MedLMs

EECS 6694: GenAI and Modern Deep Learning

Apurva Patel (amp2365)
Abhilash Praveen Kumar (ap4478)

Motivation for MedLMs

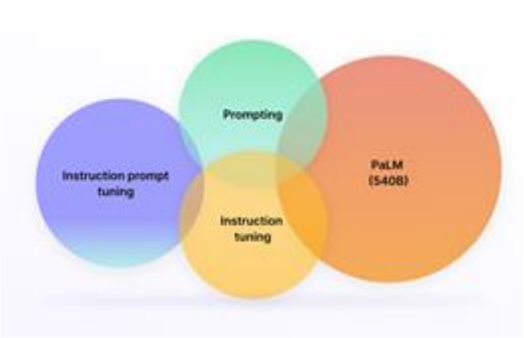
- Healthcare sector is rapidly adopting AI to assist in diagnostics, patient support, and medical research.
- Medical Question Answering (MedQA) systems can provide accurate, real-time responses to clinicians, patients, and researchers, potentially improving outcomes and reducing workload.

Existing Challenges in MedQA:

- Medical text is domain-specific, highly technical, and sensitive, requiring models to possess expert-level knowledge while ensuring reliability, safety, and ethical compliance.
- Existing general-purpose LLMs lack specialization in medical contexts and may produce hallucinated or inaccurate responses.

Existing Work

1. **MedPaLM-2:**
   - Fine-tuned from **PaLM** (Pathways Language Model), a large-scale transformer model.
   - Leveraged healthcare-specific datasets:
     - **MedQA (USMLE):** Multiple Choice Questions from US medical license exams.
     - **PubMed**: Biomedical literature for domain-specific pretraining.
     - **HealthSearchQA**: Real-world patient questions from web search.



Key techniques:

- **Instruction Tuning**:
  - Taught the model to follow structured instructions for improved task performance.
- **Patient-Friendly Summarization**:
  - Trained the model to simplify medical content while preserving accuracy for non-expert audiences.
- **Fact-Checking Mechanisms**:
  - Introduced methods to reduce hallucinations and ensure responses align with verified medical knowledge.

Performance:

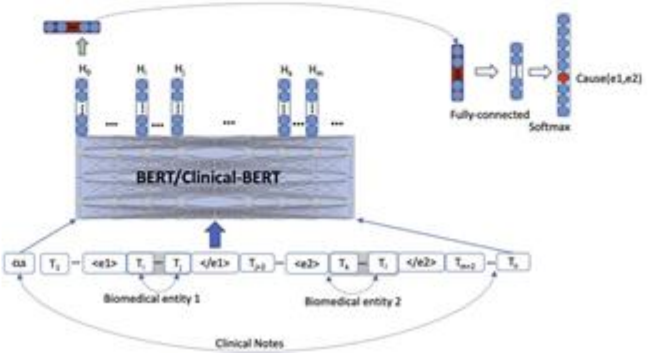
- Reduced hallucination rates compared to general-purpose LLMs.

Source: arXiv:2305.09617

Existing Work

2. Clinical BERT:

- Fine-tuned BERT on MIMIC-III clinical notes, improving performance on tasks like clinical Named Entity Recognition (NER) and question answering in medical contexts.



It can be used for (in decreasing order of precision):

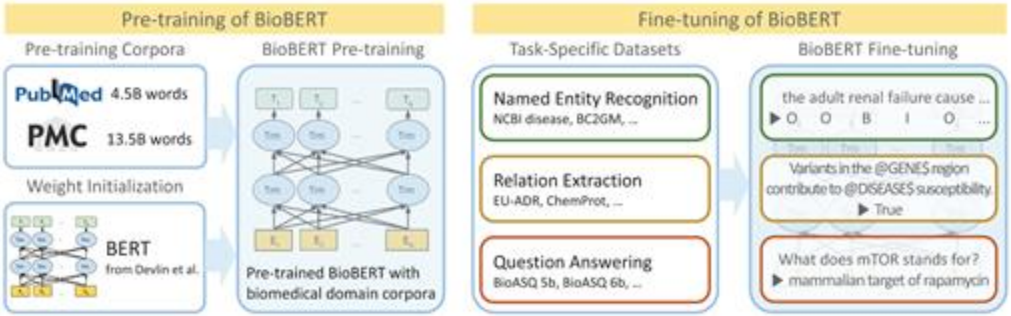
- Named entity recognition (e.g., extracting symptoms, medications, diagnoses).
- Relation extraction (e.g., identifying relationships between symptoms and diagnoses).
- Text classification (e.g., categorizing medical notes or flagging adverse drug events).
- Question answering (e.g., answering patient-related queries based on clinical text).

Source: arXiv:1904.03323

Existing Work

3. **BioBERT:**

- Adapted BERT for biomedical text mining by pretraining on PubMed and PMC articles.
- Demonstrated improvements in Named Entity Recognition (NER) and relation extraction tasks.



3. **PubMedGPT:**

- Trained specifically on PubMed articles.
- Demonstrated superior performance in domain-specific tasks compared to general GPT-2 models.
- Highlights limitations in generalization outside the PubMed dataset.
- The model struggles with new information outside its training data.

Source: arXiv:1901.08746

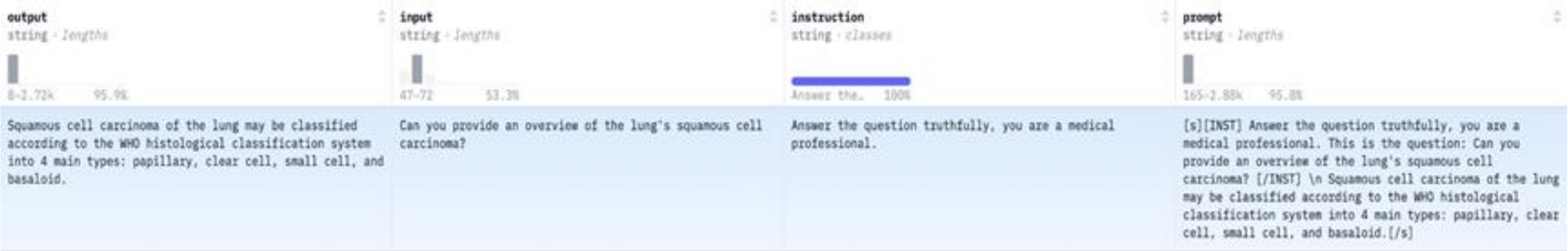
COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

Why Finetune existing models?

- **Improved Accuracy:**
  - Fine-tuning allows LLM to grasp nuances of the chosen domain - leads to more accurate and relevant responses to queries within that field.
- **Cost Savings:**
  - Training models from scratch is expensive and not cost effective.
  - Lightweight fine-tuning techniques like LoRA used.
- **Privacy and Security:**
  - Allows organizations to adapt pre-trained models with their own datasets.
  - Keeps sensitive information in-house and minimizes exposure to external threats.
- **Leveraging Pre-Trained Linguistic Knowledge:**
  - Fine-tuning taps into a model's broad linguistic knowledge, enabling it to handle both specialized medical queries and general questions effectively.

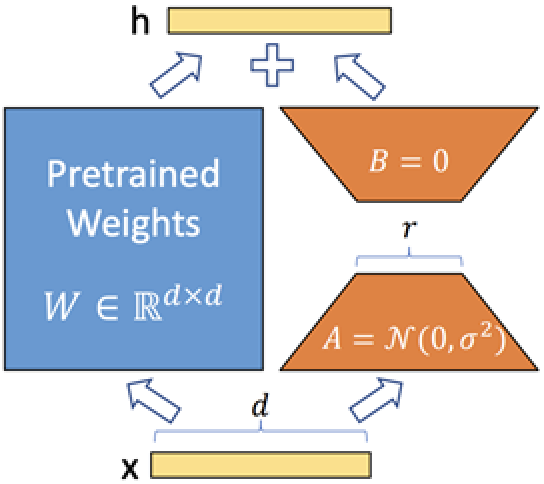
Methodology

- **Dataset:**
  - Combined medical dataset for instruction fine tuning
    - **Medical meadow wikidoc** - QA pairs sourced from WikiDoc - online platform where medical professionals collaboratively contribute and share contemporary medical knowledge.
    - **Medquad** - collection of medical QA pairs compiled from 12 authoritative sources within the National Institutes of Health (NIH).
  - 1900 data points used for fine-tuning and 100 samples used for evaluation.



Methodology

- **Fine-Tuning Techniques:**
  - **Problem** - Standard methods require modifying and storing all parameters of the model - **computationally expensive** and **memory intensive**.
  - **LoRA** - A parameter-efficient fine-tuning technique that **freezes the pre trained model parameters**
    - introduces **low-rank matrices** to learn task-specific adaptations.
  - Allows training on limited resources without compromising model performance.



https://arxiv.org/pdf/2106.09685

Model(s) Used

Mistral 7B

- **Compact & Powerful**: Handles complex tasks in smaller-scale systems.
- **Real-Time Deployment**: Suitable for low-latency healthcare applications.

Llama 3 - 8B

- **Enhanced Capabilities**: Improved pretraining for better medical reasoning.
- **Fewer Hallucinations**: Reliable for critical, factual answers.
- **Scales Complex Queries**: Handles longer contexts and multi-turn questions.

Llama 2 - 7B

- **Efficiency**: Good balance of size and performance.
- **Instruction-tuned**: Adapts well to medical datasets and factual Q&A.

Gemma 1.1 - 7B

- **High Recall Accuracy**: Excels in retrieving relevant medical information.
- **Cost-Effective**: Efficient training with strong results.

DistilGPT2

- **Lightweight**: Ideal for low-resource systems.
- **Fast & Efficient**: Great for simpler medical tasks.

Constraints

1. `max_seq_length = 2048`
    - *We limit this to keep low memory usage.*

1. `load_in_4bit = True`
    - *Done to load in quantized version and stop system crashing due to model size.*

1. `r = 16 #rank`
    - *Rank of the low-rank decomposition used in LoRA (Low-Rank Adaptation). A rank of 16 provides enough capacity to adapt the model effectively to the new task without overloading memory or computation.*

1. `target_modules = ["q_proj", "k_proj", "v_proj", "o_proj", "gate_proj", "up_proj", "down_proj",]`
    - *Targeted modules include attention and feed-forward components:*
        - *q_proj, k_proj, v_proj: Query, key, and value projections in the attention mechanism.*
        - *o_proj: Output projection in the attention mechanism.*
        - *gate_proj, up_proj, down_proj: Components of the feed-forward network (FFN).*

1. `lora_alpha = 16`
    - *A scaling factor for the LoRA weights, controlling their impact on the model's output. Maintains balance between original knowledge and task specific adaptation.*

1. `lora_dropout = 0.25`
    - *Dropout to the LoRA layers to prevent overfitting during fine-tuning.*

Note: Constraints are fixed across all different final fine-tuning process to the ones we found to be working for our system configuration, experiments might include different parameters

Results

Model	Exact match (%) (max)	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

Note: Inference speeds were measured using T4

Results

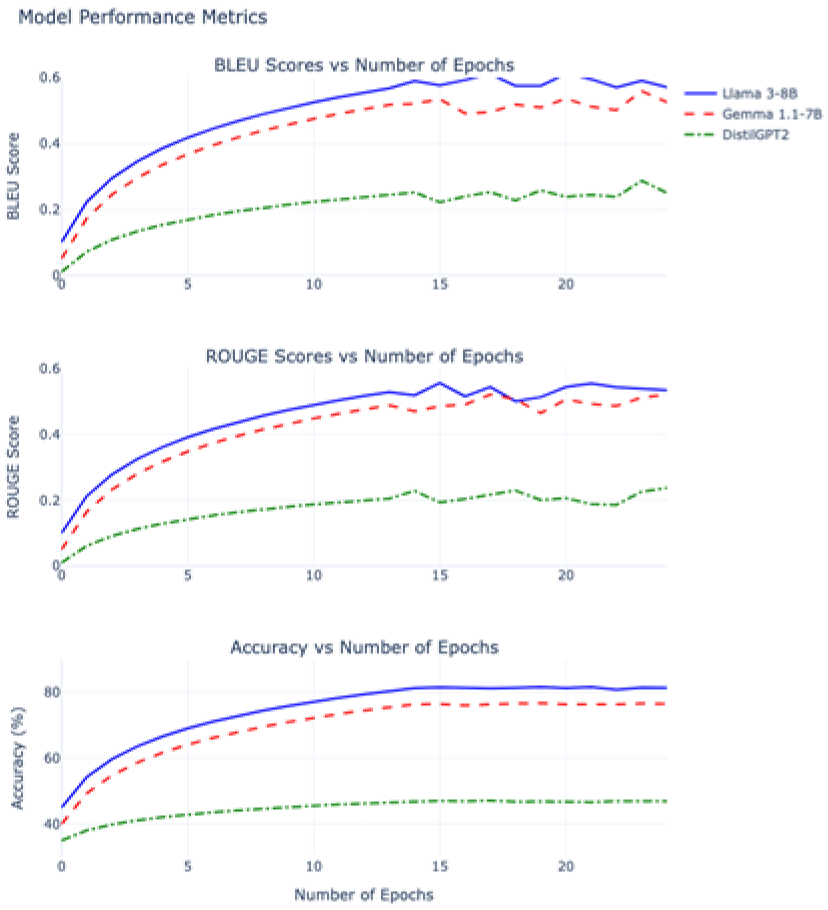
The performance gap between models is attributed to:

1. *Model capacity* - larger models (Llama, Gemma) have greater parameter counts, enabling better learning of complex patterns

2. *Architecture differences* - modern architectures in Llama and Gemma provide better feature extraction

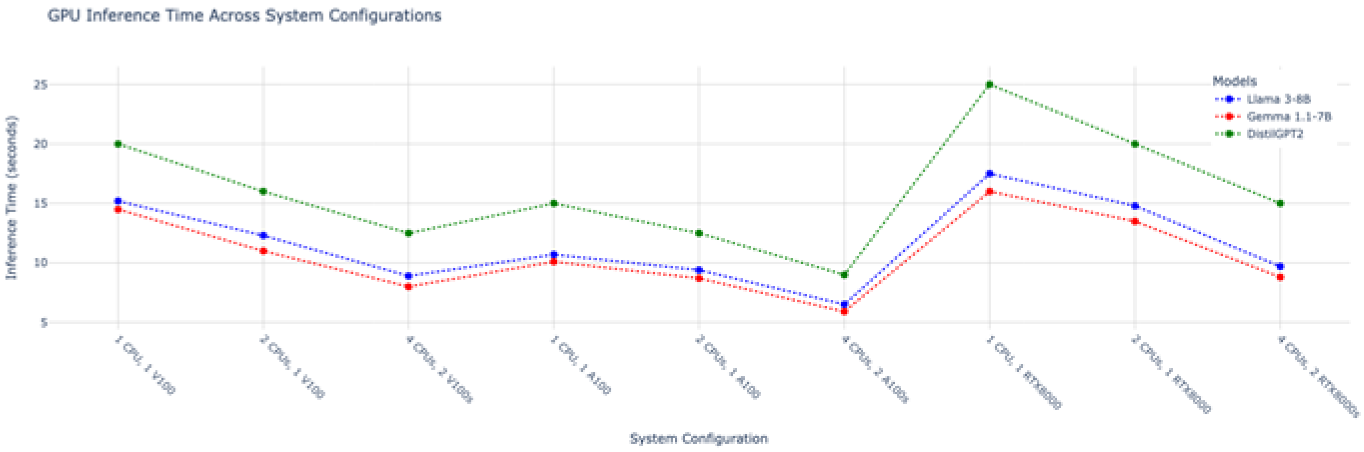
3. *Pre-training advantage* - larger models benefit from more extensive pre-training on diverse datasets

All models show rapid initial learning (~ 5-7 epochs) followed by gradual improvement, which shows effective knowledge transfer during fine-tuning.



Note: Results display only 3 models (with interesting inferences),

Results



Key Findings

- RTX8000 configurations achieve optimal performance with 4 CPUs, showing lowest inference times (9-11s)
- A100 setups unexpectedly show higher inference times than V100
- DistilGPT2 consistently requires longer inference times (15-25s) across all configurations
- Multi-CPU scaling shows diminishing returns beyond 2 CPUs

The RTX8000's superior performance likely stems from its optimized architecture for ML workloads and better memory bandwidth. The unexpected A100 performance might be due to optimization issues or system configuration bottlenecks.

Discussions

Higher ranks can capture more task-specific features but at the cost of increased memory usage.

Architecture Comparison

The performance gap between models highlights several key points:

- Llama 3-8B's superior performance suggests better architecture design for medical domain tasks
- Gemma 1.1-7B's competitive performance despite smaller size indicates efficient architecture
- DistilGPT2's limitations reflect the challenges of model compression in specialized domains

- Potential for hybrid approaches combining different model strengths
- Need for better compression techniques to improve smaller model performance

The data suggests that while larger models currently dominate performance metrics, there's significant room for optimization in both hardware configurations and model architectures for medical domain applications.

All final fine tuned models can be found here: https://huggingface.co/Apurva3509

Sample Response

```
=== distilgpt2-medical Output for Case 1 ===
Input:
Patient Information:
        [45-year-old male with persistent cough, weight loss, and night sweats.
        History of smoking 20 pack-years. No fever reported.]
        Question: [What are the potential diagnoses and next steps?]

Generated Output:


          [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]
```

Correct response

Out of context response

Incomplete response (max token length)

Thank You :)