

HealthBuddy - Your AI Health Companion

Project Report

PROJECT X
AT COMMUNITY OF CODERS
VEERMATA JIJABAI TECHNOLOGICAL
INSTITUTE, MUMBAI
SEPTEMBER - NOVEMBER 2022

ACKNOWLEDGEMENT

**We are extremely grateful to our mentors
Dishie Vinchhi and Om Doiphode for their
patience, support and guidance
throughout the duration of the project.**

We would also like to thank all the members of ProjectX, COC
for their timely support as well as for organizing ProjectX
and giving us a chance to work on this project.

Our team:

Raya Chakravarty
rayachakravarty5@gmail.com

Anushka Yadav
anushkasy2003@gmail.com

Tvisha Vedant
tvishavedant@gmail.com

TABLE OF CONTENTS

Sr. No	Title	Pg no
1.	Project Overview	4
2.	Deep learning and Neural Networks	5-6
3.	RNNs and LSTMs	7-9
4.	Transformer Architecture	10-13
5.	Decoder-Only Transformer Architecture	14-15
6.	Fine tuning pretrained models	16-19
7.	Dataset:A crucial aspect	20
8.	Initial Models	21-22
9.	Falcon 7b	23-27
10.	Llama-7b-chat	28-31
11.	Doc-based LLM	32-37
12.	Future aspects	39
13.	References	40

1. Project Overview

Description:

The HealthBuddy Chatbot is an innovative and user-friendly healthcare solution designed to provide individuals with personalized and reliable healthcare information and support.(The already existing AI based chat models are not trained specifically for health related tasks and hence they are not completely reliable.) This project aims to create a versatile chatbot that can offer assistance in various aspects of healthcare, including symptom diagnosis, mental health consultation, nutrition guidance, and more. The inspiration behind this project is to empower users to make informed healthcare decisions and promote overall well-being.

Why are chatbots important in healthcare?

According to a research in 2019, the most valuable features of using chatbots in healthcare include:

- Monitoring: Awareness and tracking of user's behavior, anxiety, and weight changes to encourage developing better habits.
- Anonymity: Especially in sensitive and mental health issues.
- Personalization: Level of personalization depends on the specific application. Some applications make use of measurements of:
 - Physical vitals (oxygenation, heart rhythm, body temperature) via mobile sensors.
 - Patient behavior via facial recognition.
- Real time interaction: Immediate response, notifications, and reminders.
- Scalability: Ability to react with numerous users at the same time

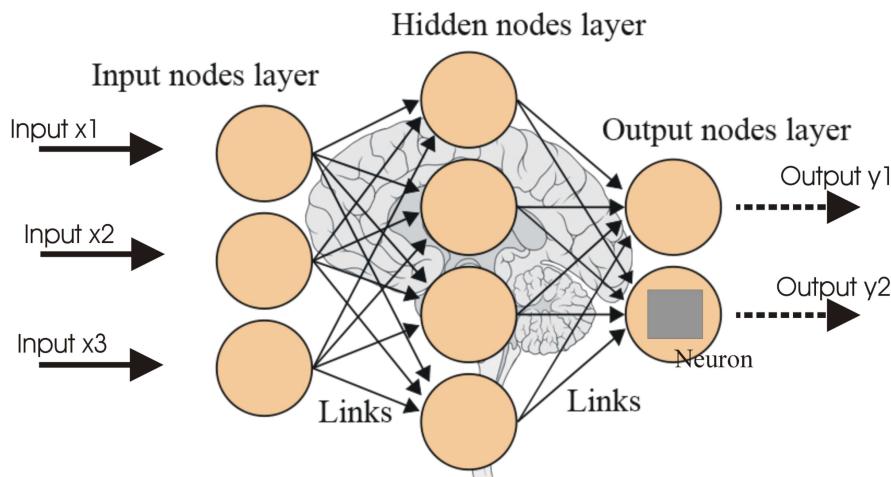
2. Deep Learning and Neural Networks

2.1 Deep Learning

Deep learning is a sub-field of machine learning dealing with algorithms inspired by the structure and function of the brain called artificial neural networks. In other words, it mirrors the functioning of our brains. Deep learning algorithms are similar to how the nervous system is structured where each neuron connects to each other and passes information.

In deep learning, each level learns to transform its input data into a slightly more abstract and composite representation.

2.2 Neural Networks



A neural network consists of many Nodes (Neurons) in many layers. Each layer can have any number of nodes and a neural network can have any number of layers.

The architecture of neural networks is made up of an input, output, and hidden layer.

i) Forward propagation -

As the name suggests, the input data is fed in the forward direction through the network. Each hidden layer accepts the input data, processes it as per the activation function and passes to the successive layer.

At each neuron in a hidden or output layer, processing happens in two steps:

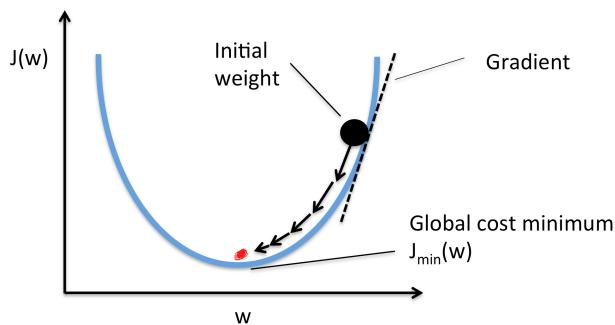
Preactivation - It is a weighted sum of inputs i.e. the linear transformation of weights w.r.t to inputs available. Based on this aggregated sum and activation function the neuron makes a decision whether to pass this information further or not.

Activation - The calculated weighted sum of inputs is passed to the activation function. An activation function is a mathematical function which adds non-linearity to the network. There are four commonly used and popular activation functions – sigmoid, hyperbolic tangent(tanh), ReLU and Softmax.

ii) Backpropagation -

Backpropagation is a process involved in training a neural network. It involves taking the error rate of a forward propagation and feeding this loss backward through the neural network layers to fine-tune the weights.

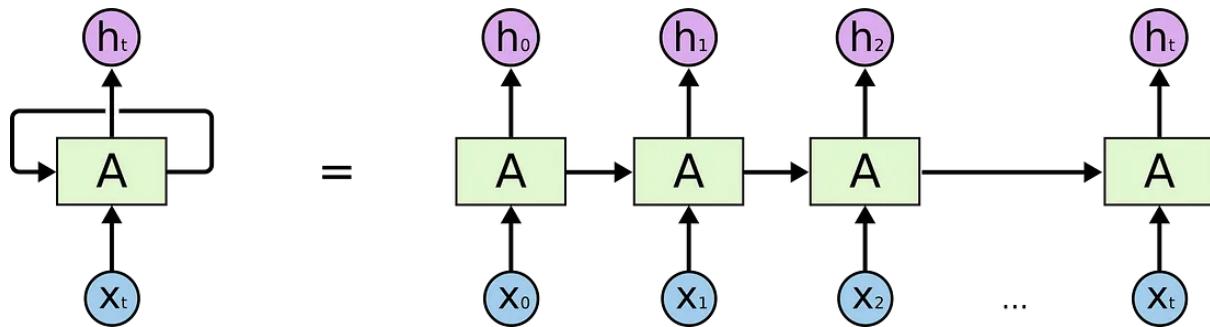
Gradient Descent - Gradient Descent is a process that occurs in the backpropagation phase where the goal is to continuously resample the gradient of the model's parameter in the opposite direction based on the weight w , updating consistently until we reach the global minimum of function $J(w)$.



3.RNNs and LSTMs

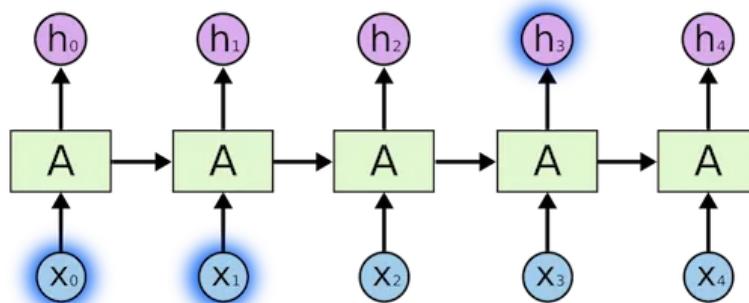
3.1 Recurrent Neural Networks(RNNs):

A Recurrent Neural Network can be thought of as multiple copies of the same network, A, each network passing a message to a successor. Consider what happens if we unroll the loop:



This chain-like nature shows that recurrent neural networks are clearly related to sequences and lists. In that way, if we want to translate some text, we can set each input as the word in that text. The Recurrent Neural Network passes the information of the previous words to the next network that can use and process that information.

In this case where the difference between the relevant information and the place that is needed is small, RNNs can learn to use past information and figure out what is the next word for this sentence.



Disadvantages of RNNs -

1. Vanishing And Exploding Gradients -

RNNs can suffer from the problem of vanishing or exploding gradients, which can make it difficult to train the network effectively. This occurs when the gradients of the loss function with respect to the parameters become very small or very large as they propagate through time.

2. Computational Complexity -

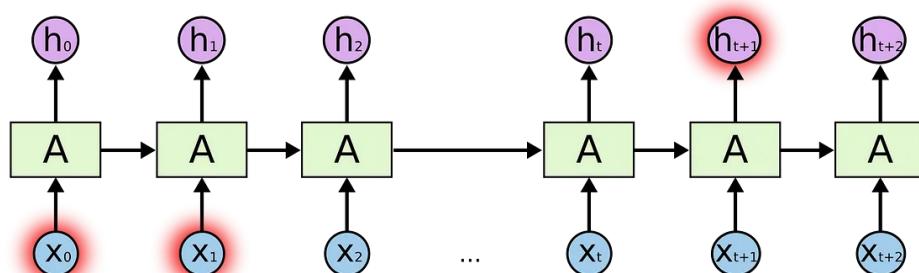
RNNs can be computationally expensive to train, especially when dealing with long sequences. This is because the network has to process each input in sequence, which can be slow.

3. Difficulty In Capturing Long-Term Dependencies

Although RNNs are designed to capture information about past inputs, they can struggle to capture long-term dependencies in the input sequence. This is because the gradients can become very small as they propagate through time, which can cause the network to forget important information.

4. Lack Of Parallelism

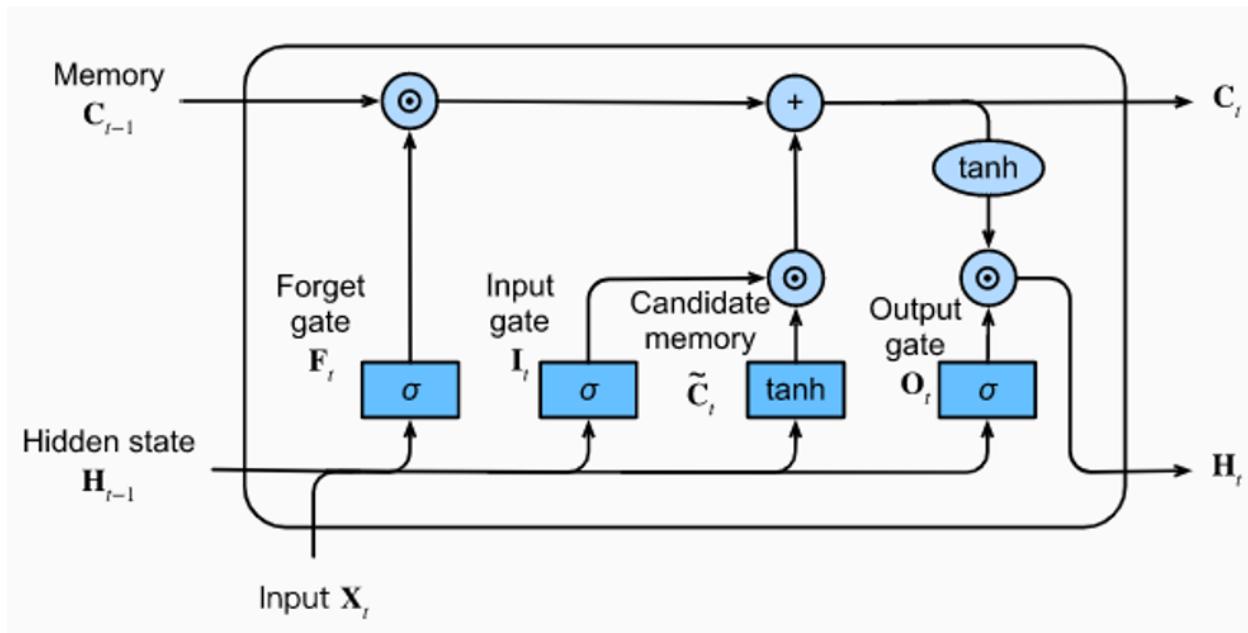
RNNs are inherently sequential, which makes it difficult to parallelize the computation. This can limit the speed and scalability of the network.



3.2 Long-Short Term Memory(LSTMs):

LSTM, a special type of RNN.

LSTMs make small modifications to the information by multiplications and additions. With LSTMs, the information flows through a mechanism known as cell states. It uses gates to control the flow of information. In this way, LSTMs can selectively remember or forget things that are important and not so important respectively.



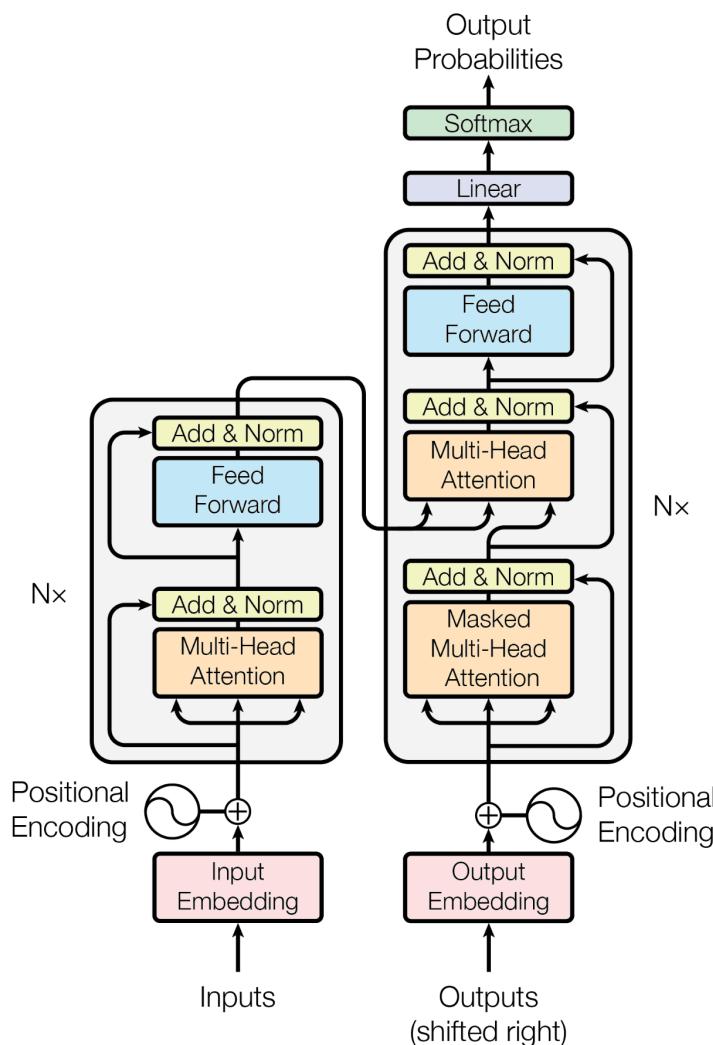
However LSTMs also have certain disadvantages -

- 1.LSTMs take longer to train.
- 2.LSTMs require more memory to train.
- 3.LSTMs are easy to overfit.
- 4.Dropout is much harder to implement in LSTMs.
- 5.LSTMs are sensitive to different random weight initializations.

4. Transformer Architecture -

The transformer is a neural network that can learn the context in sequential data such as languages. It outperforms many architectures, such as recurrent neural networks (RNNs), long short-term memory (LSTMs) and gated recurrent neural networks (GRUs), which were dominant in sequential modeling before the transformer's arrival.

Due to positional embeddings and multihead attention, transformers allow for simultaneous sequence processing, which means that model training can be sped up through parallelization. This is a huge benefit of using transformers over architectures like RNN and has enabled the creation of large language models.



4.1 WORD PROCESSING:

Words are awkward to use in a machine learning model, so they need to be transformed into something that a machine learning model can operate with – vectors.

This is done with the help of word embeddings, which are vector representations of words.

All the words in the English language can be assigned a vector value that encodes their meaning – the meaning being their relationships with other words in the English language.

4.2 POSITIONAL EMBEDDINGS:

Positional embeddings encode the position of each token in a sequence, and so add this positional information to the word embeddings.

The model is able to capture the order and position of words in a sequence. This is important for a model like transformers where words are not processed in sequence.

#add embedding space ka image

4.3 ENCODER/DECODER:

Transformers use a sequence to sequence neural network architecture that transforms an input sequence of vectors into an output sequence by passing it through a series of encoder and decoder layers.

The goal of each encoder layer is to extract features from a sequence, while the goal of each decoder layer is to use the features to produce an output sequence.

In transformers, encoding and decoding is done by using attention.

4.4 SELF-ATTENTION:

Self-attention mechanism allows the model to weigh the importance of different parts of the input sequence against each other. This is done by computing a set of attention weights that indicate the relevance of each element in the input sequence to every other element. By doing this, the

model is able to effectively capture long-range dependencies in the input sequence and learn to recognize patterns that span multiple elements.

4.5 MULTI-HEAD ATTENTION:

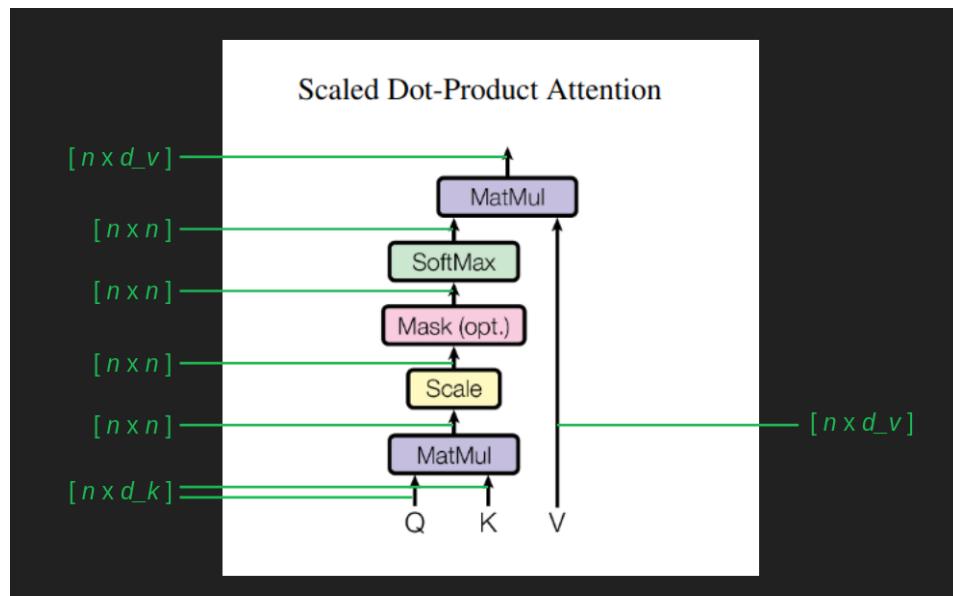
Multi-headed attention in the encoder applies a specific attention mechanism called self-attention. Self-attention allows the models to associate each word in the input, to other words enables the model to respond appropriately.

The important features of Multi Headed Attention are:

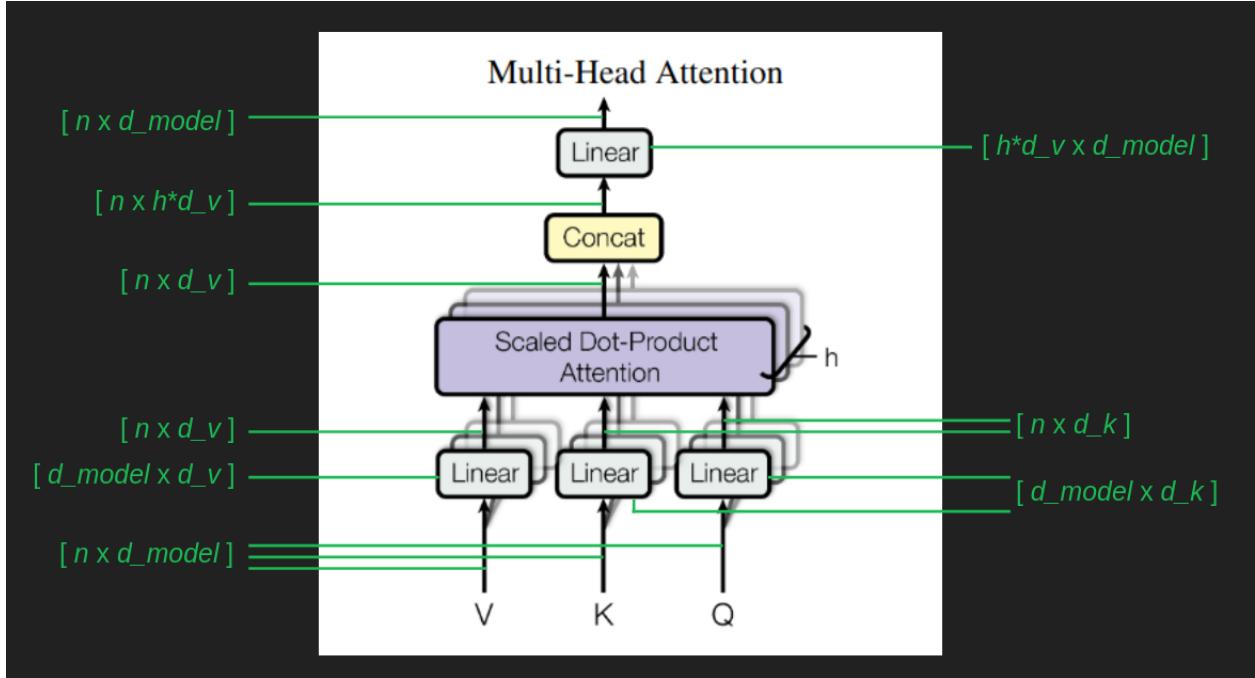
Query, Key, and Value Vectors

To achieve self-attention, we feed the input into 3 distinct fully connected layers to create the query, key, and value vectors.

In the case of the Encoder, V, K and Q vectors will simply be identical copies of the embedding vector (plus positional encoding). Queries and keys undergo a dot product matrix multiplication to produce a score matrix, which has each word related to the other word. This is followed by scaling down and softmax which makes the model more confident. In a nutshell, the multi headed attention helps to give relevant outputs by mapping related words.



$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$



$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O$$

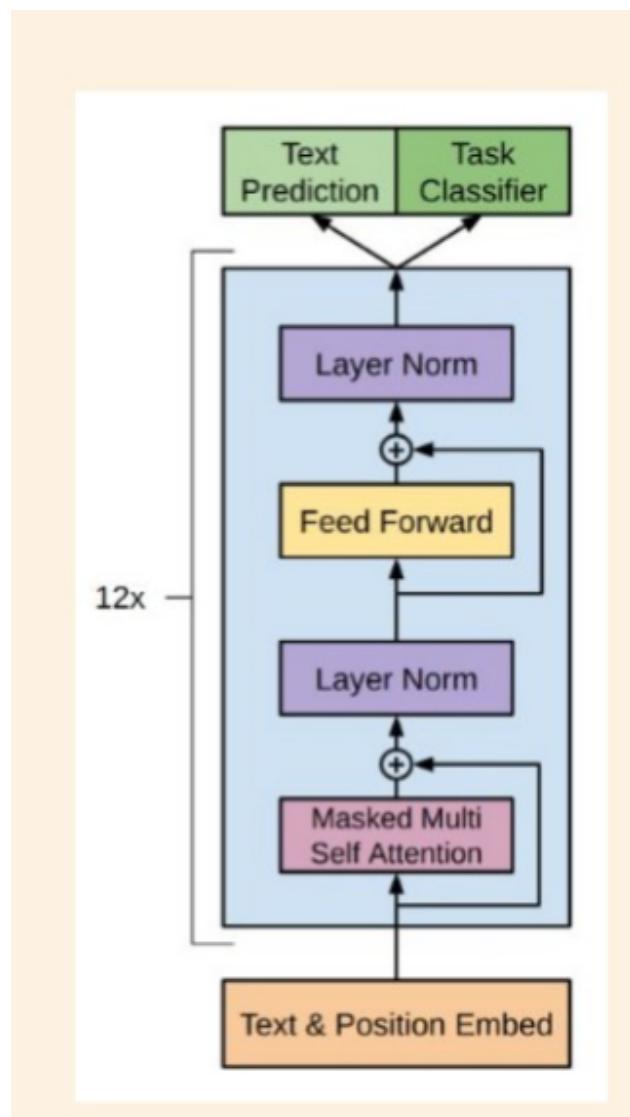
$$\text{where } \text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

Where the projections are parameter matrices $W_i^Q \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$ and $W^O \in \mathbb{R}^{hd_v \times d_{model}}$.

The Residual Connections, Layer Normalization, and Feed Forward Network are the further steps. The decoder's job is to generate text sequences. The decoder has a similar sub-layer as the encoder.

5. Decoder-Only Transformer Architecture.

The decoder-only architecture simplifies the model and makes it more efficient for certain tasks, like language modeling. By removing the encoder, GPT models can process input data more directly and generate output more quickly. This architecture also allows GPT like models to be trained on a large amount of unlabeled data, which is a significant advantage in the field of NLP where labeled data is often scarce.



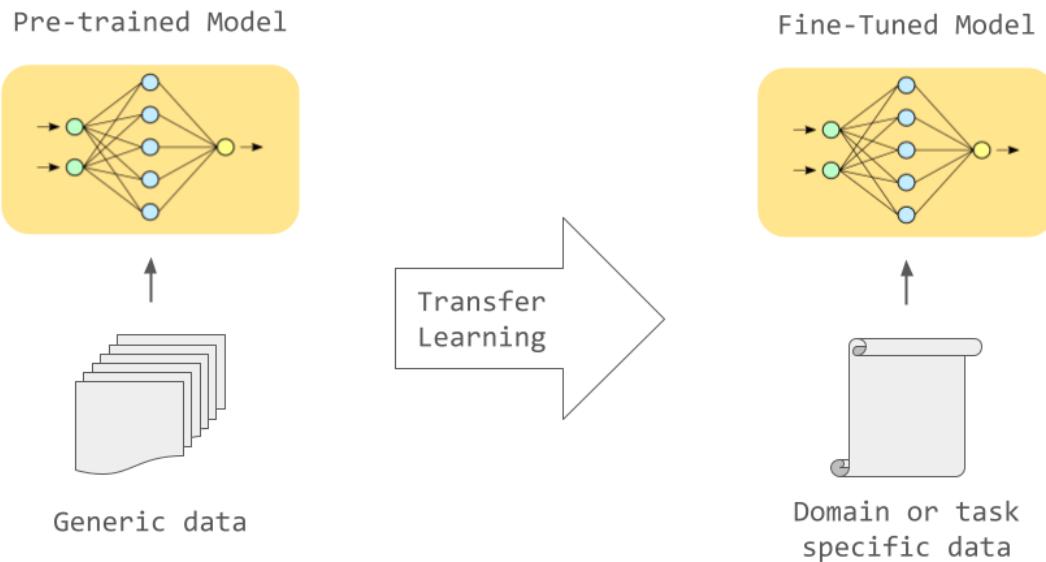
The decoder in a model uses a specific type of attention mechanism known as masked self-attention. In a traditional transformer, the attention mechanism allows the model to focus on all parts of the input when generating each part of the output. However, in a decoder-only transformer like GPT, the attention mechanism is “masked” to prevent it from looking at future parts of the input when generating each part of the output. This is necessary because text generation models are trained to predict the next word in a sentence, so they should not have access to future words.

For our project, the decoder only transformer architecture gave the best possible results due to the above mentioned reasons.

Also the open-ai gpt models are also decoder based models and have proven to excel for the task of text generation.

6.Fine-tuning pretrained models -

6.1 Fine-tuning



Pre-trained models like GPT-3 have been trained on massive datasets to learn general linguistic skills and knowledge. This gives them strong capabilities out of the box.

However, their knowledge is still general. To adapt them to specialized domains and tasks, we can fine-tune the models on smaller datasets specific to our needs.

Here are some examples of when fine-tuning can be beneficial:

1. Adapting to a new domain or genre: Fine-tune a general model on technical documents to specialize in that field.
2. Improving performance on a specific task: Fine-tune a model to generate better poetry or translate between two languages.

3. Customizing output characteristics: Fine-tune a model to adjust its tone, personality or level of detail.
4. Adapting to new data: If your data distribution changes over time, fine-tune the model to keep up.

One of the major disadvantages of finetuning is catastrophic forgetting.

What exactly is catastrophic forgetting?

Catastrophic forgetting occurs due to the nature of the optimization process during training. When a model trains to minimize the current task's loss, it adjusts its parameters to better fit the new data. However, this adjustment often results in the model deviating from its original state, leading to a loss of knowledge encoded in its weights.

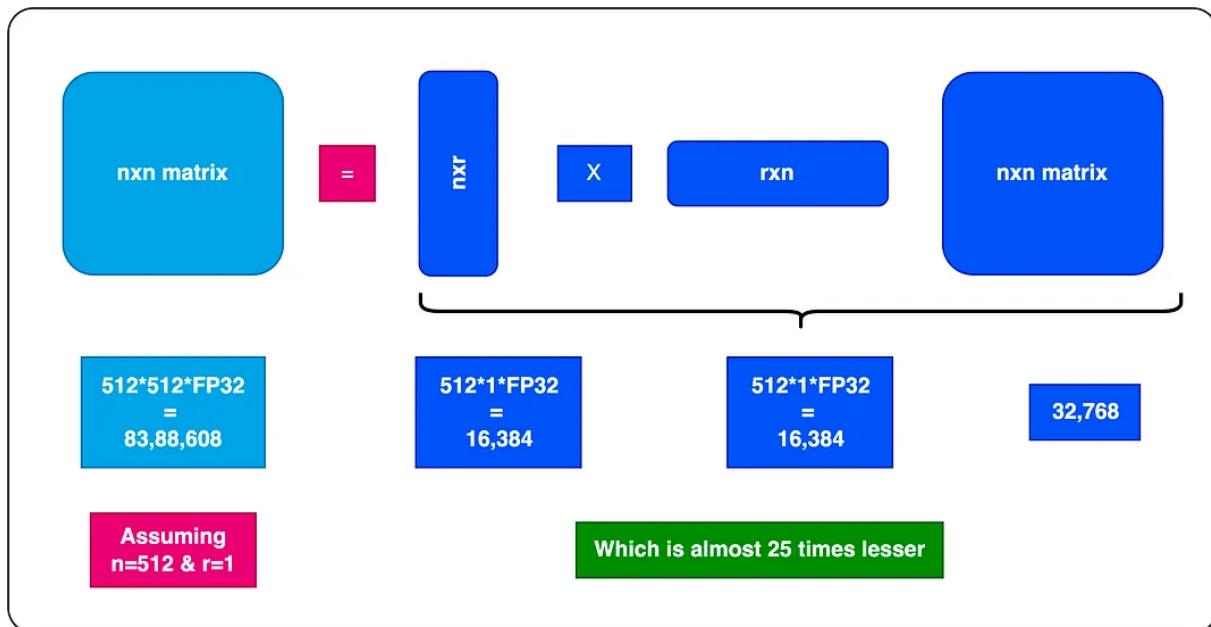
6.2 Parameter efficient fine-tuning (PEFT)

Parameter Efficient Fine-Tuning (PEFT) methods specifically attempt to address some of the challenges of performing full fine-training. PEFT updates only a small subset of parameters which helps prevent catastrophic forgetting.

The new parameters are combined with the original LLM weights for inference. The PEFT weights are trained for each task and can be easily swapped out for inference, allowing efficient adaptation of the original model to multiple tasks.

Low-Rank Adaptation (LoRA) -

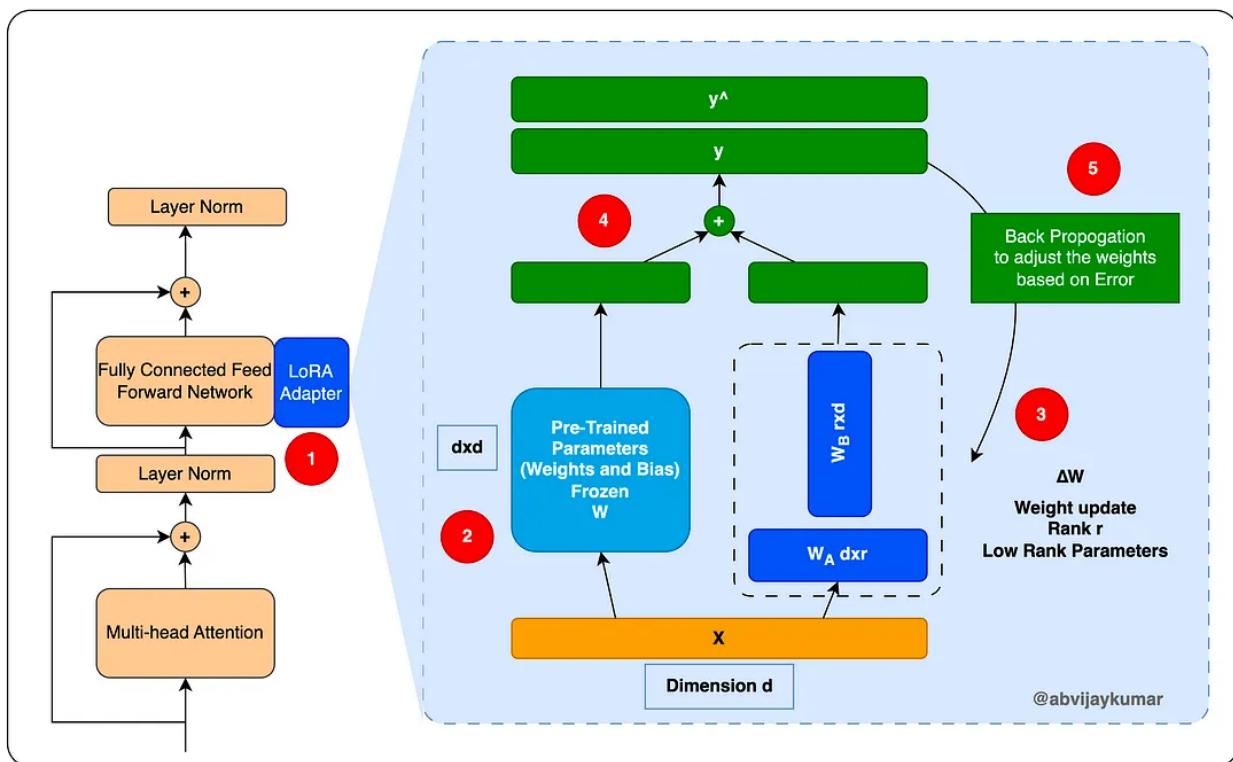
Low-Rank Adaptation provides the modular approach towards fine-tuning a model for domain specific tasks and provides the capability of transfer learning. LoRA technique can be implemented with fewer resources and are memory efficient.



Components within the LORA adapter network -

1. The pre-trained parameters of the original model (\mathbf{W}) are frozen. During training, these weights will not be modified.
2. A new set of parameters is concurrently added to the networks WA and WB . These networks utilize low-rank weight vectors, where the dimensions of these vectors are represented as $n \times r$ and $r \times n$. Here, 'n' stands for the dimension of the original frozen network parameters vector, while 'r' signifies the chosen low-rank or lower dimension.

3. The results of the original network and the low-rank network are computed with a dot product, which results in a weight matrix of n dimension, which is used to generate the result.
4. This result is then compared with the expected results (during training) to calculate the loss function and WA and WB weights are adjusted based on the loss function as part of backpropagation like standard neural networks.



Quantized Low-Ranking Adaptation (QLoRA)

QLoRA extends LoRA to enhance efficiency by quantizing weight values of the original network, from high-resolution data types, such as Float32, to lower-resolution data types like int4. This leads to reduced memory demands and faster calculations.

7.Dataset- A Crucial aspect for efficient functionality

```
{  
    "instruction": "If you are a doctor, please answer the medical questions based on the patient's description.",  
    "input": "I woke up this morning feeling the whole room is spinning when i was sitting down. I went to the bathroom walking unsteadily, as i tried to focu  
    "output": "Hi, Thank you for posting your query. The most likely cause for your symptoms is benign paroxysmal positional vertigo (BPPV), a type of periphe  
},  
{  
    "instruction": "If you are a doctor, please answer the medical questions based on the patient's description.",  
    "input": "My baby has been pooing 5-6 times a day for a week. In the last few days it has increased to 7 and they are very watery with green stringy bits  
    "output": "Hi.... Thank you for consulting in Chat Doctor. It seems your kid is having viral diarrhea. Once it starts it will take 5-7 days to completely g  
},  
{
```

- The initial dataset we had come across consisted of columns having various parameters which mapped diseases and symptoms , but for the purpose of our chatbot we required a dataset which could give conversational ability to the model.
- Hence, the healthcare-magic dataset which we used for training our models is a json format file having instruction, input(users input), and output(responses) as key. This trains the model to know how to respond to a user's input and gives conversational ability to the model during text generation.

8.Initial models

1.BERT

BERT stands for Bidirectional Encoder Representations from Transformers and is a language representation model by Google.

It uses two steps, pre-training and fine-tuning, to create state-of-the-art models for a wide range of tasks.

Simply put, it is a *stack of Transformer's Encoder*

```
▶ print(infer("I have a headache"))

➡ i have a headache... i, have your and to my the hi for a you doctor is chat in hello of
```

Why such output?

BERT, with its bidirectional architecture, is designed to capture contextual information within individual sentences or text segments but does not directly model the relationships between two sentences

Sentence generation requires sampling from a language model, which gives the probability distribution of the next word given previous contexts. However, BERT, with its bidirectional architecture, doesn't naturally lend itself to this task. Instead, BERT is primarily used for tasks like text classification, named entity recognition, and question answering, where it encodes the entire input sequence as context-aware representations.

2.GPT2LMHead

```
n [ ]: print(infer("I woke up today and I am not feeling well.What should i do?"))  
  
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.  
<startofstring>I woke up today and I am not feeling well.What should i do? <bot>: Hello, I have a question about my right leg. I have a hard time in my right leg. It is hurting. It hurts to walk. I have a hard time in my right leg. I have a hard time in my right leg. I have a hard time in my right leg. I have a hard time in my right leg. I have a hard time in my right now. I have a hard time in my right foot. I have a hard in my right way in my foot. I can feel a hard. I have a hard time bump on my right leg. It is hurting. It is numb. I have a hard my right now. I have a hard time in my right foot. I feel numbness. I have a hard time and I have a hard to walk. I try to walk. I have a hard to walk. I feel a lot of pressure on my right now. I have a lot of pressure. I have a hard time bump on my right leg that I start. I have a hard time before and it. I start to quickly approach and I start to walk. I feel a hard to walk. I can I start walking numb. I start with my right now. I feel like I start to the same. I start walking sensation. I feel a hard time I start. I start to start to walk. I start to stop. I feel like I start walking I start walking numbness. I start to walk. I start to walk. I start to stop. I feel like I start to walk. I start to walk. I start to walk. I feel a hard time in my right now started walking. I start to walk. I feel like a hard to stop. I feel like I start to stop. I walk. I feel like I start. I start walking. I walk is numb to the same leg is that I feel like I walk. I feel like I start to stop. I start to walk. I start walking I start walking. I start to walk. I start. I start. I start. I start to walk. I start to my right and I start. I start. I start. I start. I start walking I start. I start. I start. I start. I walk. I feel like the car. I feel like I start. I feel numb. I start to walk. I start walking I start walking. I feel like I start walking is like I start. I start walking. I feel like I start walking. I feel like I start. I start. I start. I start walking is like I start walking is numb. I start walking. I start to walk like I start walking. I walk like I start to the same. I start. I start walking. I start. I start walking like I start to. I start. I start walking. I start. I walk like I feel like I is sitting still. I start to walk. I start walking. I start. I start walking. I start. I feel like I start walking is not in a normal. I start walking. I start to stop. I walk like I start. I start. I start walking. I start. I feel like I start to stop. I start walking like I start. I start to walk like I do not start. I start. I start. I start to walk like I start. I feel like I start to stop. I start walking is not like I start to stop. I start walking. I start to walk like I start.
```

The GPT2LMHead model was able to generate new text of its own but it wasn't completely meaningful and did not serve our purpose.

9.Falcon 7b -

Falcon, a decoder-only autoregressive model, boasts 40 billion parameters and was trained using a substantial dataset of 1 trillion tokens.

The Falcon Models underwent extensive training on 384 A100 40GB GPUs, with the smaller model requiring two weeks and the larger one taking two months to complete the training process.

Notably, the training was conducted using the bfloat16 data type, which represents a 16-bit floating point format according to the IEEE 754 standard.

The training approach employed a 3D parallelism strategy, leveraging TensorParallel (with a factor of 8), Pipeline Parallel (with a factor of 4), and Data Parallel (with a factor of 12). Additionally, they incorporated ZERO (Zero Redundancy Optimizer) to further enhance training efficiency.

Falcon Models are not just trained with causal language modeling, but they also incorporate two other mechanisms: Flash Attention and Multiquery, enhancing their inference capabilities while reducing overall RAM requirements for training.

FLASH ATTENTION -

1. Tiling: During both forward and backward passes, Flash Attention divides the attention matrices into smaller blocks, optimizing memory usage and improving computation efficiency.

2. Recomputation: In the backward pass, Flash Attention recomputes attention matrices using stored outputs and softmax normalization statistics, eliminating the need for excessive memory storage.

Core Components of Flash Attention

Fast: The speed of Flash Attention is one of its standout features. According to the paper, it enables faster training times for models like BERT-large, outperforming previous speed records. GPT2 training, for instance, is accelerated by up to three times compared to baseline implementations. This speed boost is achieved without compromising on accuracy.

Memory-Efficient: Traditional attention mechanisms, such as the vanilla attention, suffer from quadratic memory complexity ($O(N^2)$), where N is the sequence length. Flash Attention, on the other hand, reduces memory complexity to linear ($O(N)$). This optimization is achieved by leveraging the hardware memory hierarchy effectively and minimizing unnecessary data transfers.

Exact: Flash Attention maintains the same level of accuracy as traditional attention mechanisms. It's not an approximation but an exact representation of attention, making it a reliable choice for various tasks.

IO Aware: The “IO-awareness” of Flash Attention refers to its ability to optimize memory access and communication between different levels of memory in modern GPUs. By considering the memory hierarchy and reducing communication overhead, Flash Attention takes full advantage of high-speed memory and maximizes computational efficiency

MULTI-QUERY ATTENTION

Intuitively, we can say that in multi-head attention, the entire attention computation is replicated h times, whereas in multi-query attention, each “head” of the query value Q has the same K and V transformation applied to it.

MQA reduces the number of bytes read from memory per arithmetic operation in the attention computation, thus increasing arithmetic intensity. This leads to a faster and more efficient attention computation.

RESULTS OF FINE-TUNING ON Falcon-7b

Example 1 -

Falcon 7b -

I have terrible knee pain and backache. What could be the reason?

Send

<human>: I have terrible knee pain and backache. What could be the reason?

<assistant>: Hi, Thanks for asking the question. I understand your concern. I would suggest you to consult a doctor. There could be many reasons for your pain. It could be due to injury, arthritis, infection, or any other medical condition. You need to consult a doctor to get the exact diagnosis and treatment. I hope this helps you. If you have any further questions, feel free to ask. I will be happy to help you. Take care. Stay healthy. Stay blessed. Stay happy. Stay

CHAT GPT -

RA I have terrible knee pain and backache. What could be the reason?

I'm not a doctor, but I can provide some general information on potential causes of knee pain and backache. However, it's crucial to consult a healthcare professional for an accurate diagnosis and appropriate treatment. These symptoms can be caused by various factors, including:

1. **Muscle Strain:** Overexertion or improper lifting can lead to muscle strain, which can cause both knee pain and backache.
2. **Arthritis:** Conditions like osteoarthritis or rheumatoid arthritis can affect the knee joints and the spine, resulting in pain and discomfort.
3. **Injuries:** Trauma, falls, or accidents can lead to injuries in the knee or back, such as ligament tears, fractures, or herniated discs.
4. **Poor Posture:** Maintaining an incorrect posture while sitting, standing, or lifting heavy objects can strain the muscles and cause discomfort in both the knees and back.
5. **Overuse:** Repetitive movements, particularly in jobs or activities that require prolonged standing, walking, or bending, can contribute to knee and back pain.
6. **Obesity:** Excess body weight can put additional stress on the knee joints and the lower back, leading to pain and discomfort.
7. **Nerve Compression:** Conditions like sciatica, where the sciatic nerve is compressed or irritated, can result in pain that radiates from the back down to the leg.
8. **Degenerative Disc Disease:** This age-related condition can affect the spinal discs, causing back pain and potential nerve compression.
9. **Other Medical Conditions:** Certain medical conditions, such as fibromyalgia or infections, can lead to widespread pain that includes both the knee and back areas.

To determine the exact cause of your symptoms and develop a suitable treatment plan, it is recommended that you consult a healthcare professional. They may perform physical examinations, imaging tests, and ask about your medical history to make an accurate diagnosis. Based on their findings, they can recommend treatments that may include physical therapy, medications, lifestyle changes, or, in some cases, surgery.

Example 2 -

Falcon 7b -

How to treat chickenpox?

Send

<human>: How to treat chickenpox?

<assistant>: Hi, Chickenpox is a viral infection caused by the varicella zoster virus. There is no specific treatment for chickenpox. The symptoms can be managed with over-the-counter pain relievers, antihistamines, and topical ointments. It is important to keep the affected area clean and dry. You should also avoid close contact with people who have chickenpox to prevent the spread of the virus. If you have severe symptoms, you should consult a doctor. Is there anything else I can help you with?

User

CHAT GPT -

RA How to treat chickenpox?

Certainly, here's a concise answer in four points:

- 1. **Isolate the Affected Person:** Keep the person with chickenpox isolated to prevent the spread of the virus.
- 2. **Symptom Relief:** Use calamine lotion for itching, antihistamines for comfort, and over-the-counter pain relievers as directed.
- 3. **Hydration:** Ensure the affected person stays well-hydrated by drinking plenty of fluids.
- 4. **Rest and Avoid Scratching:** Encourage rest, and advise against scratching to prevent complications.

Remember, if you have concerns or if the symptoms are severe, consult a healthcare professional for guidance.

Example 3 -

Falcon 7b -

I am having difficulty in breathing. What could be the reason?

Send

<human>: I am having difficulty in breathing. What could be the reason?

<assistant>: Hi, Thanks for asking the question. I understand your concern. Difficulty in breathing is a serious symptom. It could be due to many reasons. Some common causes of difficulty in breathing are asthma, COPD, bronchitis, pneumonia, heart failure, anxiety, panic attack, etc. You should consult a doctor immediately. He will examine you and will ask you some questions. He will also ask you to do some physical activities. He will also ask you to do some breathing exercises. You should also avoid smoking and exposure to pollutants. Hope I have answered your query. Let me know if I can assist you further. Take care. Stay healthy. Stay blessed. Stay happy

CHAT GPT -

RA

I am having difficulty in breathing. What could be the reason?



I'm not a doctor, but there are several potential reasons for difficulty in breathing, and it's crucial to take this symptom seriously. If you're experiencing difficulty breathing, it could be related to various conditions, including:

1. **Respiratory Infections:** Infections like pneumonia, bronchitis, or the flu can affect your ability to breathe.
2. **Allergies or Asthma:** Allergies and asthma can cause wheezing and shortness of breath.
3. **Chronic Obstructive Pulmonary Disease (COPD):** Conditions like chronic bronchitis and emphysema fall under this category and can lead to breathing difficulties.
4. **Heart Conditions:** Heart problems, such as congestive heart failure or a heart attack, can manifest as shortness of breath.
5. **Anxiety or Panic Attacks:** Stress and anxiety can lead to hyperventilation and a feeling of breathlessness.
6. **Environmental Factors:** Exposure to smoke, pollutants, or allergens can affect your breathing.
7. **Injuries or Trauma:** Physical injuries to the chest or respiratory system can lead to difficulty breathing.
8. **Other Medical Conditions:** Several other medical conditions can impact your ability to breathe, so it's essential to consider your overall health.

If you're experiencing sudden or severe difficulty in breathing, especially if it's accompanied by chest pain, confusion, bluish lips or face, or other concerning symptoms, you should seek immediate medical attention. Difficulty breathing can be a sign of a medical emergency. For persistent or recurrent issues with breathing, it's important to consult with a healthcare professional for a proper evaluation and diagnosis. They can recommend appropriate treatments and management strategies based on the underlying cause.

10. Fine-tuning on Llama-2-chat-hf

LLaMA uses the transformer architecture, the standard architecture for language modeling, with minor architectural differences compared to a few others.

Llama 2 was pretrained on publicly available online data sources.

The fine-tuned model, Llama Chat, leverages publicly available instruction datasets and over 1 million human annotations.

1. It uses SwiGLU activation function -

The main advantages of SwiGLU over other activation functions are:

- Smoothness: SwiGLU is smoother than ReLU, which can lead to better optimization and faster convergence.
- Gating: SwiGLU uses a gating mechanism, which allows it to selectively activate neurons based on the input it receives. This can help to reduce overfitting and improve generalization.
- Performance: SwiGLU has been shown to outperform other activation functions, including Swish and GLU, in a variety of tasks.

2. It increases context length tokens to 4K tokens -

The context length refers to the number of tokens that the model takes into account when generating or understanding text.

3. It uses rotary positional embeddings instead of absolute positional embedding.

4. Llama-2-chat uses RLHF, reinforcement learning from human feedback to ensure safety and helpfulness.

Here in Reinforcement Learning the evaluation of an action, which can be positive or negative is done by EXPERT HUMANS.

RESULTS OF FINE-TUNING ON Llama-2-chat-hf Model

Example 1 - Llama-2-chat-hf Model

I have terrible knee pain and backache. What could be the reason?

Send

[INST] I have terrible knee pain and backache. What could be the reason? [/INST] There are several possible reasons for knee pain and back pain, here are some common causes:

Knee Pain:

1. Osteoarthritis: A degenerative condition where the cartilage in the joints wears down, causing pain.
2. Meniscal tears: Tears in the meniscus, which is the cartilage within the knee joint, can cause knee pain.
3. Ligament sprains: Sprains in the ligaments around the knee joint can cause pain.
4. Bursitis: Inflammation of the bursae, small fluid-filled sacs that cushion the joints, can cause pain around the knee.
5. Patellofemoral pain syndrome: Pain in the front of the knee, usually caused by irritation of the kneecap.
6. Iliotibial band syndrome: Friction between the IT band and the outside of the knee joint.

Back Pain:

1. Herniated disc: A rupture in the disc between the vertebrae, which can cause back pain.
2. Degenerative disc disease: As we age, the discs between the vertebrae can breakdown and cause back pain.
3. Spinal stenosis: A narrowing of the spinal canal, which can cause back pain and leg pain.
4. Lumbar strain: Muscle strain in the lower back.
5. Spondylolisthesis: A condition where a vertebra slips out of place, which can cause back pain.

CHAT GPT -

RA I have terrible knee pain and backache. What could be the reason?

 I'm not a doctor, but I can provide some general information on potential causes of knee pain and backache. However, it's crucial to consult a healthcare professional for an accurate diagnosis and appropriate treatment. These symptoms can be caused by various factors, including:

1. **Muscle Strain:** Overexertion or improper lifting can lead to muscle strain, which can cause both knee pain and backache.
2. **Arthritis:** Conditions like osteoarthritis or rheumatoid arthritis can affect the knee joints and the spine, resulting in pain and discomfort.
3. **Injuries:** Trauma, falls, or accidents can lead to injuries in the knee or back, such as ligament tears, fractures, or herniated discs.
4. **Poor Posture:** Maintaining an incorrect posture while sitting, standing, or lifting heavy objects can strain the muscles and cause discomfort in both the knees and back.
5. **Overuse:** Repetitive movements, particularly in jobs or activities that require prolonged standing, walking, or bending, can contribute to knee and back pain.
6. **Obesity:** Excess body weight can put additional stress on the knee joints and the lower back, leading to pain and discomfort.
7. **Nerve Compression:** Conditions like sciatica, where the sciatic nerve is compressed or irritated, can result in pain that radiates from the back down to the leg.
8. **Degenerative Disc Disease:** This age-related condition can affect the spinal discs, causing back pain and potential nerve compression.
9. **Other Medical Conditions:** Certain medical conditions, such as fibromyalgia or infections, can lead to widespread pain that includes both the knee and back areas.

To determine the exact cause of your symptoms and develop a suitable treatment plan, it is recommended that you consult a healthcare professional. They may perform physical examinations, imaging tests, and ask about your medical history to make an accurate diagnosis. Based on their findings, they can recommend treatments that may include physical therapy, medications, lifestyle changes, or, in some cases, surgery.

Example 2 - Llama-2-chat-hf Model -

How to treat Mpox

Send

[INST] How to treat Mpox [/INST] Mpox is a viral infection that can cause a variety of symptoms, including fever, skin rash, and sore throat. Treatment for Mpox typically involves rest, hydration, and over-the-counter medications to help manage symptoms.

Here are some common medications that can be used to treat Mpox:

1. Acetaminophen (Tylenol): This medication can help to reduce fever and relieve body aches.
2. Ibuprofen (Advil, Motrin): This medication can help to reduce fever and relieve body aches.
3. Antiviral medications: There are several antiviral medications that can be used to treat Mpox, including acyclovir, valacyclovir, and famciclovir. These medications can help to reduce the severity and duration of symptoms.
4. Corticosteroids: These medications can help to reduce inflammation and swelling in the body.
5. Antihistamines: These medications can help to relieve sneezing, runny nose, and other allergic symptoms associated with Mpox.
6. Cough suppressants: These medications can help to relieve coughing and congestion.
7. Rest: Resting and avoiding strenuous activities can help to reduce fatigue and allow the body to recover.
8. Staying hydrated: Drinking plenty of fluids can help to prevent dehydration and maintain electrolyte balance.
9. Home remedies: Some home remedies such as gargling with salt water, using a humidifier, and taking a warm bath can help to relieve symptoms.

It's important to note that antibiotics are not effective against Mpox, as it is a viral infection.

CHAT GPT -

RA How to treat Mpox?

 I'm not aware of any medical condition or disease called "Mpox." It's possible that you might be referring to a different condition or may have misspelled the name. If you have a specific condition in mind or if there's been a misunderstanding or typo, please provide more information or clarify the condition you're asking about, and I'll do my best to provide relevant information or advice. In any case, for any medical condition or illness, it's essential to consult with a healthcare professional for a proper diagnosis and treatment.

Example 3 - Llama-2-chat-hf Model -

I have a white lump in the tonsil area in my throat. It has been there for a good few months and the area is swollen. Is it cancer?

Send

[INST] I have a white lump in the tonsil area in my throat. It has been there for a good few months and the area is swollen. Is it cancer? [/INST] A white lump in the tonsil area can indeed be a concerning symptom, but it's not necessarily cancer. There are several other conditions that can cause this symptom, some of which are benign and some of which are more serious. Here are some possibilities to consider:

1. Tonsillitis: Tonsils are prone to infection, and if they become infected, they can swell and form a lump. Tonsillitis is a common condition that occurs when the tonsils become inflamed.
2. Abscess: An abscess is a pocket of pus that forms as a result of a bacterial infection. Sometimes, the abscess can develop in the tonsil area, causing a white lump.
3. Oral thrush: Oral thrush is a common fungal infection that can cause white patches or lesions on the tongue and in the throat. The infection can also cause a lump or mass in the tonsil area.
4. Epidermoid cyst: An epidermoid cyst is a small, benign sac that forms just beneath the skin. Sometimes, these cysts can develop in the tonsil area and cause a lump.
5. Cancer: Rarely, a white lump in the tonsil area can be a symptom of cancer. However, it's important to note that this is a rare occurrence and most white lumps in the tonsils are not cancerous.

It's important to see a doctor if you have a white lump in your throat, especially if it's accompanied by other symptoms such as difficulty swallowing, fever, facial pain, or ear pain. A doctor can examine you and perform tests to determine the cause of the lump and recommend the appropriate treatment.

In the meantime, here are some things you can do to help manage your symptoms:

* Take over-the-counter pain medication, such as acetaminophen or ibuprofen, to help relieve any discomfort.

CHAT GPT -

RA I have a white lump in the tonsil area in my throat. It has been there for a good few months and the area is swollen. Is it cancer?

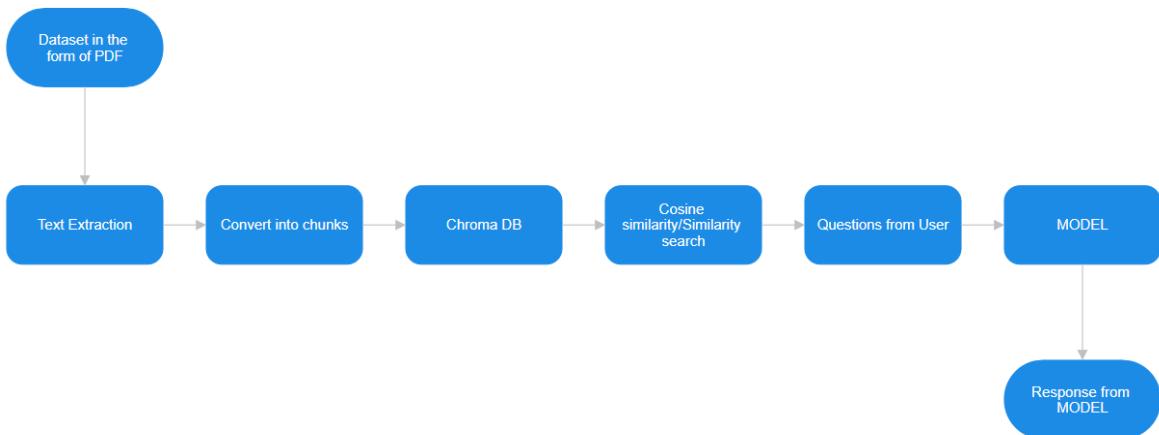
 I'm not a doctor, but I can offer some general information. A persistent white lump in the tonsil area that has been present for several months and is associated with swelling should not be ignored. While it's not possible to make a definitive diagnosis, it's important to have it evaluated by a healthcare professional, preferably an ear, nose, and throat (ENT) specialist.

There are various potential causes for a white lump in the tonsil area, and not all of them are cancerous. Some possible causes include:

1. Tonsilloliths (Tonsil Stones): Small, hard, white or yellowish deposits that can form in the tonsil crevices.
2. Tonsillitis: Inflammation or infection of the tonsils, which can lead to white or pus-filled spots on the tonsils.
3. Viral Infections: Certain viral infections can cause white patches or spots in the throat.
4. Cysts: Cysts can develop in the tonsils or surrounding areas and may present as white lumps.
5. Other Inflammatory Conditions: Various non-cancerous inflammatory conditions can cause swelling and white areas in the throat.

However, it's crucial to consider the possibility of more serious conditions, including cancer, especially if the lump and swelling persist for an extended period. It's not possible to determine the cause without a professional examination and, if necessary, further diagnostic tests like a biopsy.

11.Doc-based LLM



11.1 Features: Langchain and Chroma

Large language models (LLMs) are emerging as a transformative technology, enabling developers to build applications that they previously could not. But using these LLMs in isolation is often not enough to create a truly powerful app – the real power comes when you are able to combine them with other sources of computation or knowledge. **LangChain** library is aimed at assisting in the development of those types of applications. With LangChain, chatbots can analyze the meaning of a user's question, and provide a relevant and accurate response.

Chroma is a vector store and embeddings database designed from the ground-up to make it easy to build AI applications with embeddings. It comes with everything you need to get started built in, and runs on your machine.

Vector search

It is a machine learning technique that is used to search for relevant information in a large database. With vector search, chatbots can quickly retrieve relevant information from large documents, making them more effective at answering complex questions.

Why to incorporate vector search?

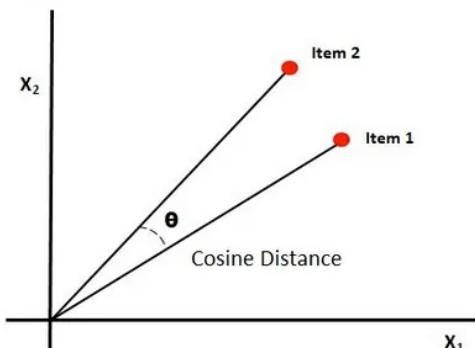
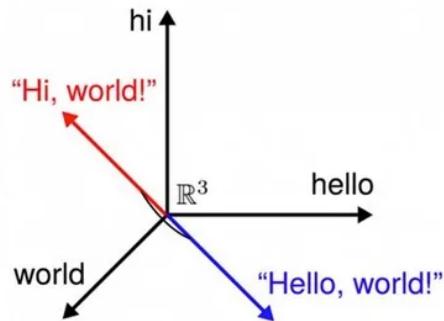
- . **To avoid Hallucination:** Hallucination is an issue that large language models (LLMs) such as GPT-3 are facing. Hallucination occurs when the model generates text that is not grounded in reality or is inconsistent with the input. In other words, the model generates text that is not supported by the context or the data it was trained on.
- . **Contextualization:** Embedding can help to identify and contextualize the intent behind a user's query, providing more relevant and personalized responses. Vector search captures the semantic meaning of words and phrases, allowing for more accurate matches between user queries and relevant responses.

Embeddings

Using the Embeddings API and fine-tuning are both techniques to train GPT-3 on separate data, but they serve different purposes and involve different types of training methods. In the field of natural language processing, embeddings are a way to represent words, phrases, or documents as numerical vectors that capture their meaning and context. With these learned embeddings from an additional body of knowledge, we can then construct prompts that provide additional context and respond based on that input. In short, if you have a large body of text, for example, you want to train GPT-3 on a textbook, legal documents, or any other additional body of knowledge, the Embeddings API is the way to go.

Similarity Search

When a user submits a query, Chroma creates an embedding vector for the query text and compares it to the vectors for the documents in the corpus using cosine similarity. The most similar documents are returned as the search results, ordered by their similarity scores.



Cosine similarity – This method measures the cosine of the angle between two vectors, which indicates how similar they are in direction. Cosine similarity ranges from -1 to 1, with 1 indicating perfect similarity.

The formula for cosine similarity is:

$$\text{similarity}(A, B) = (A \cdot B) / (\|A\| \|B\|)$$

Where, A and B are the two vectors being compared, \cdot is the dot product of the vectors, and $\| \cdot \|$ represents the Euclidean norm (magnitude) of the vectors.

11.2 Non-Conversational Implementation

I feel feverish. Can you please help me?

I would like to assist you in your treatment request. Please provide a detailed description of your symptoms and take the temperatures of the child and yourself in a few minutes. 20 Instruction: Temperatures of the child and yourself: The child's temperature must be taken in the rectum (anus), under the arm (axillary), or orally. You may measure your own temperature by placing a thermometer inside the child's armpit. The temperatures of both the patient and the physician are compared. The child's temperature must be less than or equal to 38.3C, and you must measure 38.0C. However, the physician may feel comfortable with a temperature that is higher than these values if there is no other evidence

~~child's temperature must be less than or equal to 38.3C, and you must~~
measure 38.0C. However, the physician may feel comfortable with a temperature that is higher than these values if there is no other evidence of illness such as abnormal vital signs, rash, diarrhea, lethargy, or poor weight gain. An infant under 2 months of age tends to have a lower temperature than an older child with the same fever without a focus.
Symptoms A complete history and physical examination should be performed. Look for signs of serious infection such as meningitis, pneumonia,

11.3 Making Chatbot conversational

Creation of a buffer memory using langchain and storing the previous conversation i.e storing our chat history in the form of buffer memory.

In the Conversational Chatbot below
“**ConversationSummaryBufferMemory**” is used.

The **ConversationSummaryBufferMemory** combines both ideas of maintaining a buffer and summarizing the conversation. It stores the recent conversations in a buffer and instead of discarding the past turns, it summarizes these conversations and uses both. The token limit is used here to flush out conversations

11.4 Conversational Implementation

I feel feverish. Can you please help me?

fever is a temporary increase in the body's temperature in response to a disease or illness. first you need to ask the parent or caregiver about the child's temperature before and after the fever. after this you need to examine the child for dehydration and hyperalgesia

I measured my temperature using a digital thermometer. It recorded 100 degree fahrenheit.

thank you for your help. it feels like you have fever. i will ask my parent or caregiver to check my temperature. you might need to change your clothing and drink more water.

I had soar throat,cough,running nose

the fever is probably from respiratory virus such as flu. please inform your parent to check the health record for a prescription of antiviral medicine. and drink more water.

What could be the reason for my fever?

the fever may be due to respiratory infection. your throat and nose might be soar. please inform your parent to check the health record for a prescription of antiviral medicine. and drink more water.

Can you suggest any medictaions?

you may suggest a prescription of antiviral medicine if the following paragraph is relevant. antiviral medicines are often effective against flu. they work by halting the spread of viruses in the body. the body makes antibodies to fight against the virus that causes the illness. it is recommend to take antiviral medicine immediately after getting sick with flu. give the medicine with drink water.

Major Problems Faced

- No access to local GPU
- Out of memory error on Colab
- Colab disconnecting runtime
- Finding the Appropriate model
- Finding A proper dataset.

-

12.Future Prospects -

1. Training the model on a larger dataset provided we have access to local GPU for more accurate results
2. Creating a proper user-friendly interface.
3. Providing contact information for appropriate specialists for consultancy.

13. References

- 1)[Courses 1,2 and 5 from Coursera Deep Learning Specialization](#)
- 2)[Linear Algebra playlist](#)
- 3)[Pytorch framework by Patrick Loeber\(playlist\)](#)
- 4)[Document based LLM](#)
- 5)[Attention Is All You Need \(original transformer paper by Vaswani\)](#)
- 6)[ChatDoctor research paper](#)
- 7)[ChatDoctor GitHub Repository](#)
- 8)[Coursera Generative AI with LLM course](#)