

Introduction

Language is at the heart of health and medicine, underpinning interactions between people and care providers.

Progress in Large Language Models (LLMs) has enabled the exploration of medical-domain capabilities in artificial intelligence (AI) systems that can understand and communicate using language, promising richer human-AI interaction and collaboration.

Medical question-answering (QA) systems have a significant impact on healthcare as they deliver precise and timely information to healthcare providers and patients. The development of intelligent systems for medical QA involves creating tools that can understand and provide accurate responses to questions regarding medical knowledge and healthcare practices. These systems serve a crucial function by offering valuable insights, evidence-based recommendations, and rapid access to medical information. By assisting healthcare professionals in making well-informed decisions and aiding patient education, medical QA systems contribute to enhanced healthcare outcomes. Moreover, they facilitate efficient information retrieval and knowledge discovery, bridging the divide between medical literature and practical medical decision-making.

Existing medical QA systems have made significant advancements in recent years. However, achieving high accuracy remains a challenge due to various factors such as the complexity of medical language, semantic understanding, and the need to handle a wide range of medical domains and topics. Inaccurate answers may lead to incorrect diagnoses, treatment recommendations, or patient misunderstanding. Therefore, improving the accuracy of medical QA systems is crucial to ensure reliable and trustworthy information retrieval, ultimately enhancing patient care and medical decision-making processes.

Our objectives are twofold: to enhance the accuracy of generative medical question-answering systems (where questions are answered without limiting available information to a pre-specified source) and to leverage the capabilities of the BioGPT Large [1] Model to address the limitations of existing approaches. We aim to optimize its performance for this specific task, providing accurate answers even in the absence of contextual information.

Related Work

The advent of transformers [2] and large language models (LLMs) has renewed interest in the possibilities of AI for medical extractive question-answering tasks—a long-standing “grand challenge”.

A majority of these approaches involve smaller language models trained using domain specific data (BioLinkBert [3], PubMedGPT [4], PubMedBERT [5], BioGPT), resulting in a steady improvement in state-of-the-art performance on benchmark datasets such as MedQA (USMLE) [6], MedMCQA [7], and PubMedQA [8].

However, with the rise of larger general-purpose LLMs such as GPT-3 [9] and Flan-PaLM [10] trained on internet-scale corpora with massive computing, we have seen leapfrog improvements on such benchmarks, all in a span of a few months.

Despite the significant advancements achieved by GPT-3 and Flan-PaLM, the release of BioGPT large introduced a remarkable performance breakthrough in the field., achieving an accuracy of 81% on the benchmark datasets, establishing itself as a state-of-the-art model.

Table 1 | Summary of the performance on PubMedQA dataset.

Model	Accuracy	Released date	Size	Owned By
BioGPT-Large	81	Mar 2023	1.5B	Microsoft
Med-PaLM 2	79.2	Mar 2023	NA	Google Research & DeepMind
GPT-4-base	80.4	Apr 2023	NA	Microsoft & OpenAI
GPT-3.5 + Z-Code++	79.6	Jan 2023	175B	Microsoft Azure AI
Flan-PaLM (3-shot)	79	Dec 2022	540B	Google Research & DeepMind
				Technical University of Denmark & Copenhagen University
Codex (5-shot)	78.2	Dec 2022	175B	Hospital
Human Performance	78	Sep 2019	NA	University of Pittsburgh & Carnegie Mellon University
GPT-4	75.2	Mar 2023	NA	Microsoft & OpenAI
PubMedGPT	74.4	Dec 2022	2.7B	Stanford University
PMC-LLaMA	73.4	Apr 2023	7B	Shanghai Jiao Tong University
BioLinkBERT (large)	72.2	Mar 2022	340M	Stanford University
BioLinkBERT (base)	70.2	Mar 2022	110M	Stanford University
BioBERT (multi-phase tuning)	68.1	Sep 2019	110M	University of Pittsburgh & Carnegie Mellon University
BioELECTRA	64	June 2021	110M	SAAMA AI Research Lab
PubMedBERT	55.8	July 2020	110M	Microsoft Research

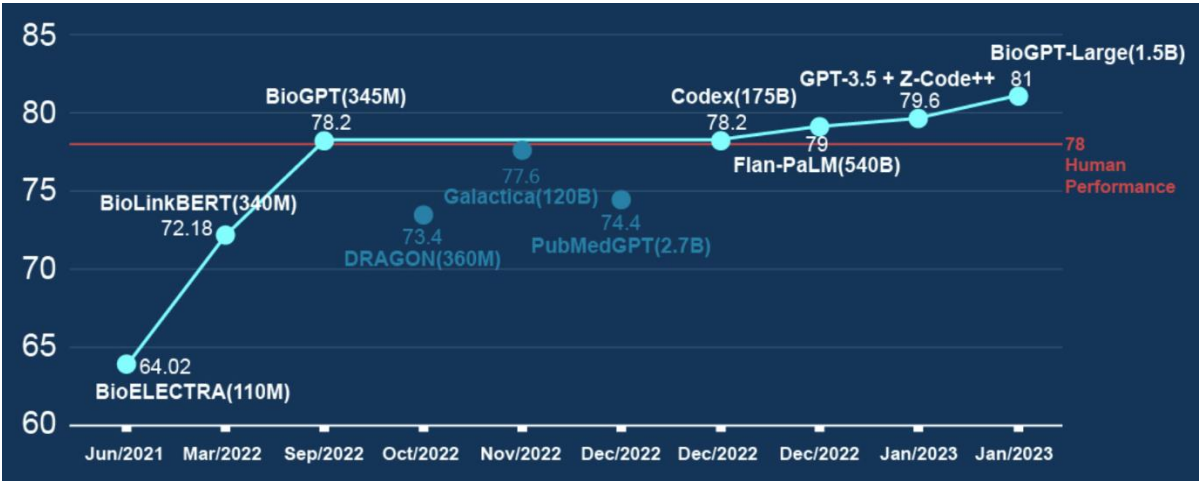


Figure 1: BioGPT large performance on PubMedQA, BioGPT large achieved an accuracy of 81%, establishing itself as a state-of-the-art model on extractive question-answering task. data source from PubMedQA website [11]

While our research primarily focuses on the generative question-answering task, we have observed certain limitations in the performance of models in the desired task, human evaluation revealed that further work was needed to ensure the AI output, including long-form answers to open-ended questions, is safe and aligned with human values and expectations in this safety-critical domain (a process generally referred to as "alignment") like Flan-PaLM which exhibit impressive capabilities in answering multiple-choice questions, their proficiency in providing accurate responses to consumer medical queries reveals notable gaps.

Additionally, BioBERT and PubMedBERT obtain superior performances than general pre-trained models on biomedical benchmarks. However, previous works mainly focus on BERT models which are more appropriate for understanding tasks, not generation tasks.

Furthermore, it is important to note that several models predominantly function as extractive question answering systems, wherein they predict a specific span of text from the source context as the answer. Additionally, they may also predict categorical labels, such as "yes" or "no," for certain types of questions like BioGPT large. However, models that possess the ability to generate answers in a more free-form and scientific manner, known as generative question answering, are relatively rare in the current landscape.

These limitations emphasize the need for further research and development in order to address the challenges posed by consumer medical questions and to enhance the capabilities of question answering models in the medical domain.

As Gu et al. [12] point, training only on in-domain data from scratch is important for a specific domain, Therefore, we adapt BioGPT large which pre-trained from scratch on medical data and established itself as a state-of-the-art model on extractive question-answering task, for our downstream task, generative question answering.

We recognized the significance of fine-tuning pre-trained language models, specifically BioGPT large, for the downstream task of generative question answering in the medical domain. By fine-tuning BioGPT large, we aimed to improve the accuracy of answering medical questions, even in the absence of provided context. This approach is crucial to ensure that the AI output adheres to human values and expectations in this critical and safety-sensitive domain.

Methodology

BioGPT, which was announced by Microsoft, can be used to analyze biomedical research with the aim of answering biomedical questions and can be especially relevant in helping researchers gain new insights. BioGPT is a type of generative language model, which is trained on millions of biomedical research articles that have already been published. This essentially means that BioGPT can use this information to perform other tasks like answering questions, extracting relevant data, and generating text relevant to biomedical.

The researchers used GPT-2 XL as the primary model and trained it on 15 million PubMed [12] abstracts before using it in the real world. GPT-2 XL is a Transformer decoder which has 48 layers, 1600 hidden size and 25 attention heads resulting in 1.5B parameters in total.

The core component of Transformer as well as BioGPT is the multi-head attention. Given the input, three linear transformations are applied to produce the query Q, the key K and the value V, and then the output is calculated as follows:

$$\text{Multihead}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_h)W,$$

$$\text{head}_i = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d}}\right) V_i,$$

where (1) h is the number of heads; (2) Q, K and V are equally split into Q_i , K_i and V_i along the feature dimension, $i \in \{1, 2, \dots, h\}$; (3) Concat denotes concatenating all inputs as a large tensor along the feature dimension; (4) W is the parameter for the affine transformation. The output of multi-head attention layer is then fed into a feed-forward layer to construct a Transformer layer (or Transformer block).

Training criteria: BioGPT large is trained via the standard language modeling task. Let $D = \{x_i\}_i$ the collection of sequences, and sequence x_i is made up of n_i tokens, i.e., $x_i = (s_1, s_2, \dots, s_{n_i})$. The training objective is to minimize the negative log-likelihood:

$$\min -\frac{1}{|D|} \sum_{i=1}^{|D|} \sum_{j=1}^{n_i} \log P(s_j | s_{j-1}, s_{j-2}, \dots, s_1).$$

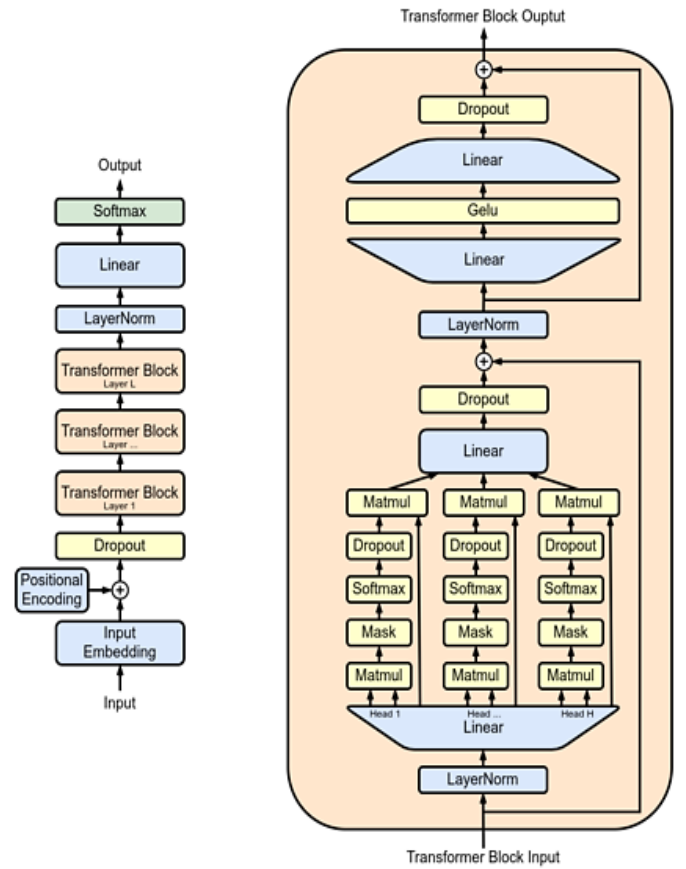


Figure 2: BioGPT architecture, decoder-only transformer model

Our research emphasizes the Generative Question Answering Task, for which we conducted fine-tuning of BioGPT large using the PubMedQA dataset. PubMedQA is a closed-domain question answering dataset specifically designed to assess questions that require comprehensive medical research comprehension skills. It encompasses a wide range of topics within the biomedical domain and utilizes PubMed abstracts as the primary source of information.

PubMedQA [8] (Closed-domain question answering given PubMed abstract) dataset contains questions on biomedical research which cover a wide range of biomedical topics, including diseases, treatments, genes, proteins, and more. PubMedQA is one of the MultiMedQA datasets (A benchmark for medical question answering). PubMedQA consists of 1k expert labeled, 61.2k unlabeled and 211.3k artificially generated QA instances with yes/no/maybe multiple-choice answer and long answer given a question together with a PubMed abstract as context.

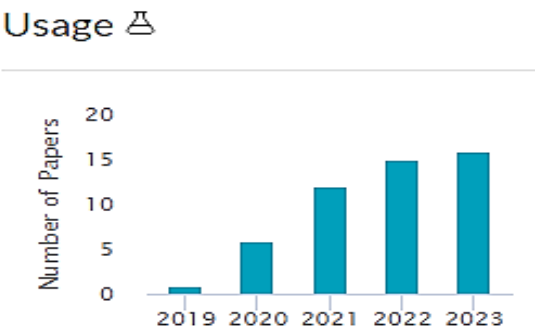


Figure 3: number of papers used PubMedQA on the medical domain per year

Table 2: PubMedQA Sample

Example
<p>Pubid: 25,444,977</p> <p>Question: "Does reference range for serum and salivary testosterone in young men of Mediterranean region?"</p> <p>Long_answer: "In men from the Mediterranean region, values of TT>9.7 nmol/L, CFT>0.22 nmol/L, and/or CBT>4.9 nmol/L make the presence of biochemical hypogonadism unlikely. According to the correlation between serum and ST, the clinical value of ST remains to be established."</p> <p>Context: {["The interassay variability found in the measurement of testosterone (T) levels warrants the need for laboratories to validate their methods to establish trustworthy cut-off points for diagnosis of male hypogonadism. The aims of this study were to validate measurement of total T (TT) at our laboratory in order to obtain reference ranges for TT, calculated free T (CFT), calculated bioavailable T (CBT), and salivary T (ST) in healthy young men from the Mediterranean region, and to evaluate the potential clinical value of ST by establishing its correlation with serum T.", "An observational, cross-sectional study with sequential sampling.", "men aged 18-30 years with body mass index (BMI)<30.", "chronic diseases, hepatic insufficiency or use of drugs altering circulating T levels. Main outcome measures TT (chemiluminescent immunoassay UniCell DXI 800 [Access T Beckman Coulter]), CFT and CBT (Vermeulen's formula), and ST (radioimmunoassay for serum TT modified for saliva [Coat-A-Count, Siemens]). Descriptive statistical analyses and correlation by Spearman's rho (SPSS 19.0 Inc., Chicago) were used.", "One hundred and twenty-one subjects aged 24±3.6 years with BMI 24±2.5 kg/m(2) were enrolled. Hormone study: TT, 19±5.5 nmol/L (reference range [rr.] 9.7-33.3); CFT, 0.38 nmol/L (rr. 0.22-0.79); CBT, 9.7 nmol/L (rr. 4.9-19.2); and ST, 0.35 nmol/L (rr. 0.19-0.68). Correlation between ST and CFT was 0.46."], "labels": ["OBJECTIVE", "METHODS", "METHODS", "METHODS", "RESULTS"], "meshes": ["Adolescent", "Adult", "Body Mass Index", "Cross-Sectional Studies", "Humans", "Hypogonadism", "Luminescent Measurements", "Male", "Mediterranean Region", "Radioimmunoassay", "Reference Values", "Saliva", "Testosterone", "Young Adult"] }</p> <p>final decision: "Yes"</p>

To adapt BioGPT large to assess the limitations we mentioned, we used the version of BioGPT large which finetuned on PubMedQA dataset and applied soft prompt in prefix tuning instead of full model finetuning on a small set of the

PubMedQA which effectively used to instruct the model to produce text generations more aligned with the requirements of the medical domain.

Soft Prompt in Prefix Tuning [13]: Because Large Language Models (LLMs) with hundreds of billions of parameters have become computationally expensive and very time-consuming to fine-tune. To address this issue, Lester et al [14]. introduced prompt tuning, a computationally efficient method for adapting LLMs to specific downstream tasks, even with limited data, a small number of labeled examples (e.g., tens)

Soft prompts in prefix tuning leverage continuous embeddings (virtual tokens) to steer the pre-trained language model by directly appending several additional virtual tokens before the text as the prompts while keeping the remaining LLM parameters frozen. Such continuous embeddings are randomly initialized and learned end-to-end on the downstream tasks to be task specific. we didn't append the virtual tokens to the very beginning of the source input, but only before the target sequence (between the source and the target). Equipped with the prompt, our final sequence is constructed as [source; prompt; target]

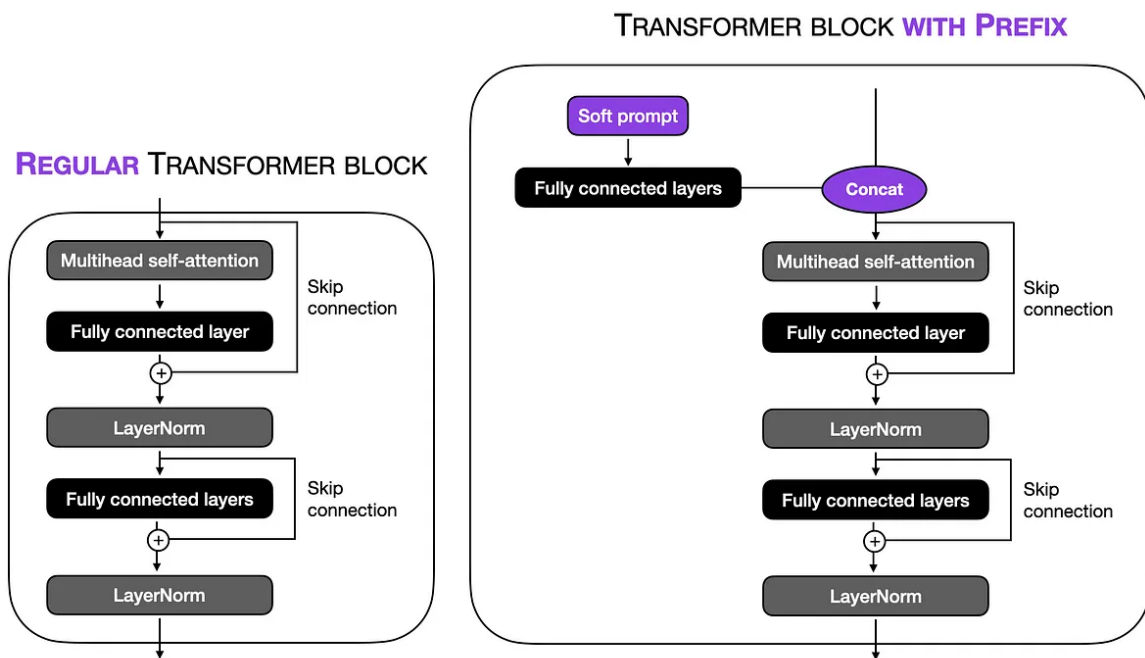


Figure 4: Illustration of prefix Tuning

Fine-Tuning Setup: we performed soft prompt in prefix tuning technique on the BioGPT large 1.5B model. The virtual tokens length was set to 10, allowing us to focus on a specific context within the input sequence. By freezing the remaining parts of the model, we limited the number of trainable parameters to 1.5 million.

During the training process, we utilized a TPU VM v3-8 with a batch size of 32. This enabled us to execute the training procedure over 80 steps, with each step involving the processing of 512 tokens. The Adam optimizer was employed, utilizing a peak learning rate of 1×10^{-3} to optimize the model's performance over the course of 5 epochs.

Experiments and Results

The following components can be included in the experimental setup:

- **Dataset:** PubMedQA
- **Model Architecture:** GPT-2 XL (Generative Pre-trained Transformer)
- **Pretrained Model:** BioGPT-Large
- **Fine-Tuning:** Soft prompt in prefix Tuning
- **Baselines:** Pretrained BioGPT and BioGPT large models without fine-tuning.
- **Evaluation Metrics:**
 - **Perplexity:** measures how well a language model predicts a given sequence of tokens. Lower perplexity indicates better performance, as it suggests that the model is more confident in its predictions. Perplexity is commonly used for evaluating language models trained on large corpora.

Perplexity is defined as the exponentiated average negative log-likelihood of a sequence. If we have a tokenized sequence $X = (x_0, x_1, \dots, x_t)$, then the perplexity of X is,

$$\text{PPL}(X) = \exp \left\{ -\frac{1}{t} \sum_{i=1}^t \log p_{\theta}(x_i | x_{<i}) \right\}$$

$$\text{PPL}(X) = \exp \left\{ -\frac{1}{t} \sum_{i=1}^t \log p_{\theta}(x_i | x_{<i}) \right\}$$

- **Human Evaluation:** Human evaluation involves soliciting human judges to rate or rank the quality of the generated text. This can be done using criteria such as fluency, coherence, relevance, and overall quality. Human evaluation provides valuable insights into the subjective aspects of text generation, which may not be fully captured by automated metrics.

In the performance evaluation methodology, the generative question answering dataset is split into training and validation. The baseline, Pretrained BioGPT and BioGPT large models without fine-tuning, are evaluated on the validation set using the chosen evaluation metrics. The best performing fine-tuned BioGPT model is selected based on evaluation results on the validation set using the chosen metrics. The selected model is evaluated on the validation set using the evaluation metrics to compare its performance against the baseline. Statistical analysis, such as t-tests or significance tests, is conducted to determine if the performance improvements of the fine-tuned BioGPT model are statistically significant compared to the baselines. Human evaluation is conducted with a subset of generated answers from the selected model and baselines, where human judges rate or rank the quality, relevance, and correctness of the answers based on predefined criteria. Additionally, a qualitative analysis is performed to assess aspects like relevance, correctness, and coherence by manually reviewing samples of generated answers and comparing them to the ground truth or desired answers.

In this section, we present the results of our model with the baselines, qualitative analysis to assess aspects of the samples generated and compare them with the baselines, and Model convergence to assess its generalization capabilities.

Table 3: Summary of the Performance of our finetuned model with the baselines

Model	Perplexity
BioGPT-prefix-tuning	35.8934
BioGPT large	32.7705
BioGPT	38.1253

Table 4: Examples of BioGPT-Fine-tuned comparison to Physician, and baselines (BioGPT and BioGPT large)

Question	Physician	BioGPT	BioGPT large	BioGPT-Fine-tuned
How can I stop diarrhea?	Without treatment, most diarrhoea resolves on its own. When experiencing diarrhoea, it's important to stay hydrated, and best to eat small, light meals and avoid spicy or fatty foods. Diarrhoea due to an infection should not be treated with imodium.	The gut microbiota (dysbiosis) and immune system have a vital role in the pathogenesis of irritable bowel syndrome (IBS), in conjunction with gut motility disorders, and this link is strongly linked to the type of food ingested by affected individuals	Patients using a polyethylene glycol-electrolyte solution for constipation, have diarrhea if the solution is left in the bowels for long enough (~ 1-2 h at most). Once diarrhea is apparent (1 2 9 g / day), reinstituting the full prescribed dose in two doses (or 1 L / day in adults for two doses) in 2 4 h will stop the diarrhea.	We recommend prescribing prokinetic drugs as first-line treatment for persistent diarrhea, and that a trial of loperamide may also be considered.
Does stress cause nosebleed?	Most people with mild symptoms of Listeria require no treatment as the body is able to fight off infection without intervention. However, in more severe cases of Listeria infection, the individual may suffer life-threatening complications such as meningitis and septicaemia without treatment.	We explored the involvement of the HPA axis in stress- and vasopressin-induced noseblee by using rats to mimic the clinical feature of primary enuresis.	several studies have pointed to stress as a pathogenic factor, especially in combination with predisposing factors. It is unlikely that stress is the main cause of attacks of nosebleed. The exact mechanisms by which stress may initiate nosebleed are not fully understood. Some patients may find the urge to nosebleed anxiety-induced and / or fear-induced, and some patients may have an autonomic imbalance.	Chronic stress resulting from long work hours and overtime is believed to increase the production of free radicals, which causes endothelial dysfunction and finally leads to hypertension. Oxidative stress might also induce tissue damage, including nasal mucosa-causing epistaxis. The aim of this study was to investigate a potential relationship between chronic stress and epistaxis. Our study included 97 soldiers who were referred to the ear nose throat clinic with ongoing nosebleed in the absence of any systemic or local trauma. Ninety-seven soldiers who were coming to the military training area

but did not have nosebleed comprised the control group. A sociodemographic data form and Beck's Depression Inventory (BDI) were administered to both groups. The number of hours worked weekly was recorded for both groups. Participants whose number of working hours exceeded the recommended level (minimum: 50 hours a week, maximum: 80 hours a week) were defined as having high work load. Participants with a BDI score equal to or greater than ten were diagnosed with probable depression. The stress levels of the study and control groups were compared. The groups were also compared with respect to the number of working hours and the existence of probable depression. There was no statistically significant difference between the groups with respect to number of working hours ($p > 0.05$).

Discussion

When fine-tuning BioGPT or any language model for generating medical answers without context using soft prompt fine-tuning, there are several strengths and limitations to consider:

Strengths:

- **Domain-Specific Knowledge:** Fine-tuning BioGPT which pretrained on medical data only enables the model to capture domain-specific knowledge from biomedical literature and research papers. This can enhance the quality and accuracy of the generated medical answers.
- **Language Understanding:** BioGPT models, when fine-tuned for medical domains, have the ability to understand and generate text specific to medical questions. This allows them to generate relevant and accurate answers in the context of medical queries.
- **Flexibility in Answer Generation:** Soft prompt fine-tuning allows for more flexibility in generating answers without context. By providing prompts or cues during fine-tuning, the model can learn to generate appropriate responses that align with the given question.
- **Context-Agnostic Generation:** Generating answers without context can be beneficial in scenarios where specific context or additional information is not available. It allows the model to generate stand-alone responses based solely on the provided question, making it applicable in a broader range of use cases.

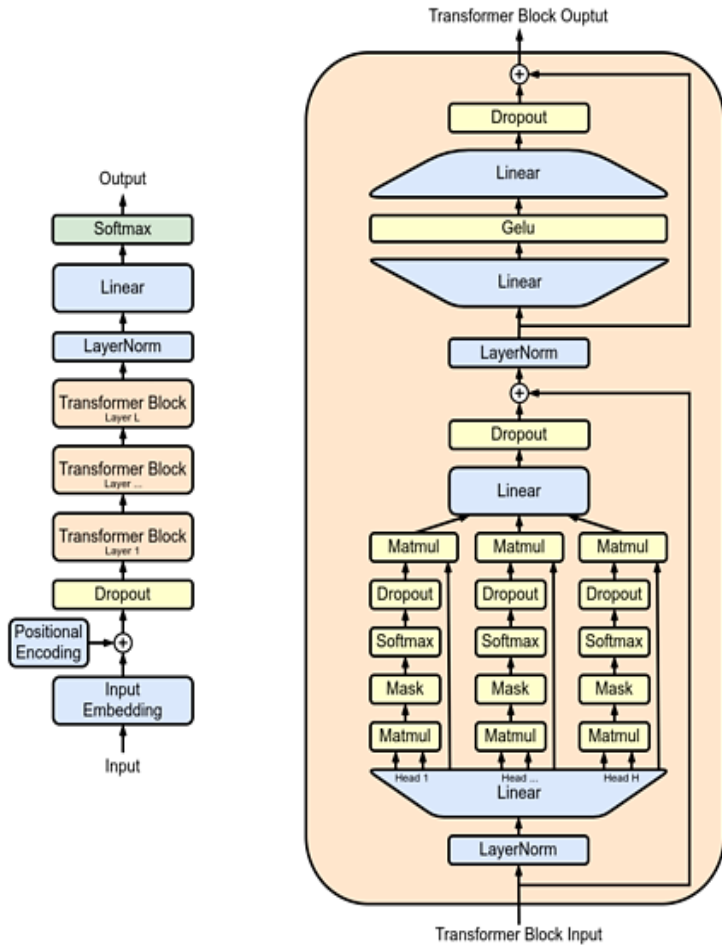
Limitations: given the broad and complex space of medical information needs, methods for measuring the alignment of model outputs will require continued development. For instance, it is likely that additional dimensions, such as the empathy conveyed by answers, are important but not measured in this study. Therefore, further research is needed to enhance the rigor of rubrics enabling human evaluation of language model (LLM) performance in medical question answering.

We encountered a significant challenge in terms of computation cost and time while attempting to fine-tune the BioGPT large model. This model consists of an extensive architecture with 48 layers and a staggering 1.5 billion parameters. As a result, the traditional fine-tuning approach proved to be computationally expensive and time-consuming.

To address this challenge, we made a strategic decision to modify our fine-tuning approach by implementing soft prompt finetuning. This alternative approach offered a promising solution to mitigate the computational burden and reduce the overall time required for fine-tuning.

By adopting soft prompt finetuning, we were able to optimize the utilization of computational resources and streamline the training process. This approach introduces a more efficient and effective utilization of computational power, resulting in substantial reductions in both computation cost and time requirements.

Diagram



Algorithm: Proposed Methodology for fine-tuning BioGPT large

1. Load the PubMedQA dataset.
2. Load the pretrained tokenizer and BioGPT model.
3. Prepare the dataset by extract question-answer pairs from the PubMedQA dataset.
4. Split the dataset into a training set and an evaluation set, with a ratio of 100:50 (train:eval).
5. Preprocess the dataset by converting it into input features suitable for the model. This includes tokenization, adding special tokens, and formatting the data.
6. Create data loaders to load the dataset in batches for efficient training.
7. Instantiate the BioGPT model with the configuration of PrefixTuning.
8. Set up the hyperparameters, including learning rate (lr), optimizer, number of training epochs (num_epochs), and batch size.
9. Perform the fine-tuning process on the training dataset for the specified number of epochs (5 epochs).
10. Evaluate the fine-tuned model using appropriate metrics (perplexity, loss, and BLEU score) to assess the performance of the model on the evaluation set.
11. Create a prompt structure to obtain inferences from the fine-tuned model.

References

1. Luo, Renqian, et al. "BioGPT: generative pre-trained transformer for biomedical text generation and mining." *Briefings in Bioinformatics* 23.6 (2022).
2. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
3. Yasunaga, M., Leskovec, J. & Liang, P. LinkBERT: Pretraining Language Models with Document Links. *arXiv preprint arXiv:2203.15827* (2022).
4. Bolton, E., Hall, D., Yasunaga, M., Lee, T., Manning, C. & Liang, P. Stanford CRFM Introduces PubMedGPT 2.7B <https://hai.stanford.edu/news/stanford-crfm-introduces-pubmedgpt-27b>. 2022.
5. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J. & Poon, H. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)* 3, 1–23 (2021).
6. Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H. & Szolovits, P. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences* 11, 6421 (2021).
7. Pal, A., Umapathi, L. K. & Sankarasubbu, M. MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering in Conference on Health, Inference, and Learning (2022), 248–260.
8. Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W. & Lu, X. PubMedQA: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146* (2019).
9. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems* 33, 1877–1901 (2020).
10. Jin et al. (2019) PubMedQA, PubMedQA Homepage. Available at: <https://pubmedqa.github.io/> (Accessed: 08 June 2023).
11. Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021
12. National Library of Medicine. (2021). PubMed Labs. PubMed Labs. <https://pubmed.ncbi.nlm.nih.gov/>
13. Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online, August 2021. Association for Computational Linguistics
14. Lester, B., Al-Rfou, R. & Constant, N. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691* (2021).
15. Ben Abacha, Asma, and Dina Demner-Fushman. "A question-entailment approach to question answering." *BMC bioinformatics* 20.1 (2019): 1-23.
16. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311* (2022).