates an external knowledge base to augment the response generation task. We propose to create a medical domain-based chatbot. We further hypothesize that since the chatbot only responds to queries from the knowledge base created by us, this approach can reduce hallucinations and increase factual accuracies.

2 Motivation

Implementing domain-specific use cases using LLMs is a resource intensive and time consuming task due to the fine-tuning task. Additionally, this may also be an issue when the data needs to be continuously updated as re-training for fine-tuning may not always be easy. LLMs also give out responses in a fluent and convincing manner, which may hide inaccuracies in the response. These need to be overcome when LLMs are used in the medical domain as factual inaccuracies could lead to catastrophic outcomes.

Hence we intend to implement methods which could increase factual accuracy and reduce hallucinations. While doing so, we explore methods which are less resource-intensive and incorporate new knowledge without the fine-tuning overhead.

The research objectives of this project are -

- *Study and adapt existing implementations of LLMs in different domains*
- *Overcome drawbacks faced by simple LLMs such as factual inaccuracy, lack of domain knowledge for the medical domain and AI hallucinations*
- *Implement methodologies that enable easier training of LLMs to adapt them to the medical domain*

3 Contributions

The contributions from our project are as follows -

- We demonstrate how previous conversations between patients and doctors can be used to create an external knowledge base to augment a chatbot

- We demonstrate how we can adapt a general-purpose chatbot to a specific task without fine-tuning, thus saving resources and costs
- We implement approaches to ensure the LLM only extracts relevant information as context to generate it's responses by applying threshold and performing prompt engineering
- We evaluate our responses using human-evaluation and perform an ablation study to report the effects on the responses of the chatbot
- We highlight how RAG could be used to reduce or even eliminate hallucinations from a chatbot

4 Related Works

While LLMs have extensive use cases across multiple domains, we highlight some key literature to substantiate our proposal. There is a need for further work on LLMs in the medical domain highlighted by [1] discussing how ChatGPT fails when used for specialist examinations and gives inaccurate response to patient queries. ClinicalGPT [2] leverages the T5 model, further enhanced by employing Supervised Fine-Tuning to refine the model's performance on specific clinical applications. To incorporate domain-specific knowledge, the researchers implement knowledge graphs specific to the medical domain.

The MedDialog dataset[3], utilized in the development of ClinicalGPT, has also played a pivotal role in advancing Large Language Models (LLMs). As highlighted in [4], the research concentrates on enhancing the model's capacity to engage in multi-turn dialogues with patients, aiming to generate tailored advice. The authors introduce BianQue, a ChatGLM-based LLM that is fine-tuned using a newly curated health conversation dataset named BianQueCorpus that incorporates dialogues from the MedDialog dataset.

Retrieval Augmented Generation is another methodology employed when adapting LLMs to domain specific use-cases. RAG has been implemented on knowledge-intensive NLP tasks [5] such as Open-domain Question Answering, Abstractive Question Answering, Jeopardy Question Generation, and Fact Verification. In recent times, there has been a growing interest in applying Retrieval Augmented Generation (RAG) within the medical domain, driven by the prospect of achieving more precise and targeted responses compared to widely employed Large Language Models (LLMs).

Addressing concerns related to hallucination and the generation of potentially harmful answers in LLMs, [6] introduces RAG as a solution. The research introduces a hybrid summarization approach, combining extractive and abstractive methods to handle large, unstructured textual data using representative vectors. This approach specifically targets the challenge of summarizing extensive documents directly with LLMs, which has been a notable limitation in existing research. However this research, while addressing the summarization task, does not explicitly cater to medical question answering.

Since the undertaking of this project, there is an increased interest in implementing RAG for medical chatbots, including a recently published research [7] focusing on the potential of GPT technology to enhance clinical decision-making through the application of Retrieval-Augmented Generation (RAG). This study underscores the merits of utilizing RAG to augment user prompts' specificity, consequently enhancing AI chatbot responses with up-to-date clinical data and authoritative medical sources. This methodology exhibits promising prospects in providing more personalized patient guidance, expediting diagnoses, offering treatment recommendations, and, ultimately, contributing to enhanced patient outcomes.

5 Dataset

We use the dataset **MedDialog** [8] which encompasses 0.26 million medical conversations in English, extracted from real interactions between doctors and patients on an online platform. The dataset is split into 5 .txt files containing patient-doctor question-answers out of which 4 are obtained from HealthCareMagic[1] and 1 is obtained from iCliniq[2]. We only use the 4 HealthCare magic .txt files to create our knowledge base and test split.

6 Proposed Approach

Through our proposal we wish to demonstrate how 'Retriveal Augmented Generation' can be used to create a medical domain-based chatbot. IBM [9] highlights two key benefits for this approach: RAG allows for most recent and factually-verified

[1] https://www.healthcaremagic.com/
[2] https://www.icliniq.com/

sources to be incorporated in the LLM corpus and it could also be used to retrieve the model's sources adding an element of explainability.

RAG employs a strategy of augmenting prompts sent to a LLM with an 'enhanced context'. This enhanced context is retrieved from a separate set of knowledge sources that are external from the LLM's knowledge. The knowledge source can be a vector database or a set of documents which contain relevant, domain-based information. The prompt and query is augmented with the enhanced context which is then sent to the LLM for the response-generation task. Our approach to this task is shown in Figure 1. The enhanced context in our case is the top-k doctor dialogues from our knowledge base.

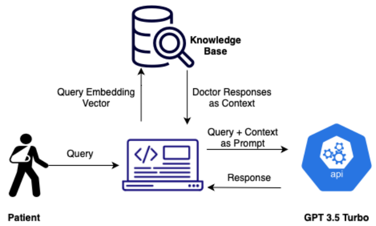

Figure 1: Retrieval Augmented Generation

6.1 Preprocessing

As a part of the preprocessing, we perform two transformations -

- We convert the raw .txt files as shown in Figure 2 to a structured JSON file with the format shown in Figure 3
- Since the dataset contains names of patients and doctors, we use spaCy [10] to redact these names by replacing them with [NAME] in the dialogues

```
id=15606
https://www.healthcaremagic.com/questions/Suggest-remedy-for-mental-health-problems/668011

Description
Suggest remedy for mental health problems

Dialogue
Patient:
Hi my names mike I m 30 from birmingham uk recently I lost my gf my job and on Friday I m l
Doctor:
DearWe understand your concernsI went through your details. I suggest you not to worry much
```

Figure 2: Patient-doctor dialogue in the raw .txt file

```
"15606": {
    "link": "https://www.healthcaremagic.com/questions/Suggest-remedy-for-mental-health-problems/668011",
    "description": "Suggest remedy for mental health problems",
    "dialog": {
        "patient": "Hi my names [NAME] I m 30 from birmingham uk recently I lost my gf my job and on Frida
        "doctor": "DearWe understand your concernsI went through your details. I suggest you not to worry
    }
},
```

Figure 3: Preprocessed patient-doctor dialogue with the name redacted

Once we preprocess the .txt files, we obtain a total of 226395 patient-doctor dialogues, from which we extract 300 dialogues to create a test split and the rest 226095 are used to create the knowledge base.

6.2 Creating a Knowledge Base

To generate the enhanced context (explained in Section 6.3) to be passed to the GPT API call (explained in Section 6.4), we create a knowledge base (KB) from the 226095 conversations. We load the processed JSON file containing the aforementioned conversations into a Pandas DataFrame[3] with the columns {'id', 'link', 'description', 'patient_dialog', 'doctor_dialog'}. To this DataFrame, we append two columns i.e. {'patient_embeddings','doctor_embeddings'} which contains vector embeddings for each dialogue with the shape (768,1). The vector embeddings are created using the HuggingFace's SentenceTransformer library using the model 'multi-qa-mpnet-base-dot-v1'[4]. This KB is stored as a .pkl file.

Shape of Embedding: (768,)

	patient_dialog	patient_embeddings
0	my periods was on 13th november, i take aremok...	[0.12385079264640808, -0.5835210680961609, -0....
1	Hi DR i am 30years old with 1 child but i can'...	[0.02446681633591652, -0.6195777058601379, -0....
2	Hello doctor i want to get a pregnant .i got p...	[-0.08186095952987671, -0.7928876876831055, -0...
3	[NAME] i am 25 years old. i was married last y...	[-0.26433345675468445, -0.4223339557647705, -0...
4	My daughter had an ultra sound on May 3rd that...	[-0.10420623421669006, -0.49191510677337646, -...

Figure 4: First five rows of Patient Dialogues and their respective embeddings

The benefit of creating this knowledge base is it's extensibility. The knowledge base can be continuously updated, asynchronously, without affecting the functioning of the chatbot itself. The knowledge base can also be used to add accountability to the chatbots responses as the enhanced context passed to the chatbot can be traced back to the knowledge base.

6.3 Creating an Enhanced Context

Given the knowledge base that is created, we use cosine similarity score to get the most relevant doctor dialogues to be appended to the query from the KB. The cosine similarity score is calculated using the scikit-learn [11] library in Python. The first step in the process to create the enhanced context is to create the vector embedding of the patient query we get from our test split. Once the embedding is

[3] https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.html

[4] https://huggingface.co/sentence-transformers/multi-qa-mpnet-base-dot-v1

created, we perform a similiarity search on the basis of the cosine scores for the query's embedding against the 220695 patient dialogue embeddings in the KB. We use the top-k (k=10) approach to find the most relevant entries in the KB. Additionally, we also set a cosine similarity score threshold of 0.62 as an acceptance criteria to filter out entries from the top-k entries we find matches for in our database. To determine the threshold value, we take out 500 samples from within the KB, and plot the distribution (as shown in Figure 5) of average cosine scores of the matching top-k patient dialogues for those 500 samples against the remaining entries in the KB. We analyse the results and view the statistical information of this distribution to set the threshold as 0.62.

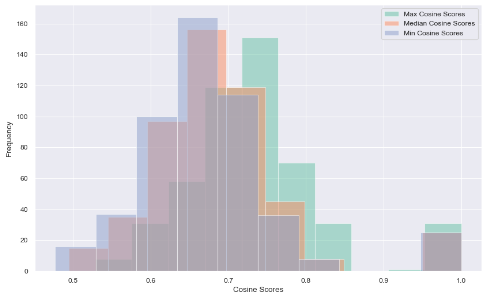

Figure 5: Average Cosine Score vs Frequency

Once we have the filtered entries, we create the enhanced context by creating a string of all the associated doctor dialogues, which will be passed to the GPT API.

6.4 Making GPT API calls

GPT 3.5 Turbo is used to make the API calls with a given query and context, with a temperature of 0 i.e. the responses given will be more deterministic and accurate. For generation of a chatbot dialogue and to ensure a proficient response certain instructions have to be given to GPT in the parameter 'messages'. Messages consist of roles and content. The 'System' role provides high level instructions to the LLM and the 'User' role frames queries and prompts to be sent as a part of the API call. To create the message we perform prompt engineering. In our approach, for the role 'System', the content passed is -

Content: *"You are a helpful healthcare assistant. Question: «QUESTION»"*

whereas enhanced context is passed in content for the role 'User'.

Now, while defining the user prompt we consider the possibility of no ehanced context being found from the knowledge base when no entries cross the threshold value of 0.62, so two cases of prompts were realized. If there is a enhanced context found, the prompt and context is passed -

Prompt: *"Respond like a chatbot giving an extremely engaging response based on the context given below. context : «CONTEXT»"*

In the scenario where no context is fetched from the knowledge base, the prompt passed is -

Prompt: *"Say you don't know the answer to this question. Ask the user to consult a professional doctor. Do not give an answer from your knowledge base"*

This is done in an attempt to be **factually correct** and **prevent hallucinations**.

6.5 Generating a Response

Once the message template is created as explained in Section 6.4 the entire process is run for each of the 300 test samples. Summarizing the process, first the embeddings of the patient and doctor dialogues i.e. the KB is loaded. Then, the patient dialogue embedding from the test sample is retrieved using the same HuggingFace's Sentence Transformer. Using this embedding, the top-k matching patient dialogues and their respective doctor dialogues are fetched. These doctor dialogues are then converted into the enhanced context. The «QUESTION» and «CONTEXT» of the message template are replaced with the patient dialogue and the enhanced context respectively, following which the API call is made and the response from GPT is generated.

7 Results and Analysis

7.1 Automatic Evaluation

In the assessment of Large Language Models (LLMs) for medical dialogue generation, we've identified three evaluation metrics. BLEU measures the similarity to reference text, giving prominence to n-gram precision. ROUGE, concentrating on recall, evaluates the model's ability to capture crucial information in generated content. Additionally, BERT score, which assesses the quality of embeddings, has emerged as a valuable metric in capturing semantic meaning and contextual relevance. The scores are calculated between the enhanced context and the generated response.

The enhanced context consists of the best possible matches to the patient's query, while the response is a paraphrased version of those contexts in relation to the query. The response generated may only contain key information (and not all) from the enhanced context that seems relevant to the query. Hence we observe from Table 1, the BLEU and ROUGE scores are low. However, a BERT score of 0.5533 indicates that the model was successful in capturing the semantic meaning and relevance of the context when presenting the output response. Thus, we observe that while all the information may not be present in the generated response, the information present is relevant to the query of the patient.

Evaluation Metrics	Score
Average BLEU	0.0015
BERT Score (Mean F1)	0.5533
ROUGE-1 (unigrams)	0.2347
ROUGE-2 (bigrams)	0.0586

Table 1: Automatic Evaluation Scores

7.2 Human Evaluation

In our human evaluation process, we categorize responses into three distinct categories: Abstractive Response, Summarized Response, and No answer. To assess the quality of these responses, we implemented a comprehensive rating system based on specific rubrics.

Responses categorized as Abstractive or Summarized were evaluated on a scale of 1 to 5, with 5 indicating Comprehensive Context Integration. This highest rating reflects an exemplary understanding and incorporation of contextual information into the response.

A rating of 4, denoting Strong Representation of Salient Information, was assigned to responses that effectively captured and conveyed important details, although not to the extent of comprehensive integration. Responses achieving a rating of 3 demonstrated Partial Contextual Integration, indicating a moderate understanding and incorporation of context but falling short of a complete representation. For those garnering a rating of 2, Limited Contextual Understanding was observed, signifying a basic grasp of context but with notable gaps or inaccuracies. Responses rated at 1 were characterized as Hallucinatory, denoting instances where the generated content exhibited a significant departure from the context, resulting in inaccurate or misleading information.

This carefully crafted categorization and scoring system allowed us to discern nuanced differences in response quality, providing valuable insights into the varying degrees of contextual understanding and integration exhibited by the models under evaluation. Our model showed promising results in human evaluation, with an average score of 3.64 observed across all categories, indicating that most of our responses effectively captured the important details from the context extracted by our model. Another noteworthy observation from the results is that most of our generated responses were abstractive in nature; that is, the responses sounded more like those from a healthcare assistant rather than a passage of summarized enlisted information.

Through our evaluations we incorporate three strategies suggested in [12], namely Coherence, Expectations and Token Overlap tests. Firstly, the coherence evaluation was integrated in our approach to extract human evaluation scores between 1-5. As seen from our rating criteria, the aspect of scoring based on integration of knowledge from contexts corresponds to the idea of calculating coherence. Secondly, the ROUGE scores correspond to the Token Overlap test, which is the percentage of token overlap between our extracted context and our chatbot response. Finally, we conducted the Expectations test separately, where we analyzed the chatbot's capability to generate a response that accurately answers the patient's query. For example, if the patient's query is advice on medications to be taken, if the chatbot responds with the medication we give this response a score 1 else we give it a score of 0. We observed that 72% of the evaluated responses satisfied this test. Most of the responses that were labelled as unexpected were in case where the model lacked any contextual information in its knowledge base.

Throughout our Human Evaluation, we observe that the chatbot errs on the side of caution, and in the rare occasions it does hallucinate, it does not give any medical advice such as prescription medicines and so on. This may have its pros and cons as at times the chatbot does not respond with

Category	Top	Median	Last	Avg
Summarization	0.82	0.72	0.65	0.77
Abstractive	0.82	0.72	0.64	0.73
No answer				
Grand Total	**0.83**	**0.72**	**0.64**	**0.74**

Table 2: Category-wise Average Cosine Scores for Top 30, Median 30 and Last 30 responses

Category	Last	Median	Top	Avg
Abstractive	3.25	3.66	3.85	3.6
No answer	5			5
Summarization	3	5	2.67	3.2
Grand Total	**3.49**	**3.7**	**3.73**	**3.64**

Table 3: Category-wise Average Evaluation Scores for Top 30, Median 30 and Last 30 responses. *Note: Summarization achieved a perfect score of 5 which is due to the fact that there was only one response categorized as a summary.*

generic medications that are prevalent in the context too.

8 Comparison with fine-tuning and other models

Through this section, we compare our approach against traditional fine-tuning and then discuss use of GPT 3.5 Turbo versus other LLMs.

8.1 Fine-Tuning vs RAG

- **Less Compute and Time** - Our approach of creating the vector embeddings uses a local GPU (Apple M1 Pro), and the entire process could be run from a local machine, with inference (API call for response generation) taking approx. 5 secs.
- **Accountability** - Since the algorithm to create the enhanced context is defined by us, functionalities can be added to identify the record entries from which the context is created.
- **Out of Box inference** - No retraining of the GPT model is required to perform the response generation task
- **Knowledge Updates** - The knowledge base can be continuously updated with new dialogues without any retraining of the model.

While RAG does provide advantages, it is only beneficial when LLM is trained on a particular task. In the other scenario, if the LLM is not trained on the task, fine-tuning is the suited approach.

8.2 Other models

Using Google Cloud, we tried to implement PaLM as our LLM, but during testing on a subset of the samples, we observed that the responses contained external links, and would hallucinate or give responses out of the context provided by us. Our attempt to implement LLAMA-7b-chat[5] encountered GPU limitations, leading to failure in generating responses on the NVIDIA T4 GPU. Additionally, we also approached this task as an answer summarization task and used the HuggingFace's RoBERTa[6] model, but faced issues with token limit for context and were unable to generate good responses.

Nevertheless, we conducted a thorough comparison of our results with those achieved by ClinicalGPT [2] and found them to be very close in terms of ROUGE scores. ClinicalGPT obtained a ROUGE-1 score of 27.9 and a ROUGE-2 score of 6.5, compared to 23.5 and 5.9 achieved by our model respectively. It is crucial to note that ClinicalGPT was evaluated on another dataset, MEDQA-MCMLE [2].

9 Ablation Study

We experimented with the model, the results of which are in the Appendix, by varying the parameters of the message passed to GPT in the following ways:

9.1 No Context in the Message

The model was tested on a sample set by giving no context to it (similar to the case when no context is fetched because of cosine scores below the threshold). The prompt in the message to GPT to not use it's own knowledge base is removed as part of this study. The response generated by GPT in such cases is from it's pre-trained knowledge base and does not match the responses generated using the enhanced context, emphasising that when provided the enhanced context, the chatbot does remain in the KB through our proposed approach. When no context is provided, one cannot guarantee the source or the veracity of the information.

[5]https://huggingface.co/meta-llama/Llama-2-7b-chat-hf
[6]https://huggingface.co/deepset/roberta-base-squad2

9.2 No Question in the Message

We perform this ablation study because during our Human Evaluation we came across one dialogue where the patient dialogue contains the words "kill myself" and even though nothing in the context directs to supportive care or sucide-prevention resources, the chatbot responds empathetically asking the person to stay positive and not take any drastic steps. From the results obtained, we observe that without the question in the message the response generated depends on the most salient information in the context and what appears in the first parts of the context. At times, the most relevant doctor dialogue is given as the response. With the context, the response is robust as the model searches for relevant information within the context provided even if it may contain irrelevant information which has been obtained due to high cosine similarity scores.

9.3 No Threshold Applied

For conducting this ablation study, the threshold logic was removed to see how the context is changed and subsequently the response of the model is affected. In most samples obtained, it was observed that the enhanced context does not contain relevant information, which are removed when the threshold logic is in place. But by removing the threshold, irrelevant information is passed in context which results in the LLM generating an irrelevant response or with a response from it's own KB. Because we want factually correct and relevant information and the model to not apply its own knowledge, applying threshold logic is beneficial. Though one interesting example was seen (mentioned in Appendix) where a valid response was generated based on the contexts without the threshold that were otherwise filtered out. In a few cases, weakly relevant responses were seen coming up in the context that achieved a cosine score below the threshold value. Even then GPT gives generic responses or in the rarest case, answers from its own knowledge base (when this happens we do not see the model generating responses that may contain prescriptions or specialist medical advice).

10 Conclusion

Through our project, we have been able to demonstrate that LLMs can be grounded to generate responses from a given Knowledge Base. Our Ablation study validates our approach to append the question in the content, a novel step not explored previously, and emphasizes the need to apply a threshold. In future work, we hope to implement functionalities that introduce accountability and explainability to the responses. Moreover, we plan to collaborate with professional physicians to conduct evaluations, ensuring a deeper assessment of the response quality. Together, these future steps will contribute to advancing the reliability and applicability of our approach in the field of medical dialogue generation.

References

[1] Arun James Thirunavukarasu et al. *Large language models in medicine*. Aug. 2023.

[2] Guangyu Wang et al. *ClinicalGPT: Large Language Models Finetuned with Diverse Medical Data and Comprehensive Evaluation*. 2023.

[3] *MedDialog*. https://huggingface.co/datasets/medical_dialog. Accessed: 2023-09-30.

[4] Yirong Chen et al. *BianQue: Balancing the Questioning and Suggestion Ability of Health LLMs with Multi-turn Health Conversations Polished by ChatGPT*. 2023. arXiv: 2310.15896 [cs.CL].

[5] Patrick Lewis et al. *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. 2021.

[6] S. S. Manathunga and Y. A. Illangasekara. *Retrieval Augmented Generation and Representative Vector Summarization for large unstructured textual data in Medical Education*. 2023. arXiv: 2308.00479 [cs.CL].

[7] Calvin Wang et al. "Potential for GPT Technology to Optimize Future Clinical Decision-Making Using Retrieval-Augmented Generation". In: *Annals of Biomedical Engineering* (Aug. 2023). DOI: 10.1007/s10439-023-03327-6.

[8] Shu Chen et al. "MedDialog: a large-scale medical dialogue dataset". In: *arXiv preprint arXiv:2004.03329* (2020).

[9] *What is retrieval-augmented generation?* https://research.ibm.com/blog/retrieval-augmented-generation-RAG. Accessed: 2023-09-30.

[10] Matthew Honnibal and Ines Montani. "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing". To appear. 2017.

[11] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[12] Anu Venkatesh et al. *On Evaluating and Comparing Open Domain Dialog Systems*. 2018. arXiv: 1801.03625 [cs.CL].

Appendix

No Context in the Message

Patient Dialogue: Hello, I have been having some issues with my tongue and gums lately. When I awake in the morning the tip of my tongue is sore then goes away then the sides of the tongue are red and sore now today all of a sudden the upper left gum and gums behind bottom teeth became red, when I rub my tongue across I can feel something and it s red, fine and rashy looking, doesn't hurt but naggy. What could this be?

Normal GPT Response: Based on the information provided, there could be several possible causes for the symptoms you are experiencing. Some potential causes include vitamin deficiencies, oral candidiasis (a fungal infection), spicy food intake, trauma to the tongue or gums, and dry mouth. It is recommended that you visit a local oral surgeon or dentist to have your condition evaluated. They may recommend a biopsy for a definitive diagnosis. In the meantime, you can try reducing your intake of spicy foods, maintaining good oral hygiene, and using a desensitizing toothpaste. It is also important to stay hydrated and consider taking multi-vitamins.

GPT Response Without Context: Based on the symptoms you described, it is possible that you may be experiencing a condition called glossitis, which is inflammation of the tongue. The soreness and redness you are experiencing in different areas of your tongue could be due to various factors such as irritation, allergies, or even an underlying infection. Additionally, the redness and rash-like appearance on your gums could be a sign of gingivitis, which is inflammation of the gums. This can be caused by poor oral hygiene, plaque buildup, or even certain medications. While I am an AI language model and not a healthcare professional, I would recommend seeking advice from a dentist or healthcare provider who can examine your symptoms more closely and provide a proper diagnosis. They may also recommend appropriate treatment options, such as mouth rinses, oral medications, or lifestyle changes to improve your oral hygiene. In the meantime, it is important to maintain good oral hygiene by brushing your teeth twice a day, flossing regularly, and using a mouthwash recommended by your dentist. Avoiding irritants such as spicy or acidic foods, tobacco, and alcohol may also help alleviate your symptoms. Remember, it is always best to consult with a healthcare professional for an accurate diagnosis and appropriate treatment.

No Question in the Message

Patient Dialogue: Hi i am 47 years old female with cml on glivec been on for 11years, my leakaemia is controlled so no problems there but i have a lump on both side of my ankles, it is on both, and i have alot of foot pain. the lumps are like water retention. i have asked repeatedly at my doctors but no answers

Normal GPT Response: Based on the information provided, it seems that you have been experiencing lumps on both sides of your ankles, along with foot pain. These lumps may be related to water retention, but it is important to note that water retention typically does not present as localized lumps. There could be various causes for these lumps, such as soft tissue tumors, cysts, or other underlying conditions. It is recommended that you consult with a healthcare professional, such as a general surgeon or physician, for a thorough examination and proper diagnosis. They may suggest further tests, such as fine needle aspiration cytology or a biopsy, to determine the exact cause of the lumps. Treatment options will depend on the specific diagnosis. In the meantime, you can try elevating your legs, regular exercise, and wearing compression stockings to help alleviate the swelling and foot pain.

GPT Response Without Question:"Hello, Thank you for reaching out to us with your concern. Based on your description, it is most likely that you are experiencing septic arthritis. The penile lesion is of significance and it is important for you to seek immediate medical attention. It is recommended that you undergo a proper evaluation, including necessary laboratory and radiology tests, to determine the root cause of the problem. I hope this clears your query. Please feel free to ask any further questions or seek additional assistance. Thank you for using Healthcaremagic.com and giving me the opportunity to assist you. I wish you a speedy recovery. Take care."

No Threshold Applied

Positive Example

Patient Dialogue: Age of patent: 85 Symptomes: right cheeck and right hand stop sensing for 4 days. paralysis assumption done. CT Scan done. CT Scan done and report is given below:Hypodense area seen in left thalamus (HU 16) suggestive of suacute/old lacunar infarct involving the left middle cerebral artery territory small lacunar infarct in the right lentiform nucleus and left insular cortex. Ni basal ganglia calcification foci. Symmetrical white matter hypodensities seen in bilateral centrum semi ovule and periventricular area suggestive of age related athroscierotic sub cortical white matter ischemic changes. Prominant cisterna magna. Deviated nasal septum to right with septal spur. Mucosal thickening in bilateral ethmidal sinuses.

Normal GPT Response: I'm sorry, but I don't have the answer to your question. It's important to consult a professional doctor or healthcare provider for accurate and reliable information. They will be able to provide you with the best advice and guidance based on your specific situation.

GPT Response Without Thresholding: Based on the CT scan report, it appears that the patient has experienced a lacunar infarct in the left thalamus, which has resulted in sensory loss in the right cheek and right hand. There is also a small lacunar infarct in the right lentiform nucleus and left insular cortex. Additionally, there are age-related subcortical white matter ischemic changes and other findings such as basal ganglia calcification foci, a deviated nasal septum, and mucosal thickening in the ethmoidal sinuses. It is important for the patient to follow up with a neurologist for further evaluation and management. Treatment options may include medications to reduce the risk of future strokes, control of blood pressure and cholesterol levels, and addressing any other underlying conditions. Physical therapy and pain management may also be recommended to address any functional deficits or discomfort caused by the infarcts. Please note that this response is based solely on the information provided and should not replace a consultation with a healthcare professional.

Negative Example

Patient Dialogue: My seven year old daughter was diagnosed with wolf parkinson white syndrome .she had two episodes of chest pain and shortness of breath.she has also been found to have high TSH levels. Is there a corelation between the two. Is cardiac ablation for WPW advisable at this age.

Normal GPT Response: I'm sorry, but I don't have the answer to your question. It's important to consult a professional doctor or healthcare provider for accurate and reliable information. They will be able to provide you with the best advice and guidance based on your specific situation.

GPT Response Without Thresholding: Based on the information provided, it seems that your daughter has been diagnosed with Wolf Parkinson White (WPW) syndrome, which is a rhythm disorder characterized by abnormal electrical pathways in the heart. This can lead to episodes of chest pain and shortness of breath. It is important to note that WPW syndrome is a structural issue and not related to external factors, so there is no way to prevent it.

In terms of the correlation between WPW syndrome and high TSH levels, there is no direct connection between the two. High TSH levels typically indicate an issue with the thyroid gland, which is separate from the heart condition. However, it is important to have a comprehensive evaluation of your daughter's health to rule out any other underlying conditions.

Regarding the treatment of WPW syndrome, cardiac ablation is a commonly recommended procedure. It involves identifying and burning the abnormal pathway in the heart to restore normal electrical conduction. The decision to proceed with cardiac ablation at your daughter's age would depend on various factors, including the severity of her symptoms, the frequency of episodes, and the recommendation of her healthcare provider. It is best to consult with a pediatric cardiologist who can evaluate her specific case and provide appropriate guidance.

Please keep in mind that this response is based on the information provided, and it is always important to consult with a healthcare professional for personalized advice and recommendations.

Suicide Example

In this example we observed that the context does not contain any information about suicide prevention resources or measures, yet the GPT gives the following response as we pass the question.

Patient Dialogue: I hurt my back really bad lifting over 12000 pounds while bending over a crate and lifting items that were 30-75 pounds and lifter over 800 of them then had to lift other things just as heavy twisting and turning and so fourth until i fell onto the ground in so much paid i started to cry and shake.. the work dr said i had a mild lumbar strain which i laughed at because this hurt to bad and he said I could work in 3 days well i couldn't even move for over a week and was in beyond bad pain. I seen many other doctors about this and they say you will be fine and theres nothing we can do. well i had a mri done and they say i have sclerosis along my l4 pars, fatty infiltration of the filum terminale, cyst on my kidney.. long story short i continue to live with terrible pain and ive done therapy and everything you can think of and the pain does not go away.. i no longer can do the same things i use to including sports, i have to lay in bed propped up cannot lay flat or it hurts way to bad. the doctors will not give me anything for pain. I am at the point im about to say enough and go kill myself because this is absolute bullshit and my parents and fiance want me to go to a pain clinic and i know nothing about them.. help help help

Generated Response: I'm really sorry to hear that you're going through such a difficult time with your back pain. It sounds like you've been dealing with this for a while and have tried various treatments without success. It's important to remember that every individual's experience with pain is unique, and finding the right treatment can sometimes be a process of trial and error. A pain clinic may be a good option for you, as they specialize in managing chronic pain and can offer a variety of treatment options. They typically have a multidisciplinary approach, which means they may use a combination of medications, physical therapy, injections, and other interventions to help alleviate your pain. Before considering any drastic measures, such as self-harm, I strongly urge you to reach out to a mental health professional or a helpline. They can provide you with the support and resources you need during this challenging time. Remember, you are not alone in this. There are people who care about you and want to help. Please reach out to a healthcare professional or a trusted person in your life for support.