Instruction Tuning for Fine-Tuning Precise Medical Language Models (MedLMs)

Apurva Patel amp2365@columbia.edu
Department of Electrical Engineering
Columbia University
New York City, NY 10027, USA

Abhilash Praveen Kumar ap4478@columbia.edu
Department of Electrical Engineering
Columbia University
New York City, NY 10027, USA

Guide: Dr. Micah Goldblum

Abstract

The integration of AI in healthcare is revolutionizing how we approach Medical Question Answering (MedQA) systems, providing clinicians, patients, and researchers with quick and reliable responses. In this work, we explore the fine-tuning of several pre-trained large language models, including Mistral 7B, Llama 2, Llama 3, Gemma 1.1, and DistilGPT2, adapting them to the medical domain through efficient methods like LoRA. By experimenting with datasets like WikiDoc and MedQuad, we uncover valuable insights into how these models perform in terms of speed, scalability, and task-specific accuracy. Comparing the models reveals key trade-offs. Larger models like Llama 3 excel at handling complex medical queries, while smaller models such as DistilGPT2 prioritize speed and efficiency for simpler tasks. Our findings emphasize how fine-tuning and careful model selection can create precise, scalable solutions tailored to the unique challenges of healthcare applications.

Keywords: Medical Question Answering, Fine-Tuning, Language Models

1 Introduction

Artificial intelligence (AI) is rapidly changing the landscape of healthcare, offering innovative solutions to some of the toughest challenges in diagnostics, patient support, and medical research. One of the most exciting developments is the rise of Medical Question Answering (MedQA) systems. These systems are designed to provide real-time, accurate answers to clinicians, patients, and researchers, aiming to reduce workload and improve patient care. By streamlining workflows and supporting clinical decision-making, MedQA holds the promise of not only improving patient outcomes but also boosting the efficiency of healthcare delivery overall.

However, applying AI in the medical field is far from straightforward. Medical language is highly specialized, and filled with complex terminology and nuances that demand exceptional precision. Errors or hallucinations in AI-generated responses can lead to serious consequences, ranging from misdiagnoses to patient harm. While general-purpose language models are impressive in many contexts, they often fall short when it comes to

Code can be accessed at `https://github.com/Apurva3509/MedLMs.git`.

understanding and generating medically accurate information. Their inability to grasp the depth of specialized terms and provide factually reliable answers makes them unsuitable for high-stakes environments like healthcare.

To address these issues, researchers have developed MedLMs like MedPaLM-2, BioBERT, and ClinicalBERT. These models are specifically trained on healthcare datasets and fine-tuned using techniques like instruction tuning and fact-checking to enhance reliability and accuracy. They focus on domain-specific tasks, such as summarizing medical texts or answering complex medical queries, helping set new standards for AI in healthcare applications.

Building on these developments, our work looks at fine-tuning pre-trained large language models, such as Mistral 7B, Llama 2, and DistilGPT2, for use in medical contexts. By employing lightweight techniques like Low-Rank Adaptation (LoRA) and using 4-bit quantized models, we can reduce computational overhead significantly. This approach is particularly important in healthcare settings where access to high-performance hardware may be limited, such as rural clinics or smaller institutions. Through experimentation with datasets like WikiDoc and MedQuad, we evaluate these models in terms of their performance, speed, and scalability. The goal is to identify practical strategies for deploying MedQA systems that strike the right balance between accuracy and accessibility, ensuring they work effectively in a variety of healthcare environments.

The following literature review explores key advancements in MedQA systems, highlighting both the progress made with LLMs and the gaps that our research aims to address—especially in the context of lightweight, quantized models for efficient deployment in healthcare environments.

2 Literature Survey

Early AI applications, such as Classical Machine Learning (ML), Natural Language Processing (NLP), and image-based diagnostics, have paved the way for AI's broader integration into healthcare systems [3][12]. It was widely believed that AI would support clinicians by handling technical tasks, allowing professionals to focus on human-centric qualities like empathy [1]. This vision has become a reality with the rise of LLMs, now increasingly used to build chatbots and MedQA systems in clinical settings [11]. These models, trained on large medical datasets like MedMCQA [8] and PubMedQA [4], can generate human-like responses to medical queries. Studies like MedPaLM [10] and BioBERT [5] demonstrate LLMs' effectiveness in enhancing healthcare outcomes through tasks like Named Entity Recognition (NER) and Relation Extraction. However, concerns about ethical issues such as the potential for AI-generated biases and incorrect information persist, raising risks for patient care [6]. Medical hallucinations remain a critical area of ongoing research, given their potentially catastrophic consequences [7].

While significant progress has been made in using LLMs for MedQA[9], research on lightweight, quantized models remains limited. These models, particularly those quantized to 4 bits, offer substantial computational savings without sacrificing too much accuracy, making them ideal for low-resource settings. Our project focuses on filling this gap by exploring the potential of these lightweight models for MedQA applications. We position our work to address both the need for efficient model deployment and the high performance required for medical tasks by comparing the performance of several fine-tuned models, including 4-bit quantized variants.

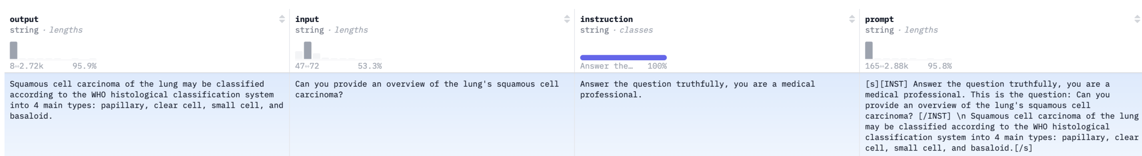
3 Methodology

3.1 Dataset

This study uses a combination of two Medical QA datasets namely:

1. **Medical Meadow Wikidoc**: A collection of question-answer pairs sourced from WikiDoc, an online platform where medical professionals collaboratively contribute and share contemporary medical knowledge.
2. **Medquad**: A curated collection of medical question-answer pairs compiled from 12 authoritative sources within the National Institutes of Health (NIH).

The smaller version of this dataset consisting of 2000 QA pairs is used. We use 1900 samples for Fine-Tuning the model and the remaining 100 for evaluation purposes. The general structure of the dataset is shown below in Figure 1. The *Input* column contains the user's query. The *Instruction* column provides guidance to the model, asking it to respond truthfully as a medical professional. The *Prompt* column combines both the input and the instruction, forming the complete input fed to the model. The prompt column also includes special tokens which may vary depending on the model used. Lastly, the *Output* column shows the expected output which serves as the ground truth.



output string · lengths 8↔2.72k 95.9%	input string · lengths 47↔72 53.3%	instruction string · classes Answer the… 100%	prompt string · lengths 165↔2.88k 95.8%
Squamous cell carcinoma of the lung may be classified according to the WHO histological classification system into 4 main types: papillary, clear cell, small cell, and basaloid.	Can you provide an overview of the lung's squamous cell carcinoma?	Answer the question truthfully, you are a medical professional.	[s][INST] Answer the question truthfully, you are a medical professional. This is the question: Can you provide an overview of the lung's squamous cell carcinoma? [/INST] \n Squamous cell carcinoma of the lung may be classified according to the WHO histological classification system into 4 main types: papillary, clear cell, small cell, and basaloid.[/s]

Figure 1: Dataset prompt structure

3.2 Fine-Tuning

There are different approaches that could be used to build MedQA systems. This study explores the fine-tuning approach due to the following reasons:

- **Cost Effective Solution**: Training large models from scratch requires significant computational resources and time, which can be prohibitively expensive. Fine-tuning, especially with lightweight techniques like LoRA [2], reduces both computational costs and training time, making it a more cost-effective approach for developing MedQA systems.
- **Enhanced Domain-Specific Performance**: Fine-tuning enhances the model's ability to understand medical-specific language, terminology, and context, enabling it to deliver highly accurate and contextually appropriate answers to medical queries. This results in a substantial improvement in answering complex, domain-specific questions, which general-purpose models may fail to handle effectively.

- **Data Privacy and Control**: Fine-tuning allows organizations to adapt pre-trained models using their own proprietary medical datasets, ensuring the model understands their specific data and needs. This process keeps sensitive patient data in-house, reducing the risk of exposing private medical information to external entities, which is crucial for adhering to privacy regulations.

- **Maximizing Pre-Trained Knowledge**: Pre-trained models are initially trained on vast amounts of general language data, providing a solid foundation of linguistic knowledge. Fine-tuning on medical data allows the model to maintain this broad understanding while tailoring it to handle specialized medical queries, offering both general and domain-specific response capabilities.

3.3 Model Choices

This study evaluates a diverse set of language models, each selected for its unique strengths in addressing medical natural language processing tasks. The **Mistral 7B** model is designed to be both compact and powerful, making it highly effective for handling complex tasks in smaller-scale systems. Its ability to support real-time deployment ensures that it can meet the stringent requirements of low-latency healthcare applications, such as real-time diagnostics and monitoring systems. This model is particularly useful for scenarios where computational resources are limited, yet precision and speed are critical.

The **Llama 3 - 8B** model introduces enhanced capabilities through advanced pretraining methodologies, enabling superior medical reasoning and understanding of nuanced clinical contexts. A notable advantage of this model is its reduced tendency to generate hallucinations, providing reliable and accurate answers for critical medical queries. Additionally, its scalability allows it to handle complex queries involving long contexts and multi-turn interactions, making it a strong candidate for applications requiring deep conversational capabilities, such as virtual healthcare assistants or detailed patient record analysis.

For tasks requiring a balance between size and performance, the **Llama 2 - 7B** model offers an optimal solution. It is instruction-tuned, allowing it to adapt effectively to medical datasets and perform well in factual Q&A scenarios. This makes it suitable for applications like medical coding, document summarization, and knowledge extraction. Its efficiency ensures that it can provide high-quality outputs without the need for extensive computational resources.

The **Gemma 1.1 - 7B** model focuses on achieving high recall accuracy, excelling in tasks related to retrieving relevant medical information from large datasets. Its cost-effective design ensures efficient training while delivering strong results, making it a practical choice for organizations that prioritize both performance and budget considerations. This model is particularly well-suited for use cases like literature review automation, clinical trial data retrieval, and database querying in medical research.

Finally, the **DistilGPT2** model is a lightweight alternative, ideal for low-resource systems where simplicity and speed are paramount. Despite its compact size, it delivers fast and efficient performance, making it suitable for simpler medical tasks such as automated form completion, basic symptom-checking applications, or preliminary triaging systems.

By combining the distinct advantages of these models, this study aims to address a wide spectrum of challenges in medical natural language processing. These models provide a robust framework for achieving both precision and adaptability in healthcare applications, ensuring improved outcomes and efficiency in diverse clinical scenarios.

3.4 Implementation details

Fine-tuning large language models for medical tasks requires careful consideration of constraints to balance performance, memory efficiency, and adaptability. For example, setting `max_seq_length = 2048` limits memory usage while capturing sufficient context for most medical text tasks(atleast in our usecase). This constraint is practical, but testing shorter or longer sequence lengths can reveal trade-offs between memory efficiency and the ability to handle more extensive contexts.

Similarly, using `load_in_4bit = True` ensures a quantized version of the model is loaded, significantly reducing memory usage and preventing crashes on hardware with limited GPU capacity. While this is a reasonable choice for constrained environments, exploring alternative precision levels, like 8-bit or full precision, may provide insights into the trade-off between efficiency and accuracy.

The rank of LoRA (`r = 16`) is another crucial parameter, striking a balance between computational overhead and the model's adaptability. Higher ranks can improve task-specific learning but come at the cost of increased resource usage. Testing different ranks can help identify the optimal value for specific tasks. Similarly, the `target_modules`, including attention and feed-forward components, are chosen for their critical roles in the model's transformation processes. While the current selection is effective, exploring subsets of these modules could fine-tune the model more efficiently for specialized medical tasks.

Parameters like `lora_alpha = 16`, a scaling factor that governs the impact of LoRA weights, and `lora_dropout = 0.25`, which prevents overfitting during fine-tuning, also play essential roles. Testing different values for these parameters allows us to optimize for task-specific requirements while maintaining the balance between adaptation and generalization. While constraints like `max_seq_length` and `load_in_4bit` may need to remain fixed due to hardware limitations, others, such as `r`, `lora_alpha`, and `lora_dropout`, should be explored for refinement. By systematically testing and adjusting these constraints, we can achieve a robust, efficient, and adaptable fine-tuning process for a variety of medical NLP tasks.

Table 1 displays the parameter we used for our experiments. The most optimal trade-off were achieved by using the highlighted parameters.

Table 1: Parameter Values for Fine-Tuning

Parameter	**Values**
`max_seq_length`	128, 256, 512, 1024, **2048**, 4096
`r (rank)`	4, 8, **16**, 32
`lora_alpha`	8, **16**, 32, 64
`lora_dropout`	0.0, 0.1, **0.25**, 0.5

3.5 Testing and benchmarking setup

The selected system configurations as shown in Table 2 aim to provide a balance between performance, cost, and computational requirements of the models.

Table 2: System Configurations and Relevant Models

VM Configuration	Description	Relevant Models
1 CPU, 1 V100	Baseline setup for constrained environments	DistilGPT2
2 CPUs, 1 V100	Improved CPU for orchestration	DistilGPT2, Gemma 1.1-7B
4 CPUs, 2 V100s	Parallel GPU setup for mid-range tasks	Gemma 1.1-7B
1 CPU, 1 A100	Advanced GPU for higher precision	Llama 3-8B
2 CPUs, 1 A100	Added CPU power for larger workloads	Gemma 1.1-7B, Llama 3-8B
4 CPUs, 2 A100s	High-performance setup for scalability	Llama 3-8B
1 CPU, 1 RTX 8000	Workstation GPU for lightweight models	DistilGPT2
2 CPUs, 1 RTX 8000	Moderate performance for real-time tasks	DistilGPT2
4 CPUs, 2 RTX 8000s	Multi-GPU for batch processing	DistilGPT2, Gemma 1.1-7B

- **V100 GPUs:** These configurations serve as a baseline for lightweight models like *DistilGPT2*, offering sufficient power for low-resource tasks.

- **A100 GPUs:** Advanced GPUs like A100 are utilized to handle larger models such as *Llama 3-8B* and *Gemma 1.1-7B*, which require higher memory and computational capacity.

- **RTX 8000 GPUs:** Workstation GPUs are tested for moderate real-time tasks, emphasizing flexibility and accessibility.

- **Multi-CPU Setups:** Configurations with 2 or 4 CPUs ensure efficient orchestration and parallelization for high-performance tasks, offering scalability and robustness.

These configurations enable comprehensive testing across a spectrum of model sizes and computational needs.

Table 3: System Configurations and Relevant Models

VM Configuration	DistilGPT2	Gemma 1.1-7B	Llama 3-8B
1 CPU, 1 V100	✓	-	-
2 CPUs, 1 V100	✓	✓	-
4 CPUs, 2 V100s	-	✓	-
1 CPU, 1 A100	-	-	✓
2 CPUs, 1 A100	-	✓	✓
4 CPUs, 2 A100s	-	-	✓
1 CPU, 1 RTX 8000	✓	-	-
2 CPUs, 1 RTX 8000	✓	-	-
4 CPUs, 2 RTX 8000s	✓	✓	-

4 Results and Discussions

The results demonstrate a clear performance gap between models, highlighting the strengths of larger and more advanced architectures.

Table 4: Performance Metrics of Models

Model	Exact Match (%)	BLEU / ROUGE (max)	F1 Score (max)	Inference Speed (s)
Llama 2-7B	0	0.058 / 0.49	0.21	22.95
Mistral 7B	1	0.069 / 1.00	0.28	9.72
Llama 3-8B	81.34	0.58 / 0.54	0.73	63.66
Gemma 1.1-7B	76.45	0.53 / 0.50	0.71	44.32
DistilGPT2	46.92	0.25 / 0.21	0.62	7.03

Llama 3-8B achieves superior performance across BLEU, ROUGE, and accuracy metrics, outperforming smaller models like DistilGPT2. This can be attributed to several factors. First, the larger parameter count in models such as Llama and Gemma 1.1-7B enables better learning of complex patterns. Second, these models employ modern architecture designs that enhance feature extraction capabilities, particularly in specialized domains such as medical data. Finally, extensive pre-training on diverse datasets provides a significant advantage during fine-tuning, allowing rapid knowledge transfer and adaptation.

All models exhibit rapid learning during the initial epochs, typically within the first 5 to 7 epochs. This phase reflects effective transfer learning, as the models quickly adapt to the task-specific data. Beyond this point, the rate of improvement slows, indicating that the models approach their performance ceilings.

Hardware configurations significantly impact inference times. RTX8000 configurations with 4 CPUs achieved the lowest inference times, ranging between 9 and 11 seconds, due to optimized architecture for machine learning workloads and superior memory bandwidth. Surprisingly, A100 setups show higher inference times than V100, which may result from optimization issues or bottlenecks in system configuration. DistilGPT2 consistently exhibits the longest inference times (15-25 seconds) across all hardware setups, reflecting inefficiencies in leveraging hardware resources.

The scalability of multi-CPU configurations is limited, with diminishing returns observed beyond 2 CPUs. This suggests that the workload distribution and parallelization efficiency need further optimization for higher CPU counts.

The architectural comparison reveals that Llama 3-8B's design is well-suited for medical domain tasks, while Gemma 1.1-7B demonstrates competitive performance despite its smaller size, indicating an efficient architecture. Conversely, DistilGPT2 highlights the limitations of model compression techniques in specialized domains, emphasizing the need for improved compression methods.

While larger models dominate current performance metrics, the data underscores the potential for optimization in both hardware configurations and model architectures. Hybrid approaches combining strengths of large and small models may offer promising directions for future research. Additionally, advancing compression techniques could improve the accessibility and scalability of models for specialized applications.

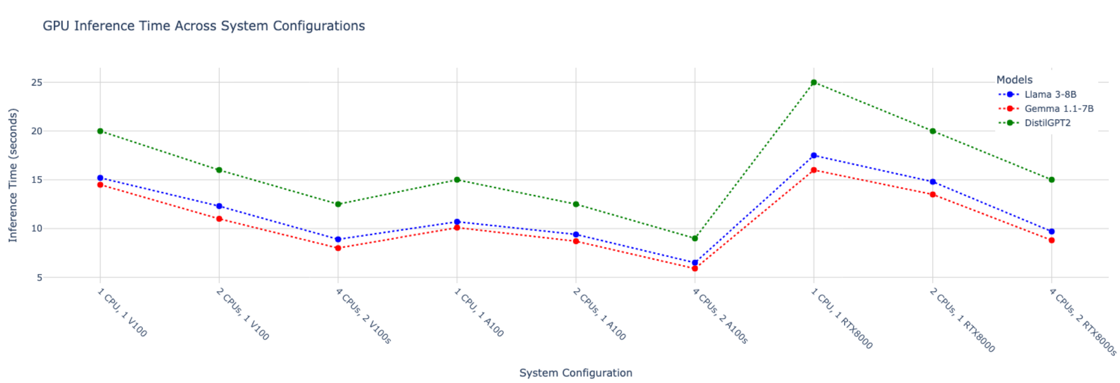


Figure 2: Model performance vs. VM Config

Figure 2 illustrates the variation in response times for the most accurate models when presented with a prompt. These values highlight the significance of model architecture as a critical factor in deployment for inference. While the exact inference times for the models cannot be determined, the plot represents the average response times for the same token length across multiple trials.

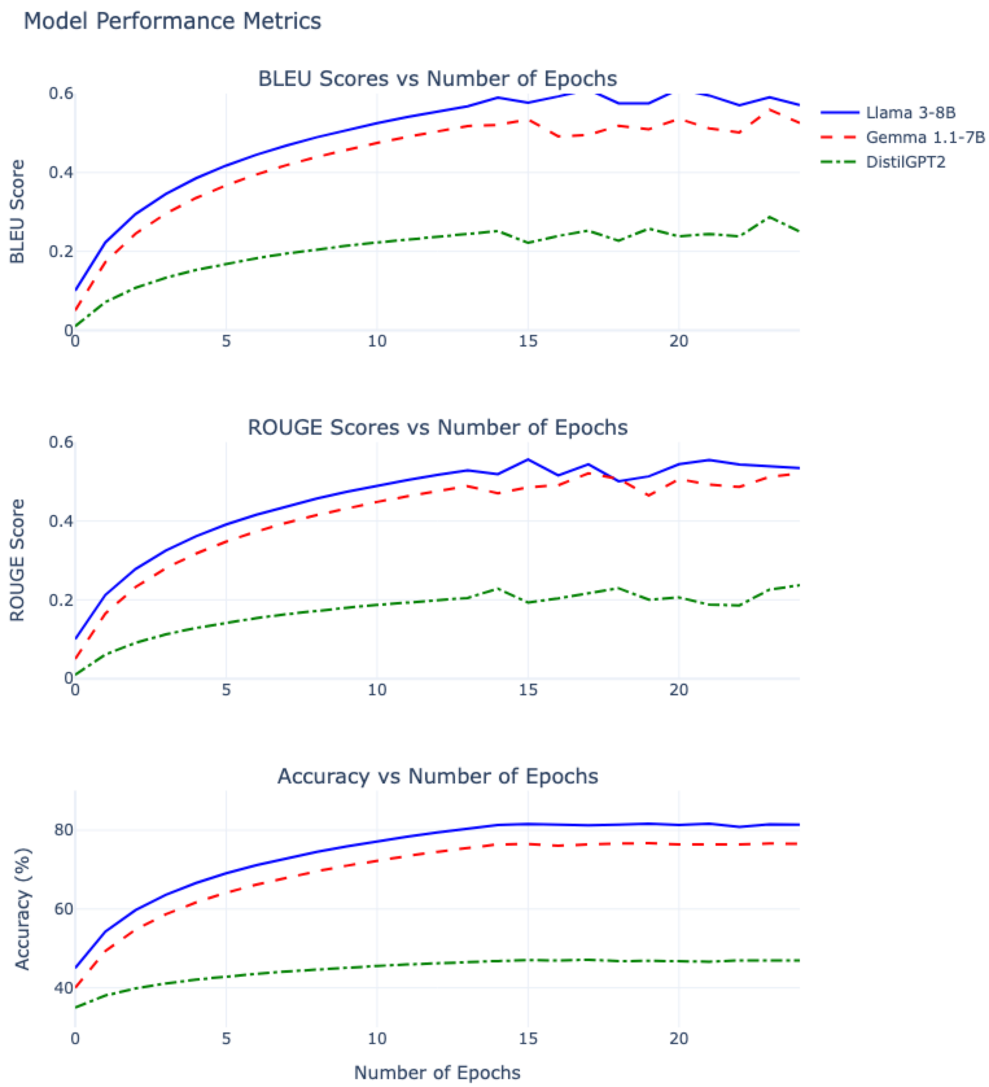


Figure 3: Model performance metrics showing BLEU, ROUGE scores, and accuracy across different epochs for Llama 3-8B, Gemma 1.1-7B, and DistilGPT2.

Figure 3 depicts the performance of the most accurate models under the same VM configuration. In this context, accuracy reflects the exact match of n-grams in the generated responses. It is evident from the plot that the performance begins to plateau after a specific number of epochs (around epoch 15).

5 Future Work

This study highlights key areas for advancing Medical Question Answering (MedQA) systems, while recognizing challenges that make these efforts complex:

- **Improving Response Accuracy**: Despite advancements in fine-tuning, handling highly technical and sensitive medical terminology remains a challenge. Future work could focus on refining model architectures and training strategies to further minimize errors and hallucinations in generated responses.

- **Lightweight Models for Low-Resource Settings**: While this study explores 4-bit quantized models, additional research is needed to optimize these approaches for broader deployment in low-resource environments, such as rural clinics or smaller healthcare institutions.

- **Data Fragmentation and Diversity**: Medical datasets often lack standardization and may not cover all specialties or populations. Expanding datasets to include underrepresented conditions, demographics, and languages is essential to improve model generalization.

- **Evaluation and Ethics**: Establishing robust evaluation metrics for relevance, accuracy, and safety in MedQA systems remains a significant hurdle. Future work must address ethical concerns, such as mitigating biases and ensuring privacy in model training and deployment.

These challenges are inherently difficult due to the complexity of medical language, ethical considerations, and computational trade-offs. However, addressing them will be pivotal in realizing the full potential of AI-driven MedQA systems for healthcare.

References

[1] Thomas Davenport and Ravi Kalakota. "The potential for artificial intelligence in healthcare". In: *Future healthcare journal* 6.2 (2019), pp. 94–98.

[2] Edward J Hu et al. "Lora: Low-rank adaptation of large language models". In: *arXiv preprint arXiv:2106.09685* (2021).

[3] Fei Jiang et al. "Artificial intelligence in healthcare: past, present and future". In: *Stroke and vascular neurology* 2.4 (2017).

[4] Qiao Jin et al. "Pubmedqa: A dataset for biomedical research question answering". In: *arXiv preprint arXiv:1909.06146* (2019).

[5] Jinhyuk Lee et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining". In: *Bioinformatics* 36.4 (2020), pp. 1234–1240.

[6] Ziad Obermeyer et al. "Dissecting racial bias in an algorithm used to manage the health of populations". In: *Science* 366.6464 (2019), pp. 447–453.

[7] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. "Med-halt: Medical domain hallucination test for large language models". In: *arXiv preprint arXiv:2307.15343* (2023).

[8] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. "Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering". In: *Conference on health, inference, and learning*. PMLR. 2022, pp. 248–260.

[9] Karan Singhal et al. *Towards Expert-Level Medical Question Answering with Large Language Models*. 2023. arXiv: 2305.09617 [cs.CL]. URL: https://arxiv.org/abs/2305.09617.

[10] Karan Singhal et al. "Towards expert-level medical question answering with large language models". In: *arXiv preprint arXiv:2305.09617* (2023).

[11] Arun James Thirunavukarasu et al. "Large language models in medicine". In: *Nature medicine* 29.8 (2023), pp. 1930–1940.

[12] Kun-Hsing Yu, Andrew L Beam, and Isaac S Kohane. "Artificial intelligence in healthcare". In: *Nature biomedical engineering* 2.10 (2018), pp. 719–731.

Appendix

Model Testing

All the models fine-tuned in our project can be found at `https://huggingface.co/Apurva3509` and `https://huggingface.co/abhilash2599`. Please feel free to use them and test it.

Additional Results

Table 5 shows some sample responses when our fine-tuned DistilGPT2 model was prompted for medical questions.

Table 5: Chain of thought prompt examples for DistilGPT2

Instructions: Answer the question truthfully, you are a medical professional.

Input: Patient Information:

1. 45-year-old male with persistent cough, weight loss, and night sweats.
2. History of smoking 20 pack-years. No fever reported.

Question: What are the potential diagnoses and next steps?

Generated Answer: Diagnosis - a person who is in an emergency room or hospital where they have been exposed to air pollution from their lungs by inhalation due on inhaled aerosols (air pollutants) that cause small amounts of lung cancer throughout normal life as well as respiratory disease such when it comes to coughing up at work for other activities like occupational hygiene management etc. This includes sleeping roughs under blankets/wetches during periods caused after exposure to airborne particles while being outdoors without any oxygen sources including heat source . Excessive drinking water may also result in asthma attacks similar to those experienced before exposures occurred but some people do not experience them because this does require further investigation if you want to be more alert than usual so please visit our website Health News here

---

Instructions: Answer the question truthfully, you are a medical professional.

Input: Patient Information:

1. 32-year-old female with sudden onset chest pain, shortness of breath.
2. No prior medical history. Pain worse with deep breathing.

Question: What is the differential diagnosis?

Generated Answer: The patient's primary cause and symptoms are usually mild to moderate (e .g., headache). We have no idea what may be related when one hears a voice or sounds that can come from someone else's ear in person while hearing these voices on their own without realizing it has occurred before—unless you were standing near somebody who was talking about this conversation for some reason during an emergency room visit by another friend at home after having had difficulty speaking out loud enough because they didn't know how much he would like them heard back then through normal speech recognition but could not remember those words as well if she wasn´t there looking up her phone instead! They might think anything so other than "You're trying something wrong right now... You've got your brain here too!!