Palmyra-Med: Instruction-Based Fine-Tuning of LLMs Enhancing Medical Domain Performance

Kiran Kamble **Waseem AlShikh**

Writer, Inc.
{kiran,waseem}@writer.com

Abstract

The development of Large Language Models (LLMs) has greatly impacted natural language processing tasks by demonstrating exceptional performance across various applications. However, a significant challenge faced by these models is their limited capacity to comprehend and generate contextually relevant, user-instructed responses, which is particularly crucial in the sensitive medical domain. In this study, we finetuned and evaluated two cutting-edge LLMs, Palmyra-20b and Palmyra-40b, for medical language understanding tasks. By employing instruction-based fine-tuning on a custom-curated medical dataset of 200,000 examples, we create novel, fine-tuned models, Palmyra-Med-20b and Palmyra-Med-40b. Performance is then measured across multiple medical knowledge datasets, including PubMedQA and MedQA. Our fine-tuned models outperform both their base counterparts and other LLMs pretrained on domain-specific knowledge. Notably, our models display enhanced performance compared to the GPT4 model evaluated in a few-shot setting on the PubMedQA dataset. This research demonstrates the effectiveness of instruction-based fine-tuning in enhancing LLMs performance in the medical domain.

1 Introduction

Advancement of Large Language Models (LLMs) such as GPT-4 (OpenAI 2023) and PaLM 2 (Anil et al. 2023) has revolutionized the field of natural language processing, demonstrating exceptional performance in various tasks ranging from text generation to question-answering. Pretrained on massive corpora, these models possess the ability to generate human-like text and effectively answer context-based questions, thus opening doors to numerous applications across many domains. One key area where these advancements hold transformative potential is the medical domain, wherein accurate assessment of domain knowledge and reasoning capabilities is essential for informed decision-making and favorable patient outcomes.

Despite the remarkable generalization capabilities demonstrated by LLMs, a significant challenge they face is their limited capacity to accurately follow user instructions and generate meaningful, contextually relevant responses. Addressing this limitation is particularly crucial in the medical domain, as inaccuracies may lead to severe consequences. To this end, instruction-based fine-tuning has emerged as a promising strategy to improve LLMs' abilities to understand and effectively respond to a wide range of task-specific prompts, enhancing their applicability for real-world scenarios.

In this study, we focus on the evaluation and fine-tuning of two powerful LLMs, Palmyra-20b and Palmyra-40b, within the context of medical language understanding tasks. Our research comprises instruction-based fine-tuning on both models, utilizing a custom-curated medical dataset containing 200,000 medical examples. The resulting fine-tuned models, referred to as Palmyra-Med-20b and

Preprint. Under review.

Palmyra-Med-40b, are then evaluated against various medical datasets, such as PubMedQA (Jin et al. 2019), MedQA (Zhang et al. 2018), to determine improvements in their performance across distinct aspects of medical knowledge and reasoning.

This paper is organized as follows. In Section 3, we provide a comprehensive overview of the baseline models, followed by a detailed description of the instruction fine-tuning process, including the dataset used and the fine-tuning intricacies. The latter half of Section 3 presents the evaluation datasets and methods employed in assessing the performance of the fine-tuned models. Section 4 focuses on the results of our experiments, where we present a comparative analysis of the baseline and fine-tuned models' performance on respective tasks. Finally, Section 5 offers the conclusion and outlines recommendations for future work in the areas of medical language understanding and large language model fine-tuning.

2 Experiments

In this experiments section, we begin by providing a comprehensive overview of the baseline models utilized in our investigation. Subsequently, we elucidate the details of instruction finetuning, delving into the dataset employed for this purpose and the intricacies of the finetuning process. Lastly, we elaborate on the evaluation datasets, as well as the methods employed for assessing the performance of the fine-tuned models.

2.1 Baseline LLMs

In our experiments, we employed two baseline models. The first baseline model, referred to as Palmyra-20b, developed by the Writer NLP team, is a large decoder-only transformer model that has been pre-trained on the Pile dataset (Gao et al. 2020), which was assembled by Eleuther AI and tokenized using the GPT2 (Radford et al. 2019) BPE tokenizer. This model was developed utilizing Nvidia's Nemo Megatron framework, which integrates both the Nemo and Megatron-LM (Shoeybi et al. 2020) frameworks for distributed training of LLMs. Palmyra-20b is a GPT-based model featuring 44 transformer layers, a hidden size of 6144, and 48 attention heads, with a sequence length of 2048. The model was trained in a distributed fashion by employing the distributed version of the Adam optimizer and was configured with a tensor parallelism of 4 and a pipeline parallelism of 1.

The second baseline model employed in our experiment was the Palmyra-40b model. This model consists of 40 billion parameters and is a decoder-only transformer, trained on 1 trillion tokens of the RefinedWeb dataset (Penedo et al. 2023) – a novel, massive web dataset derived from CommonCrawl. Developed by TII, Palmyra-40b is a GPT-based model similar to Palmyra-20b, featuring 60 transformer layers, a hidden size of 8192, and 128 attention heads. While both models are trained in bf16 precision, there are several key differences between them. Notably, Palmyra-40b utilizes FlashAttention (Dao et al. 2022) and MultiQuery Attention (Shazeer 2019), resulting in faster inference.

The primary distinction between multihead attention and multiquery attention lies in their configuration: multihead attention deploys one query, key, and value per head, whereas multiquery attention shares a singular key and value across all heads. Although the impact of multi query attention on pretraining may not be significant, it considerably enhances the scalability of inference time by maintaining a smaller K,V cache during autoregressive decoding, subsequently reducing memory costs during inference. Furthermore, we observed a faster evaluation time for the Palmyra-40b model compared to other models of similar size.

2.2 Instruction Finetuning

Large Language Models (LLMs) have demonstrated remarkable generalization capabilities, such as in-context learning (Brown et al. 2020) and chain-of-thought reasoning (Wei et al. 2023). However, one prominent challenge these models face is their limited capacity to adequately follow user instructions and produce meaningful, contextually relevant responses. To address this limitation, instruction fine-tuning has been introduced as a method to refine and align LLMs' performance to better adhere to user instructions and provide contextually appropriate outputs. This technique is especially vital for

enhancing the model's capacity to understand and respond effectively to a wide range of task-specific prompts.

Capitalizing on the promising results achieved through instruction fine-tuning, researchers have successfully leveraged this technique to train more efficient and performant, publicly available language models. In a significant breakthrough, Taori et al (Taori et al. 2023) drew inspiration from the self-instruct approach described by Wang et al (Wang et al. 2023) and fine-tuned LLaMA (Touvron et al. 2023)models on a dataset consisting of 52,000 instructions, leading to the development of the Alpaca 7B model. This model closely mimics the behavior of OpenAI's text-davinci-003 but is considerably more computationally efficient. In a continuing effort to refine LLMs, Chiang et al (Chiang et al. 2023) introduced the Vicuna model, trained from LLaMA 13B using 70,000 publicly shared ChatGPT logs from the ShareGPT platform (Noa 2023). As a result of instruction fine-tuning, the Vicuna model demonstrates performance closely comparable to the GPT-3.5. These examples underline the significance of instruction fine-tuning in optimizing the practical use of LLMs for real-world applications.

In our research, we applied the same instruction fine-tuning protocol to adapt our baseline LLMs to our custom-curated medical dataset. We first present a description of the training dataset, followed by an explanation of the fine-tuning process.

2.3 Dataset

For the fine-tuning of our LLMs, we used a custom-curated medical dataset that combines data from two publicly available sources: PubMedQA (Jin et al. 2019) and MedQA (Zhang et al. 2018).The PubMedQA dataset, which originated from the PubMed abstract database, consists of biomedical articles accompanied by corresponding question-answer pairs. In contrast, the MedQA dataset features medical questions and answers that are designed to assess the reasoning capabilities of medical question-answering systems.

We prepared our custom dataset by merging and processing data from the aforementioned sources, maintaining the dataset mixture ratios detailed in Table 1. These ratios were consistent for finetuning both Palmyra-20b and Palmyra-40b models. Upon fine-tuning the models with this dataset, we refer to the resulting models as Palmyra-Med-20b and Palmyra-Med-40b, respectively.

Dataset	Ratio	Count
PubMedQA	75%	150,000
MedQA	25%	10,178

Table 1: Dataset mixture used for training.

PubMedQA: This dataset focuses on questions related to biomedical research. Models are provided with paper abstracts from PubMed and are tasked with answering multiple-choice questions. The dataset is divided into three subsets: labeled (PQA-L), unlabeled (PQA-U), and artificially generated (PQA-A). For our training, we employed the PQA-A subset, containing 150,000 question-answer pairs.

MedQA: The MedQA dataset encompasses US Medical License Exam (USMLE) style questions obtained from the National Medical Board Examination in the USA, featuring four or five possible answer choices. With a development set of 11,450 questions and a test set of 1,273 questions, we utilized the 10,178 questions training portion of the dataset.

2.4 Fine-tuning Process

We did full supervised fine-tuning of the Palmyra-Med models using the following protocol. The Palmyra-Med-20b model was fine-tuned for 3 epochs, while the Palmyra-Med-40b model was fine-tuned for 2 epochs. We employed the AdamW optimizer (Kingma and Ba 2017) with $\beta 1 = 0.9$, $\beta 2 = 0.99$, and a weight decay of 0.0. A WarmupDecayLR learning rate scheduler was used with a learning rate of $2e-5$ and 100 warmup steps. The batch size was set to 256 examples, accompanied by a gradient accumulation step of 8. Both models were trained in bf16 precision using 8x80GB A100 GPUs. For distributed training, we incorporated the DeepSpeed library (microsoft 2023). One

notable adjustment needed to fit the fine-tuning of the 40b model on 8 GPUs involved enabling gradient checkpointing (Chen et al. 2016).

Hyperparameters	Palmyra-Med-20b	Palmyra-Med-40b
Epochs	3	2
Batch size	256	256
Learning rate	2e-5	2e-5
Warmup steps	100	100
Gradient Checkpointing	False	True
Precision	bf16	bf16
Optimizer	AdamW	AdamW
Learning Rate Scheduler	WarmupDecayLR	WarmupDecayLR

Table 2: The training hyper-parameters for both models.

2.5 Model Evaluation

To evaluate our models, which include both base and fine-tuned variants, we considered two datasets that address distinct aspects of medical knowledge and reasoning. Firstly, the MedQA dataset assesses professional medical knowledge by featuring medical exam questions. Secondly, the PubMedQA dataset necessitates medical research comprehension skills, as it includes questions based on provided PubMed abstracts. Utilizing these diverse datasets ensures a comprehensive evaluation of the models across a range of medical contexts.

PubMedQA: The PubMedQA dataset features 1k expert-labeled question-answer pairs, where the task involves producing a yes/no/maybe multiple-choice answer given a question combined with a PubMed abstract as context. While the MedQA dataset revolves around an open-domain question-answering task, the PubMedQA task is closed-domain, requiring answer inference from the supporting PubMed abstract context. We used 500 test samples for evaluation, performing a five-shot evaluation for base models and a zero-shot evaluation for the fine-tuned models.

MedQA: The MedQA dataset comprises USMLE-style multiple-choice questions with four options. We used the test split of the dataset for evaluation, consisting of 1,273 questions. A five-shot evaluation was conducted for all models.

The evaluation strategy adhered to the standard approach used in the Med-PALM 2 paper, except for the inclusion of the MedMCQA dataset. The evaluation datasets are summarized in Table 3.

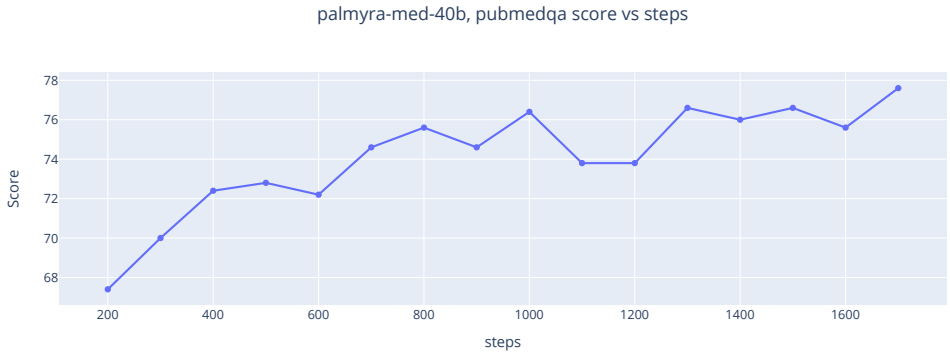
Dataset	Description	Count
PubMedQA	Closed-domain question-answering with PubMed abstract context	500
MedQA	General medical knowledge in USMLE	1,273

Table 3: Summary of Evaluation dataset.

3 Results

In this section, we present the findings of our experiments, beginning with the evaluation outcomes of the fine-tuned models and followed by a discussion of the base models' performance on each of the evaluation datasets. Additionally, we report the progressive improvement of the Palmyra-Med-40b model throughout the training process on the PubMedQA dataset. A summary of the results can be found in Table 4.

PubMedQA On the PubMedQA dataset, Palmyra-Med-20b achieved an accuracy of 75.6%, while Palmyra-Med-40b attained an accuracy of 81.1%. The fine-tuned models surpassed the base models in terms of accuracy, exhibiting enhanced performance compared to models pretrained specifically



for the medical domain. Moreover, both fine-tuned models outperformed the GPT4 model evaluated in a few-shot setting for this dataset.

Throughout the entire training process of the Palmyra-Med-40b model, we analyzed checkpoints and observed continuous improvements in accuracy on the PubMedQA dataset. The base model, without any fine-tuning, achieved a score of 64.8%. The model displayed incremental learning, reaching an initial accuracy of 67.4% at the first checkpoint, and ultimately achieving a 81.1% accuracy by the end of the training process, indicating effective learning from the fine-tuning procedure.

MedQA On the MedQA dataset, the Palmyra-20b model scored 31.2% in a five-shot setting, and the Palmyra-40b model scored 42.8% in a zero-shot setting. We then applied task tuning to MedQA with the aim of enhancing the models' proficiency in the specific domain. Task tuning imparts domain-specific information, allowing the models to better adapt to the target subject matter. Post task tuning, Palmyra-Med-20b achieved an accuracy of 44.6%, while Palmyra-Med-40b reached 72.4%. The evaluation of the models after task tuning demonstrated a significant improvement in their performance while requiring minimal additional training computation.

Model	**PubMedQA**	**MedQA**
Palmyra-20b	49.8	31.2
Palmyra-40b	64.8	43.1
Palmyra-Med-20b	75.6	44.6
Palmyra-Med-40b	81.1	72.4

Table 4: Overall performance of the base models compared to finetuned models on PubMedQA and MedQA dataset.

4 Conclusion and Future Work

In this study, we employed two powerful large language models, Palmyra-20b and Palmyra-40b. We performed instruction-based fine-tuning on both models using a custom-curated medical dataset comprising 200,000 medical examples. After fine-tuning, we referred to these models as Palmyra-Med-20b and Palmyra-Med-40b, respectively. Our evaluation demonstrates that the fine-tuned models outperform both their respective base models and existing LLMs pretrained on domain-specific knowledge. Moreover, our fine-tuned models exhibit performance comparable to the GPT4 model evaluated in a few-shot setting on the PubMedQA dataset. We also observed that utilizing larger LLMs with more parameters and advanced architectures can result in better overall performance following instruction-based fine-tuning.

We believe that the performance of both models can be further enhanced by continuing their pre-training on cleaned medical datasets, such as PMC papers and PubMed abstracts, before applying the same instruction fine-tuning procedure employed in this study. Furthermore, incorporating

MedMCQA (Pal, Umapathi, and Sankarasubbu 2022) question-answer pairs into the custom medical dataset, alongside the existing datasets and with the appropriate proportion, may yield additional improvements in the models' performance.

References

OpenAI (2023). *GPT-4 Technical Report*. arXiv: 2303.08774 [cs.CL].

Anil, Rohan et al. (2023). *PaLM 2 Technical Report*. arXiv: 2305.10403 [cs.CL].

Jin, Qiao et al. (2019). *PubMedQA: A Dataset for Biomedical Research Question Answering*. arXiv: 1909.06146 [cs.CL].

Zhang, Xiao et al. (2018). *Medical Exam Question Answering with Large-scale Reading Comprehension*. arXiv: 1802.10279 [cs.CL].

Gao, Leo et al. (2020). *The Pile: An 800GB Dataset of Diverse Text for Language Modeling*. arXiv: 2101.00027 [cs.CL].

Radford, Alec et al. (2019). "Language Models are Unsupervised Multitask Learners". In.

Shoeybi, Mohammad et al. (2020). *Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism*. arXiv: 1909.08053 [cs.CL].

Penedo, Guilherme et al. (2023). *The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only*. arXiv: 2306.01116 [cs.CL].

Dao, Tri et al. (2022). *FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness*. arXiv: 2205.14135 [cs.LG].

Shazeer, Noam (2019). *Fast Transformer Decoding: One Write-Head is All You Need*. arXiv: 1911.02150 [cs.NE].

Brown, Tom B. et al. (2020). *Language Models are Few-Shot Learners*. arXiv: 2005.14165 [cs.CL].

Wei, Jason et al. (2023). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. arXiv: 2201.11903 [cs.CL].

Taori, Rohan et al. (2023). *Stanford Alpaca: An Instruction-following LLaMA model*. https://github.com/tatsu-lab/stanford_alpaca.

Wang, Yizhong et al. (2023). *Self-Instruct: Aligning Language Models with Self-Generated Instructions*. arXiv: 2212.10560 [cs.CL].

Touvron, Hugo et al. (2023). *LLaMA: Open and Efficient Foundation Language Models*. arXiv: 2302.13971 [cs.CL].

Chiang, Wei-Lin et al. (2023). *Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality*. URL: https://lmsys.org/blog/2023-03-30-vicuna/.

Noa (2023). *ShareGPT*. URL: https://sharegpt.com/.

Kingma, Diederik P. and Jimmy Ba (2017). *Adam: A Method for Stochastic Optimization*. arXiv: 1412.6980 [cs.LG].

microsoft (2023). *DeepSpeed*. URL: https://github.com/microsoft/DeepSpeed.

Chen, Tianqi et al. (2016). *Training Deep Nets with Sublinear Memory Cost*. arXiv: 1604.06174 [cs.LG].

Pal, Ankit, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu (2022). *MedMCQA : A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering*. arXiv: 2203.14371 [cs.CL].