MediQ-Ai: Medical Question Answering(QA) systems using Fine-tuned LLMs

Akshat Khare
SITE - Faculty of Engineering
University of Ottawa
Ottawa, Canada
akhar075@uottawa.ca

Yash Keswani
SITE - Faculty of Engineering
University of Ottawa
Ottawa, Canada
ykesw031@uottawa.ca

Siddhant Tiwari
SITE - Faculty of Engineering
University of Ottawa
Ottawa, Canada
stiwa017@uottawa.ca

***Abstract*—With the advent of advancement and development of Question Answering (QA) systems fitted and utilized in specific domains based on NLP techiques. These systems have proved to be of immense potential in the medical and healthcare field to assist some healthcare professionals such as doctors, nurses in securing relevant insights into the patient's disease efficiently. This paper introduces MediQ-Ai, a QA model designed specifically for the medical domain, leveraging the comprehensive MedQuad dataset. MediQ-Ai is built upon some of the latest and state-of-the-art NLP architectures which was further trained on the MedQuad dataset, which basically comprises a diverse range of medical questions and their corresponding answers sourced from reliable medical literature and expert annotations. This model employs an amalgamation of pre-trained language representations and some of the domain-specific fine-tuning techniques to achieve robust performance in medical QA tasks. Some of the MediQuadA's Key features include its ability to comprehend and decode complex medical terminology, in order to accurately interpret nuanced questions, and provide contextually appropriate answers. Through an extensive experimentation and evaluation on the benchmark datasets, MedQuadQA demonstrates superior performance compared to other existing QA models, especially in tasks requiring a deep understanding of medical concepts and literature. We employ a diverse set of pre-trained language models, including BioGPT Causal, and LLAMA 2, GPT-2 and fine tune them to explore the effectiveness of various architectures in capturing the nuanced semantics and domain-specific knowledge prevalent in medical texts. One of the main aims of this paper is to present a viable approach to fine-tune models on resourcefully restricted systems like Colab Notebooks and Kaggle. Model performance and Evaluation metric tests such as ROUGE and Perplexity have been studied for the text generation task. The Rouge metric evaluates the recall of n-grams between the model outputs and expected answers, reflecting the overlap accuracy, while perplexity measures the model's surprise on encountering new data, indicating predictive fluency.**

***Index Terms*—BioGPT Causal, Llama 2, GPT-2, MedQuad**

I. Introduction

With this paper we attempt to solve and delve into the intricate challenges and methodologies involved to build a sophisticated Question Answering (QA) model specifically tailored for the medical field, MediQ-Ai. This model(s) uses few of the state-of-the-art pre-trained language models fine-tuned with cutting-edge techniques, which are the backbone of MediQ-Ai's training performed on the MedQuad dataset [1]. This is a comprehensive repository of medical questions paired with expert-annotated answers derived from authoritative sources(hospitals, other researchers and authors who performed studies in the same field). This study explores the capabilities of several pre-trained models—including BioGPT Causal, and LLAMA 2, GPT-2—to capture the nuanced semantics and domain-specific knowledge that are critical in the medical texts. We also explore ways to efficiently train models in computationally restricted set-ups which can be very useful for smaller companies and businesses who want to train their own models at lower costs.

We bring an innovative MediQ-Ai model that harnesses these potent pre-trained GPT and LLM architectures, which upon refining by fine-tuning them on MedQuad dataset. This process utilizes the models adaptability to easily navigate through complex medical terminology, while also accurately interpreting nuanced inquiries, and produce medically relevant responses that are contextually appropriate.

MediQ-Ai utilizes the powers of these pre-trained GPT and LLM models through fine-tuning based on the MedQuad dataset, enabling them to accurately comprehend complex medical terminology, interpret nuanced queries, and generate contextually medically relevant responses. Furthermore, these models incorporate interpretability mechanisms, which enhances trust and transparency by providing deep and useful insights into their reasoning process.

Our methodology comprises of fine-tuning these pre-trained models on the MedQuad dataset, by leveraging its comprehensive collection of these medical questions and annotated answers. Throughout the development process, we prioritize these two key performance metrics: ROUGE and perplexity. ROUGE, a metric commonly used in evaluating text summarization and generation tasks, measuring the quality of generated responses by assessing their similarity to reference answers. Perplexity, on the other hand, quantifies the model's ability to predict the next immediate token in a sequence, serving as a proxy for its overall language understanding and coherence.

Some specific challenges that we address in this paper are :

- **Domain-specific understanding**: Medical language is highly specialized and nuanced. Training LLMs to un-

derstand medical jargon and domain-specific knowledge poses a significant challenge.

- **Handling long-form content**: Medical information often involves lengthy and detailed explanations. Models must capture the key information while maintaining coherence and relevance in their responses.
- **Handling rare and complex cases**: Medical conditions vary widely in rarity and complexity, and QA systems must be capable of addressing a diverse range of queries.
- **Reliability and accuracy**: Ensuring that LLMs provide reliable and accurate information is crucial in medical QA systems.

  Addressing specific challenges such as the need for domain-specific accuracy, handling of lengthy and detailed medical content, and the ability to manage rare and complex cases, this paper underscores the critical attributes required for successful deployment of LLMs in the medical domain. Our exploration also prioritizes the reliability and accuracy of the generated answers, which are paramount in medical applications where the stakes are high.

II. Problem Statement

In times progress, in natural language processing (NLP) has brought about the creation of large pre trained language models that greatly improve the comprehension and generation of human language. In the field of medicine these models show promise in enhancing decision support systems and advancing patient care by offering contextually relevant responses to medical inquiries. This study aims to bridge this gap by developing a question answering (QA) model that utilizes cutting edge trained language models through fine tuning techniques. Specifically the research delves into the effectiveness of four trained models—BioGPT Causal, LLAMA 2 GPT 2—in capturing the intricacies of medical texts and producing contextually relevant responses.

The main goal is to create a QA model that excels at interpreting queries and understanding complex medical terminology while generating responses that are not just precise but also clinically significant. Moreover the model should be trained using limited resources to explore cost reduction in cloud subscriptions.

Key evaluation metrics for assessing these language models include ROUGE (Recall Oriented Understudy, for Gisting Evaluation) and perplexity, which gauge the quality of generated answers and the models grasp of language understanding abilities respectively.In tackling these obstacles and harnessing the potential of language models this research seeks to transform the way medical information is accessed and decision making is supported leading to patient care results and improved efficiency, in clinical processes.

III. Literature Review

With the Existing QA models, they demonstrate impressive performance in general domain. However they fall short when applied to medical literature due to the specialized nature of medical language. Additionally, the lack of interpretability and transparency in these models also hinders their adoption in clinical settings, where trust and accountability are paramount.

For the strong professionalism of knowledge in the medical field, the construction of knowledge graph should be targeted. How the system needs to accurately collect the questions that users want to ask is a big difficulty in the question answering system of deep learning. To solve the above problems, Authors [2] collected the data from the Tianchi Chinese data set Toyhom, the BERT model is used for word segmentation and vocabulary construction, Neo4j realizes the organization and storage of knowledge, and the naive Bayesian machine learning method is used for intention recognition. Based on the above technology, the medical knowledge graph is built, and the visual QA window is completed. Finally, this kind of medical automatic question-answering robot was realized.

K.Huang et al. [3] proposed a TensorFlow architecture-based approach to medical text generation that uses sequential models in Keras to transform the task into a classification problem by treating the text generation problem as a prediction problem. By training and experimenting with the model on a large-scale Chinese medical question-and-answer dataset, the results show that their model was a good fit with applications in this specific domain. Their model can aggregate medical domain knowledge, extract useful treatment information and generate medical knowledge text.

Lubna et al, [4] discuss about the work on medical image visual question answering done on the ImageCLEF 2019 medical VQA dataset. Visual question answering is a task where an image and a related question is given as input to the machine and we get a correct answer to the question as output. In their problem statement, both the input image and question are from medical domain. In medical imaging, VQA has applications like providing a second opinion to radiologists about their analysis of the image. It can also be used by the patients for getting a basic information about the image without consulting the doctor. Authors considered the problem of answering modality based questions for medical images like X-ray, Computed Tomography(CT), ultra sound(US), magnetic resonance imaging(MRI) etc. The approach used here is to use a Convolutional Neural Network(CNN) to classify the input image to its modality class and thus generate the answer according to the CNN output. The proposed model shows a testing accuracy of 0.838 which is comparable with state of the art.

We found Lora and Peft approaches from this paper, "Local LoRA: Memory-Efficient Fine-Tuning of Large Language Models," [10]. Here, the authors present a new method of memory-efficient fine-tuning for large language models (LLMs), which can be adapted to run even on consumer-grade hardware. The key innovation, to which we bestow the name "Local LoRA," is segmenting a model into chunks and fine-tuning each chunk independently by using localized loss functions. This will allow for the fine-tuning of much larger models than could be done normally for a given level of hardware, with respect to available memory. The results of this

experiment outlined in this paper indeed show that while Local LoRA cannot always compete with the high-level performance of end-to-end fine-tuning methods such as E2E LoRA, the baseline models have been outperformed in each case and thus must be considered a valuable choice for users with not so many resources at their disposal. The authors suggest that more work could be done in testing more sophisticated proxy models to come up with better approximations of gradients across model chunks, hence making Local LoRA more effective. The work also fits well into the broader discourse of making training large-scale models easier and less reliant on resources.

"Performance Analysis of LoRA Finetuning Llama-2" [5] focused on the exploration and open-sourcing of Llama-2, a significant LLM, through fine-tuning with the Low-Rank Adaptation (LoRA) technique. The satisfactory results obtained from the LoRA fine-tuning of Llama-2 have laid a foundation for further research in this domain. Its status as a transformer model, coupled with refined hyperparameter tuning, positions Llama-2 as a pivotal tool for research and practical applications in the foreseeable future. Their study demonstrated the enhanced performance of finetuned Llama-2, suggesting its potential for broader applications and furthering its significance in cutting-edge research. The dataset was imported from the hugging face space on which they tried to finetune the model.

N. Kazi et al. [6] explored the effectiveness of the RoBERTa Large model, an LLM trained on an extensive text corpus for language comprehension. By fine-tuning the model on the Multi-Genre Natural Language Inference (MNLI) corpus for semantic inference and subsequently on the SciEntsBank dataset, with a focus on the 3-way labels of correct, incorrect, and contradictory, They achieved a weighted Fl-score of 0.77, 0.72, and 0.72 on unseen answers, questions, and domains, respectively. Their model significantly benefits from fine-tuning on the MNLI corpus, particularly in enhancing its performance on the contradictory class through transfer learning leading to significant improvements on the more challenging test sets: unseen questions and unseen domains.

"Enhancing Transfer Learning of LLMs through Fine- Tuning on Task - Related Corpora for Automated Short-Answer Grading" [7] introduced LLaMA-Reviewer, a framework for automating the code review process using large language models (LLMs) and parameter-efficient fine-tuning (PEFT) techniques. Demonstrating the use of the smallest version of LLaMA with only 6.7B parameters and less than 1

Authors [8] evaluated the models based solely on their final decision accuracy, overlooking the long answer generation. Curie-Fine-Tuned achieved 68.1 accuracy on PubMedQA, falling short of Bio-GPT's 81. However, it excelled in BioASQ-task7b with a remarkable 99.6 score, surpassing BioLink-BERT. These results indicate GPT-3's fine-tuned models excel in closed-domain question-answering tasks. The weaker performance on Pub-MedQA may stem from its limited dataset size and fuzzy decision requirements. Conversely, Bio-GPT's superior performance may be attributed to its training on PubMed data. BioASQ's larger dataset and binary output requirement likely contributed to Curie-Fine-Tuned's exceptional performance. Overall, these findings underscore the potential of GPT-3's transfer learning in biomedical question-answering tasks.

K.P. Saikia et al. [9] performed a Empirical analysis that demonstrates supremacy of semantic search over the conventional approach of fine-tuned Large Language Models (LLMs) and ChatGPT-3.5. This superiority extends across two vital dimensions: the relevance of answers provided and the efficiency of response times. Semantic search consistently garners elevated cosine similarity values with ground truth answers, an indicator of its ability to furnish contextually enriched responses. This advantageous trait can be attributed to its bedrock of vector-based semantic comprehension, which facilitates meticulous matching and ranking of outcomes.

In the prior model performance study [11], the authors used three Bloomz models of various sizes , Bloomz-396M , Bloomz-1b4, Blooms-6b4, which were further optimized to decrease the size of their parameters . LoRA modules utilize Incremental Fine-tuning for several task types. As mentioned earlier, we group 23 unique task types and observe a substantial drop in task-type mixing perfor-mance for given tasks i.e., LoRA, based on OpenQA tasks, with all other tasks experiencing difficulty, while LoRA held out combined Ultimately, significantly lower perplexity of our proposed chain-of-LoRA approach is demonstrated. This can be seen as evidence of the effectiveness of their framework, which builds on this out an exploit the generative power of Language Model architectures, generating coherent responses, even when the mechanism for selecting the right task is inexact. The following experiment, which aims to evaluate the effectiveness of our task-selection module, referred to as P could be a substantial success. Using the Knuth-Morris-Pratt algorithm and a Top- k operation, we were able to achieve good precision in predicting instruction labels. Evaluation across different parameter sizes revealed promising results, with Incremental Fine-tuning (IF) showing slight improvements, albeit with increased computational demands. Despite these benefits, we maintained architectural consistency by employing LoRA modules consistently within the framework, ensuring operational simplicity and coherence. Our framework's effectiveness was further validated through response quality evaluations conducted by both GPT-3.5-Turbo and human annotators. Utilizing a scoring system ranging from 1 to 6 to gauge response helpfulness, GPT-3.5-Turbo evaluations demonstrated results akin to typical direct IF methods, highlighting the efficacy of our approach in enhancing response quality. Additionally, human evaluators' assessments revealed significantly reduced failure rates and optimized pass rates, reinforcing the efficacy of our methodology. These findings corroborate previous observations and emphasize the strong correlation between human and automatic evaluations, validating the robustness of our framework.

In this paper [12], Authors focused on generating a synthetic QA dataset using an adapted Translate-Align-Retrieve method. They created the largest Serbian QA dataset, which we name SQuAD-sr. To acknowledge the script duality in Serbian, they

generated both Cyrillic and Latin versions of the dataset. Authors also investigate the dataset quality and use it to fine-tune several pre-trained models. Best results were obtained by fine-tuning the BERTić model on Latin SQuAD-sr dataset, achieving 73.91

The Narrative question answering (QA) problem involves generating accurate, relevant, and human-like answers to questions based on the comprehension of a story consisting of logically connected paragraphs. Developing Narrative QA models allows students to ask about inconspicuous narrative elements while reading the story. However, this problem remains unexplored for the Arabic language because of the lack of Arabic narrative datasets. To address this gap, Authors [13] present the Arabic-NarrativeQA dataset, which is the first dataset specifically designed for machine-reading comprehension of Arabic stories. This dataset consists of two parts: translation of an English NarrativeQA dataset and a collection of new question-answer pairs based on Arabic stories. Furthermore, they implement the Arabic-NarrativeQA system using the Ranker-Reader pipeline, exploring and evaluating various approaches at each stage to identify the most effective ones. To avoid the need for an extensive data collection process, they utilize cross-lingual transfer learning techniques to leverage knowledge transfer from the English Narrative QA dataset to the Arabic-NarrativeQA system. Experiments show that incorporating cross-lingual transfer learning significantly improved the performance of the reader models. Furthermore, the question's evidence information provided in the Arabic-NarrativeQA dataset enables the learnable rankers to effectively identify and select the pertinent paragraphs. Finally, they examine and categorize challenging questions that require a deep understanding of the stories. By incorporating these question types into the introduced dataset, Authors show that existing reading comprehension models struggle to answer them, and further model development should be conducted

IV. Dataset Description

MedQuAD , the Medical Question Answering Dataset, which presents an extensive database of 16,413 instances of question-answers, carefully selected from the dataset of 12 prominent National Institutes of Health's websites . The dataset exhibits question and their answers, in the form of written text, that are the epitomes of information specific to knowledge on diseases. The information has been structured in such a way to cater to various enquires on health which include the treatment aspects, the modes of diagnosis, ailment consequential impacts and many more . Categorizing these instances, the dataset further contains a voluminous collection of 37 different question types, all structured and associated with resourceful questions and their answers. Whether delving into treatment modalities, unraveling diagnostic mysteries, or exploring the nuances of side effects, this dataset presents a new and diverse array of inquiries designed to illuminate and spread light on the complexities of healthcare.

Each question and answer traversed while navigating through the dataset includes the dimensions of diseases, medications, and medical tests. The varied spectrum of medical queries studied spans all the way from information regarding specific diseases to questioning the use of drugs and methods of diagnosis . The dataset is formatted in CSV and comprises four main components per question-answer pair: precisely, the question asked, the answer received, the source, and the focus disease of a question. Such organization substantially simplifies the search process and, hence, allows for more targeted analysis. As a result, the dataset maximizes the utilization of extensive medical experience to assist researcher and expert in analyzing it in a goal-oriented way and in aiding the AI-system realize its potential in medical research, study, and treatment.

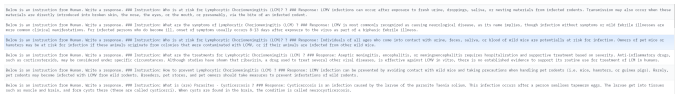

Fig. 1. Question and Answer

V. Methodology Description

This section provides more information about the different methodologies employed.

A. *Data pre-processing*

In this NLP task of creating the question-answering model using causal language modeling, the tokenizer plays an essential role in converting raw text into a structured format that can be processed by Large language models. First, the text is broken into (tokens). After that, a rich vocabulary is built, possibly domain-specific, with all the words having their unique numerical IDs for efficient processing and interpretation of the input by the model. Special tokens are also inserted by the tokenizer. These tokens help the model to understand the location of the text boundaries and maintain the logical flow; this is important for keeping coherent and contextually relevant responses. Broadly, tokenization serves to optimize the training of the model, affects performance during use, and ensures the text or the answers generated are coherent and contextually relevant to the learning domain.

B. *Domain specialized approach: BioBERT*

BioBERT is a variant of BERT (Bidirectional Encoder Representations from Transformers) pre-trained on biomedical text data, allowing its easy adaption in the biomedical domain with improved performance on given NLP tasks [21]. Unlike BERT that is pre-trained on general text data from the web, BioBERT is fine-tuned on biomedical literature, including articles from PubMed and other relevant sources [21]. This specialized training allows BioBERT to capture domain-specific terminology, syntax, and even the huge part of the json's semantic nuance involved in biomedical text and is hence valuable for any kind of task related to biomedical text mining, clinical natural language processing, or biomedical question answering.

C. *Large Language Models*

The following large language have been employed for this study.

– *GPT-2*

GPT-2(Generative Pre-trained Transformer 2) is an advanced model of the GPT model developed by OpenAI [18]. GPT features 117 million parameters whereas GPT-2 has several versions with the largest having up to 1.5 billion parameters [18]. GPT-2 is a transformer-based neural network model primarily designed for causal language modeling where it predicts the next word in a sequence given the previous words. Pre-trained on a wide-variety language pattern understanding from a large dataset known as WebText, GPT-2 learns and further sharpens when fine-tuned to tasks or datasets [18]. This model is great for tasks ranging from content creation to the generation of dialogues and any other applications that require the production of fluent, contextually suitable text that stretches over lengthy, coherent discourses.

– *Domain specialized approach: BioGPT*

BioGPT is a specialized variant of GPT [19]. It has been adapted for biomedical and scientific text generation tasks [19]. This transformer based model is specifically tailored to handle the complexities and unique characteristics of biological language [19]. Unlike GPT that is trained on general data this is trained on large datasets comprising scientific literature, such as research papers, medical texts, and other domain-specific documents [19]. This provides BioGPT with the capability to understand and generate text containing technical terminologies, complex sentence structures, and the context-specific nuances usually found in the biomedical field [19].

– *LLaMa-2*

LlaMa-2 is also a transformer-based language model developed by Meta [20]. It is designed for casual language modeling. It is also trained on data from books, websites and texts available up to a cut off date. Meta has made it more accessible to be part of academic and research work. it has released the model weights to researchers upon conditions for usage, making it more accessible than the latest versions of GPT. Like GPT, it is built in more than one size but focuses more on efficiency and performance optimization even in the small models, which are intended to be more computationally efficient while still delivering high performance [20]. In essence, LLaMa tries to deliver competitive or superior performance at lower computational costs and with fewer parameters. The difference between architecture of both the models is demonstrated in figure 2 [14].

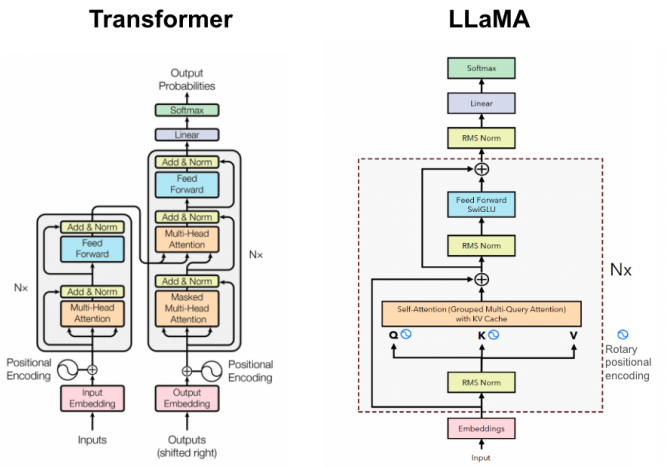

Fig. 2. Transformer Vs LLama

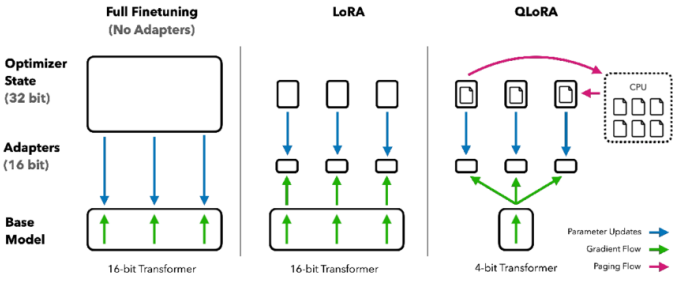

Fig. 3. Different finetuning methods and their memory requirements.

D. *QLoRA: Quantized Low Rank Adaption* QLoRA is a cutting-edge adaptation approach for fine-tuning large language models. The idea behind it is a low-rank matrix adaptation combined with quantization that largely reduces both computational demands and memory usage [15]. Where LoRA (Low-Rank Adaptation) uses only the low-rank matrices to effectively adapt the pre-trained model weights with minimum changes, QLoRA enhances it to quantize the weights in lower-bit formats [15]. This extra step reduces the resources required in processing and consequently further improves the suitability of the model for deployment on low-resource-constrained devices, offering gigantic advantages in terms of model size, computational speed, and energy consumed. This factor contributes much to the value QLoRA brings to the most resource-limited and capacity-restricted environments in high-efficiency applications [15]. Figure 3 [15] illustrates different fine-tuning methods commonly used and their corresponding memory requirements.

VI. Implementation

As we already specified the methods in the previous section, this section details our implementation approaches

with the various models for the QA answering.
We have used a method called Quantization which is used to reduce the precision of the model's weights to optimize memory usage and inference speed. We configured quantization parameters using the "bits and bytes" library, setting it to use 4-bit precision respect to every model. A trainer is configured with various training arguments, such as the output directory, number of epochs, batch size, and other hyperparameters.
Additionally, parameters specific to supervised fine-tuning, such as max sequence length and packing, are set here. After configuring the dataset, models, tokenizers, and trainers, The models are trained on the custom dataset for a specified number of epochs. While training the models learn to generate text relevant to the task based on the patterns it identifies in the dataset. After training trained models can be evaluated using metrics relevant to the task. This could involve measuring its performance on a validation set or testing it on unseen data to assess its generalization capabilities. Depending on the performance of the trained model, further fine-tuning iterations may be necessary. This could involve adjusting hyperparameters, dataset augmentation, or incorporating additional data to improve the model's performance.
We imported the respective model from hugging face that we want to train and we set the QLoRA parameters which consist of attention dimension, alpha parameter for LoRA scaling which is set to 16, dropout probability for LoRA layers which is set to 0.1. Additionally, We also set the activation of 4-bit precision of base model loading to True and set the quantization type to "nf4"(weights initialized using a normal distribution), while also enabling the nested quantization for 4-bit based models. Finally set the Training argument parameters with 3 epochs. Enabling fp16/bf16(retains more precision for both the weight and gradient) for A100 GPU supported training in colab. By setting the training and validation batch size to 4, the authors ensure that models doesn't exceed the available memory while training. After setting the optimizer and it's parameters(weight decay, learning rate) and SFT parameters, we load the models and initiate the fine tuning the models.
QLORA(Figure 4. [15]) was introduced to design multiple innovations in order to reduce memory use without sacrificing performance:

- 4-bit NormalFloat, an information theoretically optimal quantization data type for normally distributed data that yields better empirical results than 4-bit Integers and 4-bit Floats.
- Double Quantization, a method that quantizes the quantization constants, saving an average of about 0.37 bits per parameter (approximately 3 GB for a 65B model).
- Paged Optimizers, using NVIDIA unified memory to avoid the gradient checkpointing memory spikes that occur when processing a mini-batch with a long sequence length.

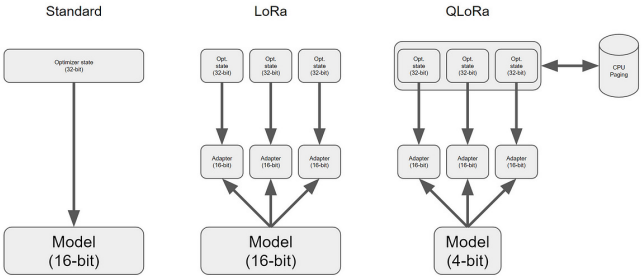

Fig. 4. QLoRA: Fine-Tune a Large Language Model

Fine-tuning GPT-2(Figure 5 [16]) was challenging and time-consuming, however after applying parameter efficient fine-tuning and eventually, it helped us improve the results and efficiency. Additionally, data augmentation techniques such as paraphrasing, synonym replacement, or back-translation helped in increasing the size and diversity of our data. Regularization techniques like dropout, weight decay, or gradient clipping were utilized to reduce complexity and variance of our pre-trained model. Finally, early stopping criteria such as the validation loss, accuracy, or perplexity helped determine to stop the training process.
A note on Llama Model: We spent a significant amount of time fine-tuning the llama model and setting optimal parameters using Lora and Peft for it. We finally trained llama over 3 epochs and the model fine tuning kept crashing initially due to CUDA Out of Memory errors. To solve this our research led to Lora and PEFT as defined in the recent sections. We were then able to fine tune the model using A100 GPU on Kaggle (limited to 30 hours per week), but Inferencing the LLaMA model on a test dataset like MedQuAD with 1000 questions and answers still posed significant challenges. Even on powerful platforms like Google Colab Pro equipped with an A100 GPU boasting 40GB of VRAM and 80 GB of system RAM, we couldn't gather the prediction set for a 1000 datapoints. The primary issue stems from the massive memory requirements of LLaMA, especially for handling large batches of data or complex queries that increase the computational load substantially. Even with a robust setup, the CUDA out-of-memory error indicates that the GPU memory was insufficient to accommodate the model and the data simultaneously. Attempts to mitigate this by switching to CPU-based computation resulted in system RAM being overwhelmed, leading to a total crash of the notebook. That is the reason we do not have results from the Llama model but here is a link to a fine-tuned model on the MedQuad dataset that could be useful for future research work and comparison. [17]

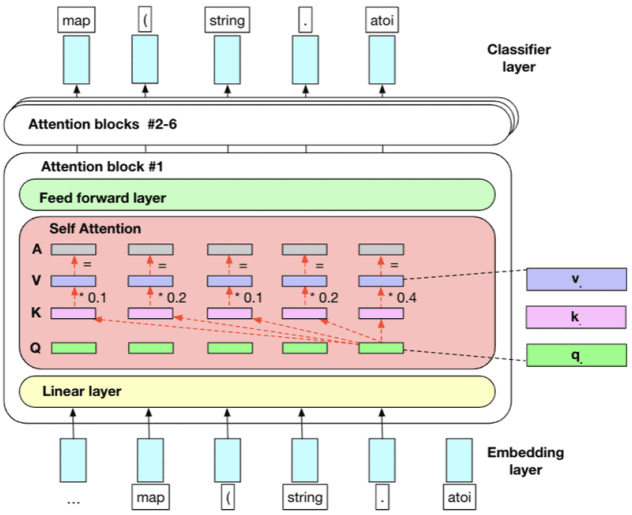

Fig. 5. Fine tuning GPT-2

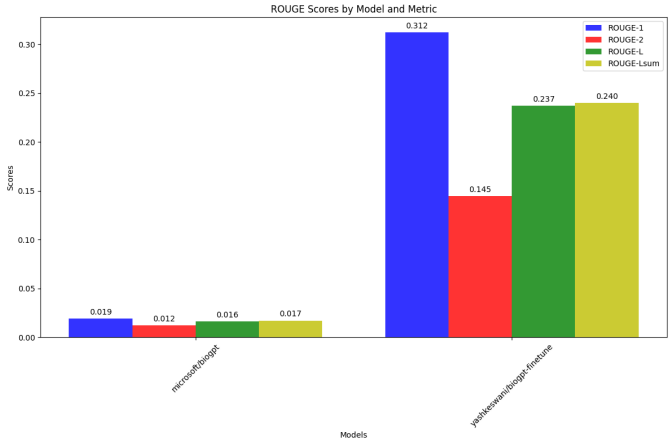

Fig. 6. BioGPT - ROUGE Scores

VII. EVALUATION AND RESULTS

A. *Evaluation Metrics*

As discussed throughout this paper, one of our main tasks was to take a deep dive into evaluation metrics for the sequence generation task. For this evaluation we used two popular metrics, ROUGE (Recall-Oriented Understudy for Gisting Evaluation) and Perplexity. These are prominently used in evaluating sequence generation tasks such as text summarization and language modeling. Both ROUGE and perplexity offer improvements over simpler evaluation metrics like accuracy or error rate, which fail to capture the nuanced aspects of language such as semantic coherence, relevance, and fluency in generated texts. These metrics address specific qualities essential for the evaluation of complex sequence generation tasks, providing more detailed and relevant assessments of model performance in natural language processing applications.

- **ROUGE**: It is one of the most popular automatic metrics developed by Chin-Yew Lin to compare summary documents or machine-translated sentences with a set of model human summaries or translations. ROUGE is essentially a measure of how much of the content in the reference texts is captured by the generated text, making it particularly useful for tasks like summarization, where the goal is to condense the text but leave key information preserved. Most of the variants—ROUGE-N (measuring n-gram overlaps), ROUGE-L (measuring the longest common subsequence), and ROUGE-S (measuring skip-bigram co-occurrence statistics)—have been proposed to capture different dimensions of text similarity. The recall-based nature of ROUGE and its variants provides a rich methodology to assess the completeness and relevance of the generated text systematically with respect to the source content.
- **Perplexity**: Proposed by researchers in the field of information theory and later modified for language modeling, is a statistic measure of a language model with regard to its prediction ability over a string of words. These words are used for the calculation of the normalized inverse probability of the test set. Perplexity provides a clear measure of the likelihood of a sequence of words occurring according to the model, especially for translation tasks, machine translation, or continuous text generation. A lower perplexity score shows a more accurate text sequence prediction in a model, which suggests the higher linguistic fluency of a model and fewer surprises in word choice. This allows comparing performances of different models on the same dataset, giving intuition of which model best understands the language patterns behind the data.

The next section discusses results obtained on the basis of the studied metrics.

B. *Results Obtained*

ROUGE:

We have taken 4 models for evaluation - base versions of BioGPT and GPT along with fine-tuned versions of the same models. We have already specified our fine-tuning approach in the previous sections. Figure 6 and Figure 7 demonstrate the individually obtained rouge scores on a test set of 1000 data points borrowed from the MedQuad Dataset.

The provided bar chart in Figure 8. illustrates the ROUGE scores of four models on the Medquad QA dataset for sequence generation evaluation. The models being compared are

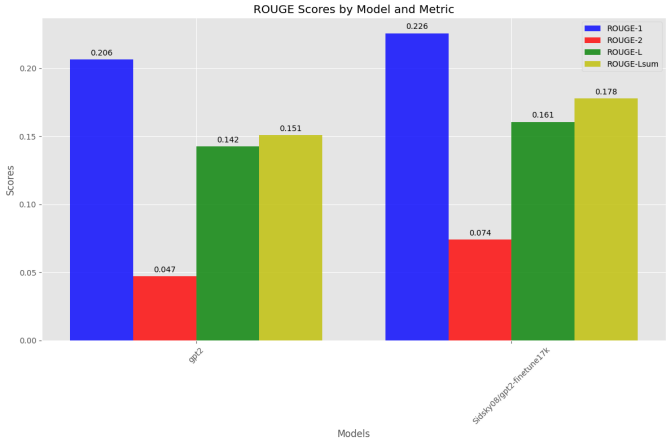

Fig. 7. GPT - ROUGE Scores

microsoft/biogpt, gpt2, yashkeswani/biogpt-finetune, and Sidsky08/gpt2-finetune17k. The scores presented are for ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-Lsum, which are standard metrics for evaluating the quality of generated text against reference texts.

Here's a breakdown and analysis of the ROUGE scores: **ROUGE-1 (Blue)** measures the overlap of unigrams between the generated text and the reference text. It primarily evaluates the models' ability to capture the most important keywords and concepts from the reference.

The Sidsky08/gpt2-finetune17k model has the highest ROUGE-1 score, suggesting that it is most effective at capturing key unigrams present in the reference summaries.

ROUGE-2 (Red) measures the overlap of bigrams between the generated text and the reference text. It is indicative of the models' ability to generate coherent phrases and handle the order of words. The Sidsky08/gpt2-finetune17k model outperforms others in ROUGE-2 as well, indicating its strength in creating coherent bi-gram sequences that are present in the references.

ROUGE-L (Green) measures the longest common subsequence and is often used as a proxy for assessing the fluency of the generated text. Here, the yashkeswani/biogpt-finetune model has a comparable score to Sidsky08/gpt2-finetune17k, suggesting it has similar capabilities in generating fluent and coherent text.

ROUGE-Lsum (Yellow) is like ROUGE-L but applied to the entire summary instead of sentence by sentence. It reflects the models' ability to produce structured, long-form content.

Again, Sidsky08/gpt2-finetune17k scores the highest, indicating it generates summaries that are structurally closest to the references.

From this analysis, it is evident that the fine-tuned models (yashkeswani/biogpt-finetune and Sidsky08/gpt2-finetune17k) generally perform better than their non-fine-tuned counterparts. This improvement is likely due to the adaptation of these models to the specifics of the Medquad dataset through fine-tuning, which allows them to generate more contextually relevant and coherent responses. The Sidsky08/gpt2-finetune17k model, which has been fine-tuned with Qlora and PEFT parameters, shows the best overall performance across all ROUGE metrics. This suggests that the fine-tuning process, which presumably included a learning phase specifically tailored to the Medquad dataset's domain, has significantly enhanced the model's text generation capabilities, aligning it more closely with the expert annotations present in Medquad.

In conclusion, the fine-tuned models demonstrate superior performance in generating sequences that are more aligned with the expected responses. This indicates the effectiveness of fine-tuning strategies like Qlora and PEFT for the task of sequence generation on domain-specific datasets such as Medquad.

Perplexity:

The Figure 9 shows the perplexity scores with which models evaluate on the Medquad QA dataset. Perplexity is among the measurement criteria to gauge the effectiveness of a probability model when forecasting a sample, hence in language models, a low perplexity score means good performance.

Following is the analysis of the perplexity scores:

microsoft/biogpt (Red Bar): This model returned the highest perplexity score of 48.66 out of the four. Consequently, a higher perplexity score suggests that the model has less certainty in its predictions and is generally not well-tuned, especially with regard to the specifics of this dataset, compared to others.

Gpt2 (Blue Bar): Standard model gpt2 scores 23.11 in perplexity, far better than the perplexity of the previous microsoft/biogpt, but still remained on the higher end. It shows that there is still room for improvement in its prediction capability for this dataset.

yashkeswani/biogpt-finetune (Light Blue Bar): It registered a per-PleXity score of 4.91 for the fine-tuned Biogpt, which is significantly lesser when compared to its non-fine-tuned counterparts. Such a huge improvement reflects the success of fine-tuning on domain-specific data, as it gives the model more skills to make predictions about the samples.

Sidsky08/gpt2-finetune17k (Navy Blue Bar): This model shows the lowest perplexity score of 4.36, hence falls under Sidsky08/g. This may mean that

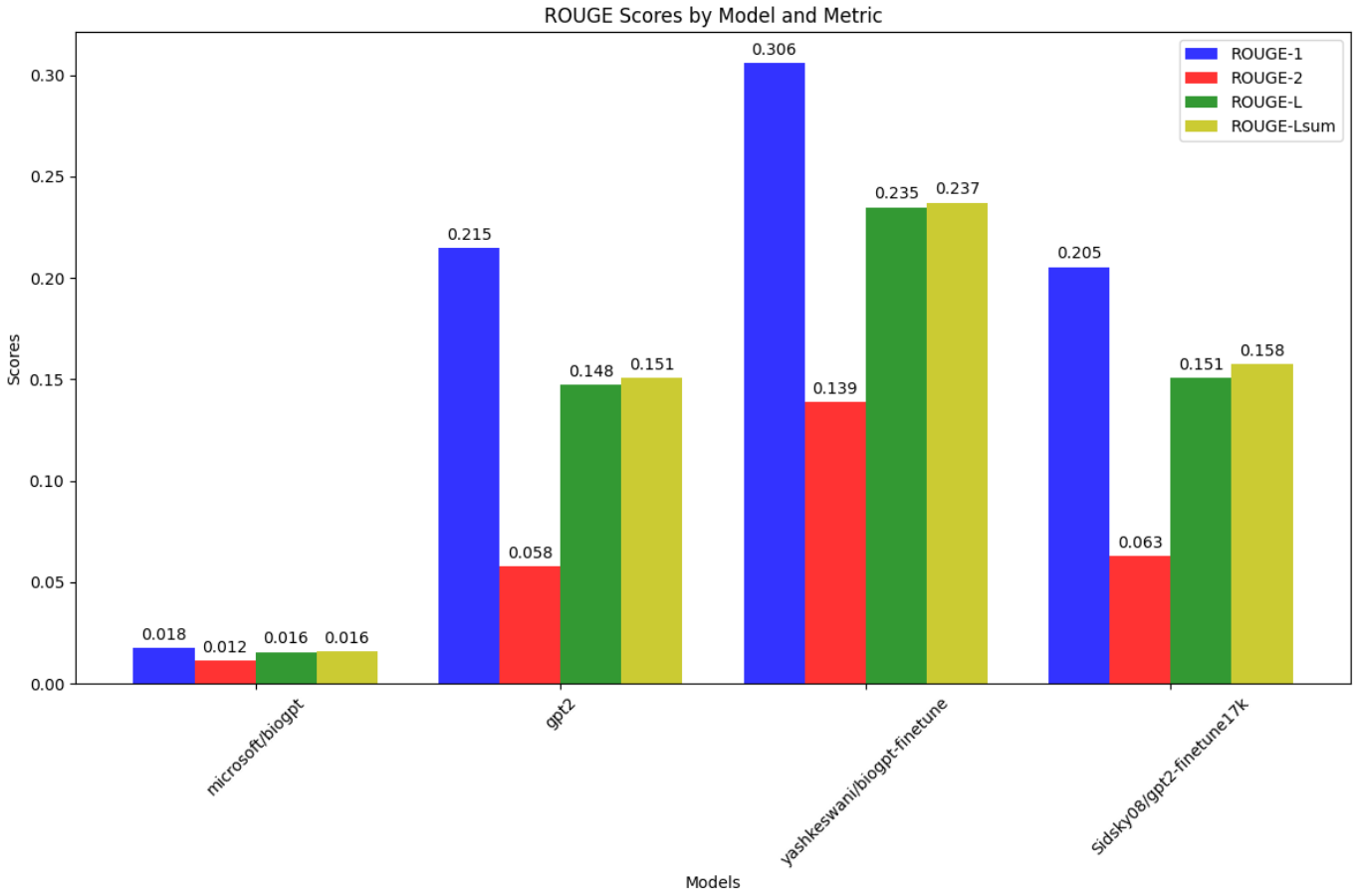

Fig. 8. ROUGE Scores - All models

a fine-tuned process, probably using techniques and data that are domain-specific in the Medquad set, could have made this model very proficient in predicting the sequences within this domain, giving a better-fitted model that has the greatest predictive power.

General Evaluation: While Both fine-tuned models (yashkeswani/biogpt-finetune and Sidsky08/gpt2-finetune17k) also clearly outperform the non-fine-tuned models, as is visible in lower perplexity and higher ROUGE scores for these. The results of this work match the implications on the importance and effectiveness of fine-tuning large language models for task-specific specialized cases.

Especially, the Sidsky08/gpt2-finetune17k model gives the best result, showing its fine-tuning process is very optimized for the Medquad QA dataset. Moreover, the color gradient of the chart from the maximum score's red to the minimum score's blue visually emphasizes the relative performance level of each model, and it clearly gives the idea that the fine-tuning process is very important for the performance enhancement of the models.

Can you make future work and conclusion for the same work. In terms of Loran and PEFT and use cases along with how this work of fine-tuned LLMs for specific tasks such as Medquad is useful.

VIII. CONCLUSION

Fine-tuning LLMs with techniques like Qlora and PEFT has markedly enhanced their performance with limited resources on specialized tasks, such as medical question answering demonstrated by the Medquad dataset. The fine-tuned models not only achieved lower perplexity but also scored higher on ROUGE metrics, indicating a more precise alignment with reference answers. The standout model, based on the finetuning of gpt2 - Sidsky08/gpt2-finetune17k, exemplifies the potential for tailored fine-tuning to yield models with superior predictive accuracy. These advancements illustrate the transformative impact that fine-tuned LLMs can have on healthcare, offering professionals refined tools for information retrieval and decision support. PEFT and Lora offer smaller companies cost-effective strategies to tailor large language models to their specific needs without the prohibitive expense of full

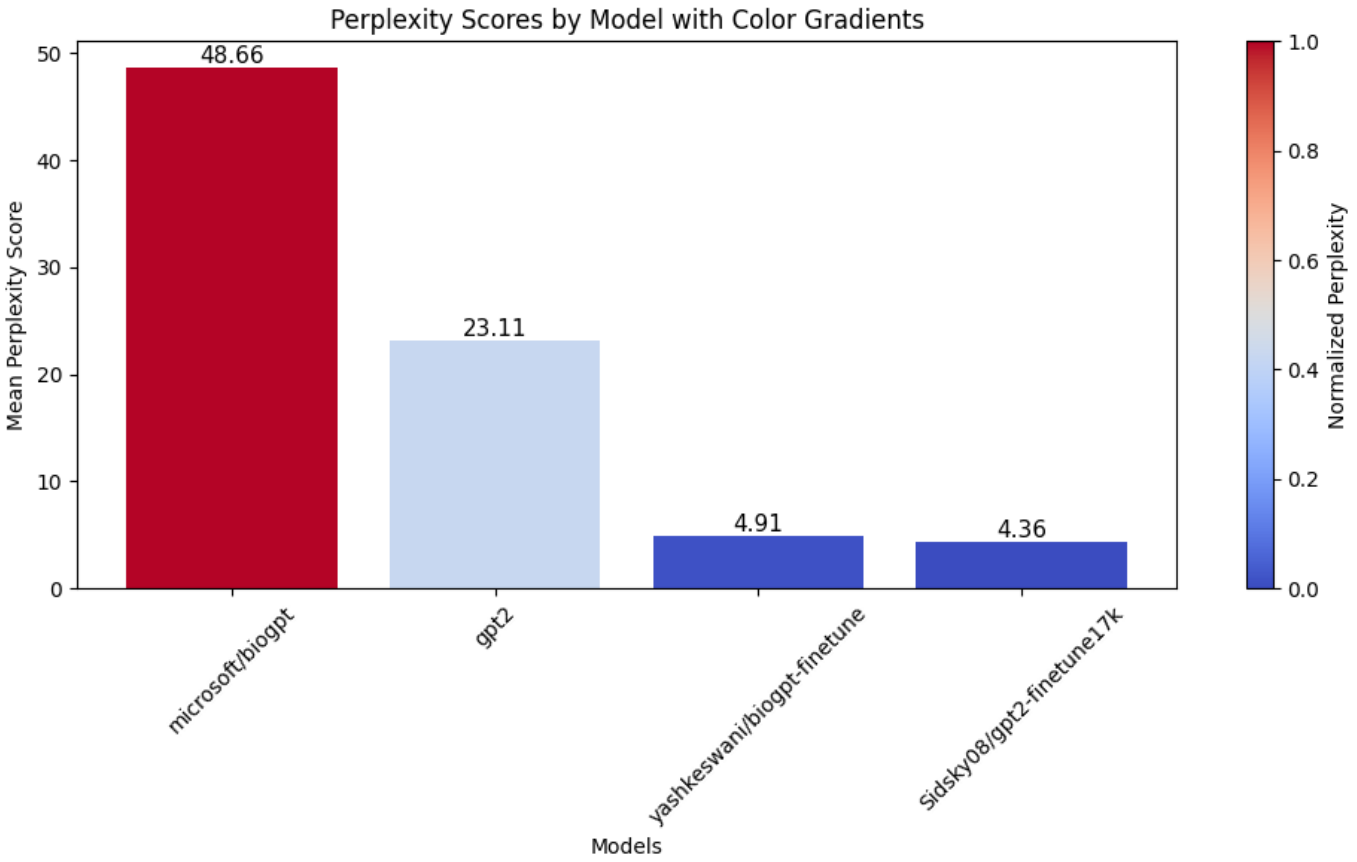

Fig. 9. Perplexity - All models

model training, allowing them to leverage cutting-edge AI within their budget constraints for enhanced performance and competitiveness in their respective markets. The journey from data to diagnosis may become more efficient as AI becomes a more integrated part of the medical field, underscoring the significance of ongoing research in this realm.

IX. FUTURE WORK

The current success of applying LoRA and PEFT to fine-tune large language models on the Medquad QA dataset opens several pathways for future research. As we explore the potential of large language models (LLMs) in specialized domains, we're entering new frontiers. Building on the success of Qlora and PEFT fine-tuning for the Medquad QA dataset, future research could delve deeper into parameter-efficient methodologies, such as prompt tuning and adapter layers, to enhance model accuracy. The goal of making these fine-tuning techniques work across different fields encourages us to apply them in diverse areas such as legal and financial applications apart from the medical domains. This broad application ensures that the methods are universally useful and adaptable.. Enhancing interpretability, particularly in healthcare, and exploring hybrid models could improve predictive modeling and diagnostic accuracy. Moreover, optimizing hyperparameters through methods like neural architecture search could refine model configurations to boost efficiency. Extending these advancements to multilingual settings could democratize access to quality healthcare information, setting new standards for AI in global service.

X. REFERENCES

REFERENCES

[1] https://huggingface.co/datasets/Tonic/medquad

[2] H. Shi, X. Liu, G. Shi, D. Li and S. Ding, "Research on medical automatic Question answering model based on knowledge graph," 2023 35th Chinese Control and Decision Conference (CCDC), Yichang, China, 2023, pp. 1778-1782, doi: 10.1109/CCDC58219.2023.10327124. keywords: Deep learning;Epidemics;Computational modeling;Knowledge graphs;Medical services;Big Data;Question answering (information retrieval);Knowledge Graph;Retrieval Question Answering;Chinese Medical Text;Transformer,

[3] K. Huang, F. Ji, W. Lu and Y. Xiao, "Research on Text Generation of Medical Intelligent Question and Answer Based on Bi-LSTM and Neural Network Technology," 2022 IEEE/ACIS 22nd International Conference on Computer and Information Science (ICIS), Zhuhai, China, 2022, pp. 54-59, doi: 10.1109/ICIS54925.2022.9882349. keywords: Training;Information science;Computational modeling;Neural networks;Transforms;Predictive models;Question answering (in-

formation retrieval);text generation;smart question and answer;medical;TensorFlow;sequential model,

[4] A. Lubna, S. Kalady and A. Lijiya, "MoBVQA: A Modality based Medical Image Visual Question Answering System," TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), Kochi, India, 2019, pp. 727-732, doi: 10.1109/TENCON.2019.8929456. keywords: Biomedical imaging;Task analysis;Knowledge discovery;Feature extraction;Visualization;Natural languages;Training;visual question answering;medical image analysis;deep learning;artificial intelligence;natural language processing,

[5] A. Pathak, O. Shree, M. Agarwal, S. D. Sarkar and A. Tiwary, "Performance Analysis of LoRA Finetuning Llama-2," 2023 7th International Conference on Electronics, Materials Engineering Nano-Technology (IEMENTech), Kolkata, India, 2023, pp. 1-4, doi: 10.1109/IEMENTech60402.2023.10423400. keywords: Adaptation models;Computational modeling;Force;Transformers;Artificial intelligence;Faces;Tuning;LLMs;Llama-2;LoRA;GPT-4;Hugging Face,

[6] N. Kazi and I. Kahanda, "Enhancing Transfer Learning of LLMs through Fine- Tuning on Task - Related Corpora for Automated Short-Answer Grading," 2023 International Conference on Machine Learning and Applications (ICMLA), Jacksonville, FL, USA, 2023, pp. 1687-1691, doi: 10.1109/ICMLA58977.2023.00255. keywords: Deep learning;Transfer learning;Semantics;Benchmark testing;Task analysis;Tuning;Automated Short Answer Grading;ASAG;Large Language Models;Transfer Learning;Semantic Inference;Recognizing Textual Entailment;MNLI;SemEval;SciEntsBank,

[7] J. Lu, L. Yu, X. Li, L. Yang and C. Zuo, "LLaMA-Reviewer: Advancing Code Review Automation with Large Language Models through Parameter-Efficient Fine-Tuning," 2023 IEEE 34th International Symposium on Software Reliability Engineering (ISSRE), Florence, Italy, 2023, pp. 647-658, doi: 10.1109/ISSRE59848.2023.00026. keywords: Codes;Automation;Quality assurance;Software reliability;Task analysis;Tuning;Software engineering;Code Review Automation;Large Language Models (LLMs);Parameter-Efficient Fine-Tuning (PEFT);Deep Learning;LLaMA;Software Quality Assurance,

[8] H. Bousselham, E. H. Nfaoui and A. Mourhir, "Fine-Tuning GPT on Biomedical NLP Tasks: An Empirical Evaluation," 2024 International Conference on Computer, Electrical Communication Engineering (ICCECE), Kolkata, India, 2024, pp. 1-6, doi: 10.1109/ICCECE58645.2024.10497313. keywords: Adaptation models;Technological innovation;Biological system modeling;Predictive models;Propulsion;Parallel processing;Transformers;Large Language Models;GPT-3;Fine-tuning;NLP;Embeddings;Pretrained models,

[9] K. P. Saikia, D. Mukherjee, S. Mahapatra, P. Nandy and R. Das, "Unveiling Deeper Petrochemical Insights: Navigating Contextual Question Answering with the Power of Semantic Search and LLM Fine-Tuning," 2023 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), Greater Noida, India, 2023, pp. 881-886, doi: 10.1109/ICCCIS60361.2023.10425564. keywords: Semantic search;Soft sensors;Petrochemicals;Information retrieval;Question answering (information retrieval);Time factors;Business;semantic search;finetuning;large language model (LLM);petrochemical;generative artificial intelligence,

[10] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models.

[11] X. Qiu, T. Hao, S. Shi, X. Tan and Y. -J. Xiong, "Chain-of-LoRA: Enhancing the Instruction Fine-Tuning Performance of Low-Rank Adaptation on Diverse Instruction Set," in IEEE Signal Processing Letters, vol. 31, pp. 875-879, 2024, doi: 10.1109/LSP.2024.3377590. keywords: Task analysis;Training;Computational modeling;Adaptation models;Predictive models;Optimization;Graphics processing units;Large language models;low-rank adaptation;instruction fine-tuning,

[12] A. Cvetanović and P. Tadić, "Synthetic Dataset Creation and Fine-Tuning of Transformer Models for Question Answering in Serbian," 2023 31st Telecommunications Forum (TELFOR), Belgrade, Serbia, 2023, pp. 1-4, doi: 10.1109/TELFOR59449.2023.10372792. keywords: Adaptation models;Natural languages;Benchmark testing;Transformers;Question answering (information retrieval);Telecommunications;Synthetic data;Natural Language Processing;Question Answering;Neural Networks;Transformer,

[13] M. A. Ateeq, S. Tiun, H. Abdelhaq and N. Rahhal, "Arabic Narrative Question Answering (QA) Using Transformer Models," in IEEE Access, vol. 12, pp. 2760-2777, 2024, doi: 10.1109/ACCESS.2023.3348410.

[14] Pickleprat, "Tweaking the Transformer: LLaMa," Medium, Apr. 04, 2024. https://medium.com/@pickleprat/tweaking-the-transformer-llama-95d77e747b91 (accessed Apr. 28, 2024).

[15] Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L. (2023). QLoRA: Efficient Finetuning of Quantized LLMs. arXiv [Cs.LG]. Retrieved from http://arxiv.org/abs/2305.14314

[16] zhats, "Training and Fine-Tuning GPT-2 and GPT-3 Models Using Hugging Face Transformers and OpenAI API," It-Jim, Jun. 15, 2023. https://www.it-jim.com/blog/training-and-fine-tuning-gpt-2-and-gpt-3-models-using-hugging-face-transformers-and-openai-api/ (accessed Apr. 28, 2024).

[17] Sidsky08, Llama-2-7b-chat-finetune17k. Hugging Face Model Hub, 2023. [Online]. Available: https://huggingface.co/Sidsky08/Llama-2-7b-chat-finetune17k.

[18] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models are Unsupervised Multitask Learners," 2019. Available: https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf

[19] R. Luo et al., "BioGPT: generative pre-trained transformer for biomedical text generation and mining," Briefings in Bioinformatics, Sep. 2022, doi: https://doi.org/10.1093/bib/bbac409.

[20] H. Touvron et al., "LLaMA: Open and Efficient Foundation Language Models." Available: https://research.facebook.com/file/1574548786327032/LLaMA–Open-and-Efficient-Foundation-Language-Models.pdf

[21] J. Lee et al., "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," Bioinformatics, vol. 36, no. 4, Sep. 2019, doi: https://doi.org/10.1093/bioinformatics/btz682. keywords: Task analysis;Question answering (information retrieval);Decoding;Transformers;Transfer learning;Context modeling;Vocabulary;Arabic question answering;answer generation;cross-lingual transfer learning;reading comprehension;narrative QA,