Deep Learning Project Report

Title: Comparative Analysis of Large Language Models-LLMs in Healthcare

Introduction

Large Language Models such as DeepSeek, GPT, LLaMA, BERT and specialize biomedical LLMs like ClinicalBERT, BioBERT and the recent one Me-LLaMA are progressively used in processing and analyzing healthcare tasks. However, the comparative analysis of different LLMs in healthcare; Medical Q/A task is a topic of examination. In this project we systematically compare and assess LLMs in the medical field. The analysis is only done on five models which can help optimize the models' use in healthcare and research settings by shedding light on which ones are most appropriate for medical applications. The dataset I have used is QA having complex_CoT (complex chain of thoughts), therefore before and after fine tuning the models: DeepSeek (unsloth/DeepSeek-R1-Distill-Llama-8B) and Mistral (unsloth/mistral-7b-instruct) perform well before and after fine tuning Whereas BioBert did well on “MedQA-USML” dataset with accuracy of 51%.

Methodology

Dataset:

I have used medical reasoning Q/A dataset available on huggingface library. Publically available without any special permission. It contain almost 90k entries having columns: Questions, Complex_CoT(Chain of thoughts), and Response(Answer). This dataset is created using GPT-24o and validated by medical verifiers. I have also verify this data from my sister she is also MBBS Dr. currently doing FPSC-II training. This dataset is useful to perform complex medical reasoning e.g. diagnostic decision making or treatment planning. The dataset contains ~90K entries and only one split “train”. As shown in the Figure 1 there are three columns; Question, Chain of Thoughts (Complex_CoT), Response.

Question string · lengths 51 2.38k	Complex_CoT string · lengths 834 5.54k	Response string · lengths 4 4k
Given the symptoms of sudden weakness in the left arm and leg, recent long…	Okay, let's see what's going on here. We've got sudden weakness in the…	The specific cardiac abnormality most likely to be found in this scenario…
A 33-year-old woman is brought to the emergency department 15 minutes afte…	Okay, let's figure out what's going on here. A woman comes in with a sta…	In this scenario, the most likely anatomical structure to be injured i…
A 61-year-old woman with a long history of involuntary urine loss…	Okay, let's think about this step by step. There's a 61-year-old woman…	Cystometry in this case of stress urinary incontinence would most…
A 45-year-old man with a history of alcohol use, who has been abstinent…	Alright, let's break this down. We have a 45-year-old man here, who…	Considering the clinical presentation of sudden onset dysarthria, shufflin…
A 45-year-old man presents with symptoms including a wide-based gait…	Okay, so here's a 45-year-old guy who's experiencing some pretty…	Based on the clinical findings presented—wide-based gait, blank…
A patient with psoriasis was treated with systemic steroids, and upon…	I'm thinking about this patient with psoriasis who's been on systemic…	The development of generalized pustules in a patient with psoriasis…
What is the most likely diagnosis for a 2-year-old 70 kg child who present…	Okay, so we're dealing with a 2-year-old child here. But wait, this kid i…	Based on the described symptoms and the unusual weight for a 2-year-old…

‹ Previous 1 2 3 ... 198 Next ›

Figure 1 Medical QA dataset with questions, reasoning, and responses.

Data creation process [1]:

The objective of "*HuatuoGPT-o1, Towards Medical Complex Reasoning with LLMs*" paper [1] is to create open-ended, verifiable problems requiring complex reasoning. It has two stages; 1- Filtering: Remove easy, short, or ambiguous questions Ensure unique, objective answer and 2- Reformat: Convert MCQ to open-ended format (x). Retain only the correct answer (y*). The process is explained in Figure 2.

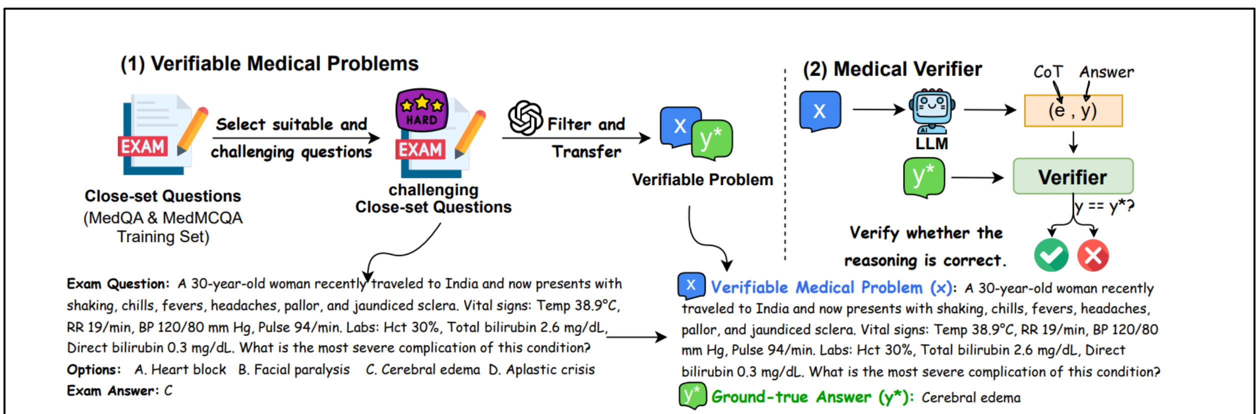


Figure 1: **Left:** Constructing verifiable medical problems using challenging close-set exam questions. **Right:** The verifier checks the model's answer against the *ground-truth answer*.

Figure 2 Data Creation Process [1]

Preprocessing:

Subset section of data:

There are 04 subsets:

- en – 19.k rows: only English
- en_mix - 24.9k rows : English and Chinese mix
- zh – 20.2krows :only Chinese

- Zh_mix – 25.4k rows : Chinese and English mix

I choose subset of “en” rows as shown in Figure 3; 2000 entries for training which are used for fine tuning whereas as the 600 rows were used for fine-tuned model’s inferences and evaluation.

```
from datasets import load_dataset
dataset_train = load_dataset("FreedomIntelligence/medical-o1-reasoning-SFT","en", split = "train[0:2000]",trust_remote_code=True)
dataset_train = dataset_train.map(formatting_prompts_func, batched = True,)
dataset_test = load_dataset("FreedomIntelligence/medical-o1-reasoning-SFT","en", split = "train[2001:2601]",trust_remote_code=True
dataset_test = dataset_train.map(formatting_prompts_func, batched = True,)
```

```
README.md: 100% 1.97k/1.97k [00:00<00:00, 239kB/s]
medical_o1_sft.json: 100% 58.2M/58.2M [00:00<00:00, 104MB/s]
Generating train split: 100% 19704/19704 [00:00<00:00, 20046.51 examples/s]
Map: 100% 2000/2000 [00:00<00:00, 25351.27 examples/s]
```

Figure 3 Train/Test split

Prompt Formatting

The below Table 1 shows prompt style helps instruction-following LLMs (like Mistral, DeepSeek) understand how to answer in CoT style.

```
prompt_style = """Below is an instruction that describes a task, paired with
an input that provides further context.
Write a response that appropriately completes the request.
Before answering, think carefully about the question and create a step-by-
step chain of thoughts to ensure a logical and accurate response.

### Instruction:
You are a medical expert with advanced knowledge in clinical reasoning,
diagnostics, and treatment planning.
Please answer the following medical question.

### Question:
{}

### Response:
<think>{}"""
```

Table 1 Prompt style for CoT format

BioBert prompt format is given below in Table 2:

```
def format_for_biobert_batch(batch):
    return {
        "question": batch["Question"],
        "context": batch["Complex_CoT"],
        "reference": batch["Response"]
    }
```

Table 2 BioBert prompt format function

BioBERT (baseline model) is not trained on CoT-style reasoning by default therefore it was generating incorrect response as shown in Figure 7. To ensure fair comparison or improved performance, use the original dataset used for generating CoT-formatted reasoning examples. I have used MedQA-USMLE dataset (Figure 4) for biobert.

id string · lengths 11 11	sent1 string · lengths 66 3.58k	sent2 string · classes 1 value	ending0 string · lengths 1 230	ending1 string · lengths 1 206	ending2 string · lengths 1 251	ending3 string · lengths 1 212	label int64 0 3
train-00000	A 23-year-old pregnant woman at 22 weeks gestatio…		Ampicillin	Ceftriaxone	Doxycycline	Nitrofurantoin	3
train-00001	A 3-month-old baby died suddenly at night while…		Placing the infant in a supine position on a firm…	Keeping the infant covered and…	Application of a device to maintain…	Avoiding pacifier use during sleep	0
train-00002	A mother brings her 3-week-old infant to the…		Abnormal migration of ventral pancreatic bud	Complete failure of proximal duodenum t…	Abnormal hypertrophy of the pylorus	Failure of lateral body folds to move…	0
train-00003	A pulmonary autopsy specimen from a 58-year-ol…		Thromboembolism	Pulmonary ischemia	Pulmonary hypertension	Pulmonary passive congestion	0
train-00004	A 20-year-old woman presents with menorrhagia…		Hemophilia A	Lupus anticoagulant	Protein C deficiency	Von Willebrand disease	3
train-00005	A 40-year-old zookeeper presents to the emergency…		Aspirin	Oral contraceptive pills	Scorpion sting	Hypothyroidism	2
train-00006	A 25-year-old primigravida presents to her physician…		Bilirubin assessment	Coagulation studies	Leukocyte count with differential	24-hour urine protein	3

Figure 4 MedQA-USMLE dataset

```
[ ] dataset["train"][0]

{'id': 'train-00000',
 'sent1': 'A 23-year-old pregnant woman at 22 weeks gestation presents with burning upon urination. She states it started 1 day ago and has been worsening despite drinking more water and taking cranberry extract. She otherwise feels well and is followed by a doctor for her pregnancy. Her temperature is 97.7°F (36.5°C), blood pressure is 122/77 mmHg, pulse is 80/min, respirations are 19/min, and oxygen saturation is 98% on room air. Physical exam is notable for an absence of costovertebral angle tenderness and a gravid uterus. Which of the following is the best treatment for this patient?',
 'sent2': '',
 'ending0': 'Ampicillin',
 'ending1': 'Ceftriaxone',
 'ending2': 'Doxycycline',
 'ending3': 'Nitrofurantoin',
 'label': 3}
```

Figure 5 Training Q/A format of MedQA-USML dataset for BioBert

Many preprocessing steps were performed automatically by the HuggingFace unsloth library such as Tokenization etc. as shown in Figure 17.

Models Used:

I have used five LLMs as shown below. For fine tuning LLMs we use unsloth which make the process easier and faster.

General Purpose LLMs:

1- DeepSeek
   a. Modal Name: unsloth/DeepSeek-R1-Distill-Llama-8B
   b. Model type: it is distilled version of DeepSeek's LLaMA-style model
   c. Model has ability to explore chain-of-thought (CoT) for solving complex problems
2- LLaMA
   a. Model name: unsloth/llama-3-8b-Instruct-bnb-4bit
   b. It is designed for commercial and research use in English, with instruction-tuned models for assistant-style chat and pretrained models for broader language generation tasks.
3- Mistral-based
   a. Model name: unsloth/mistral-7b-instruct
   b. It is instruct type model which means it can support chat style prompt-response (Q/A)

c. It has 7B parameters
d. Decoder only transformer model (like - GPT)

Medical Domain LLMs

1- BioBert:
As shown in the above snapshot of clinicalbert research papers abstract; Clinical Bert works on summarization dataset not the reasoning question/answer.

2- Me-LLaMA:

Model	Domain	Type	Arch.	Params	Max Seq Len
DeepSeek	General	Distilled LLaMA	Decoder	8B	2048
LLaMA-3	General	Instruct LLaMA-3	Decoder	8B	8192 (used 2048)
Mistral	General	Instruct	Decoder	7B	2048
BioBERT	Medical	BERT-based	Encoder	110M	512 (used 384)
Me-LLaMA	Medical	Instruct	Decoder	7B+	2048

Table 3 Overview of LLMs with domain, architecture, and key features.

Architectural Diagram

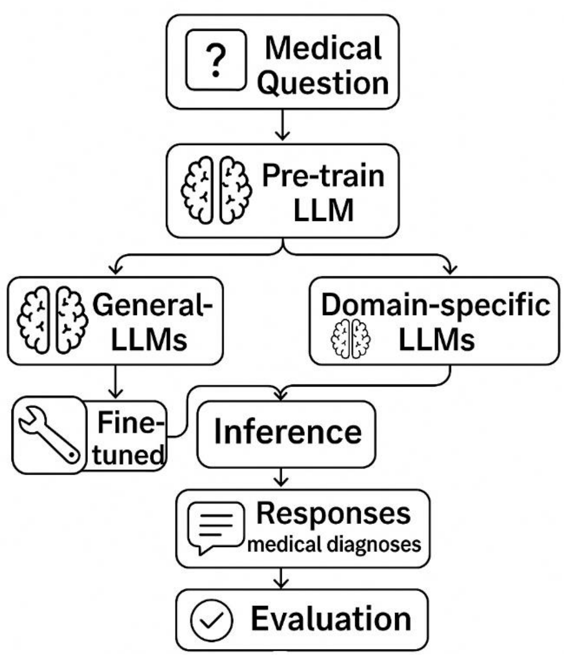


Figure 6 High level flow diagram of project

Results

This section presents the evaluation of five LLMs on medical question-answering datasets. Five LLMs were evaluated on medical QA tasks using qualitative analysis due to GPU constraints. Fine-tuned DeepSeek and Mistral models produced effective chain-of-thought responses, with DeepSeek showing strong clinical reasoning. Me-LLaMA performed well without fine-tuning, benefiting from domain-specific training.

BioBERT handled MCQ-style questions reasonably but struggled with open-ended prompts. LLaMA-3 failed to generate outputs due to prompt format issues.

1. BioBert:

Below Figure 7 shows the comparisons of predicted answer by BioBert and the reference answer of dataset (ground truth) and the fig rouge scores shows that model performed bad on the Complex CoT reasoning Q/A dataset.

	Question	Context	Reference_Answer	Predicted_Answer
0	A 3 month old child has moderate fever and non...	Alright, so we're dealing with a very young ch...	The most probable diagnosis for the 3-month-ol...	s what I
1	In a 65-year-old female presenting with recurr...	Alright, let's dive into what could be going o...	In a 65-year-old female presenting with recurr...	wouldn't there be more going on, like swelling...
2	List the characteristics of papillary carcinom...	Okay, let's start by thinking about papillary ...	Papillary carcinoma of the thyroid is the most...	to wrap it all up neatly, what this all
3	A 28 years old woman having limited cutaneous ...	Alright, let's see what we have here. The pati...	Based on the details provided, the patient is ...	to ILD
4	What condition is likely responsible for a pat...	The symptoms this patient is experiencing soun...	The condition likely responsible for the sympt...	.

Figure 7 Comparison of predicted responses by LLMs with reference answers from the dataset.

```
rouge_result = rouge.compute(predictions=predictions, references=references)
print("ROUGE:", rouge_result)
```

```
ROUGE: {'rouge1': np.float64(0.03771508216529726), 'rouge2': np.float64(0.0014285714285714286), 'rougeL': np.float64(0.035263890670344265), 'rougeLsum': np.float64(0.03394810119666006)}
```

Figure 8 Inferences (before fine tuning) Rouge Score of BioBert (Using CoT Q/A dataset)

Since the BioBERT (baseline model) is not trained on CoT-style reasoning by default therefore MedQA-USMLE dataset (Figure 4). BioBERT was fine-tuned using the Huggingface trainer API for 10 epochs with a learning rate of 1e-5 on the MedQA-USMLE dataset. Below Figure 9 shows that training/validations loss is reducing and accuracy is increased to 51%.

[1500/1500 31:56, Epoch 10/10]

Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1
1	No log	2.162738	0.412500	0.305276	0.319023	0.279735
2	No log	1.566764	0.485833	0.403851	0.395089	0.365586
3	No log	1.425991	0.510000	0.423719	0.417907	0.386979
4	2.232100	1.366492	0.528333	0.428170	0.419664	0.390411
5	2.232100	1.326433	0.519167	0.421433	0.406005	0.380258
6	2.232100	1.333301	0.529167	0.436324	0.426079	0.398285
7	1.085100	1.322855	0.515833	0.422197	0.406635	0.381527
8	1.085100	1.349827	0.515000	0.417795	0.406735	0.377801
9	1.085100	1.381859	0.513333	0.424837	0.407908	0.381069
10	0.913800	1.382810	0.508333	0.419173	0.404846	0.375500

Figure 9 Fine-tuning performance of BioBERT on MedQA-USMLE over 10 epochs

```
results
```

```
{'eval_loss': 1.3828099966049194,
 'eval_accuracy': 0.5083333333333333,
 'eval_precision': 0.41917298257265934,
 'eval_recall': 0.40484556805089716,
 'eval_f1': 0.3754997522711574,
 'eval_runtime': 14.4833,
 'eval_samples_per_second': 41.427,
 'eval_steps_per_second': 2.624,
 'epoch': 10.0}
```

Figure 10 Final results of Fine-tuned BioBERT on MedQA-USMLE

Figure 11 Evaluation metrics during BioBERT fine-tuning on the MedQA-USMLE dataset, showing trends in accuracy, loss, runtime, and throughput over training steps

Figure 12 Training metrics of BioBERT showing loss, learning rate, gradient norm, and training progression over epochs

2. DeepSeek:

Number_training_epochs = 5

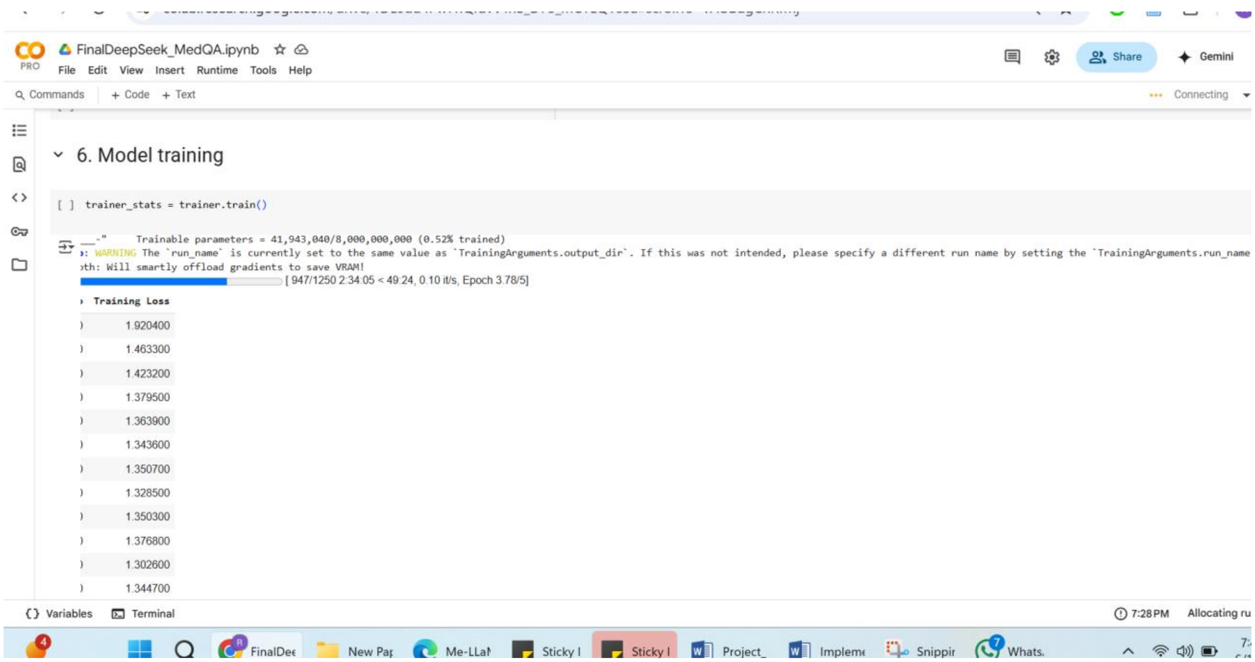


Figure 13 Training loss output during fine-tuning of the DeepSeek model using 5 epochs

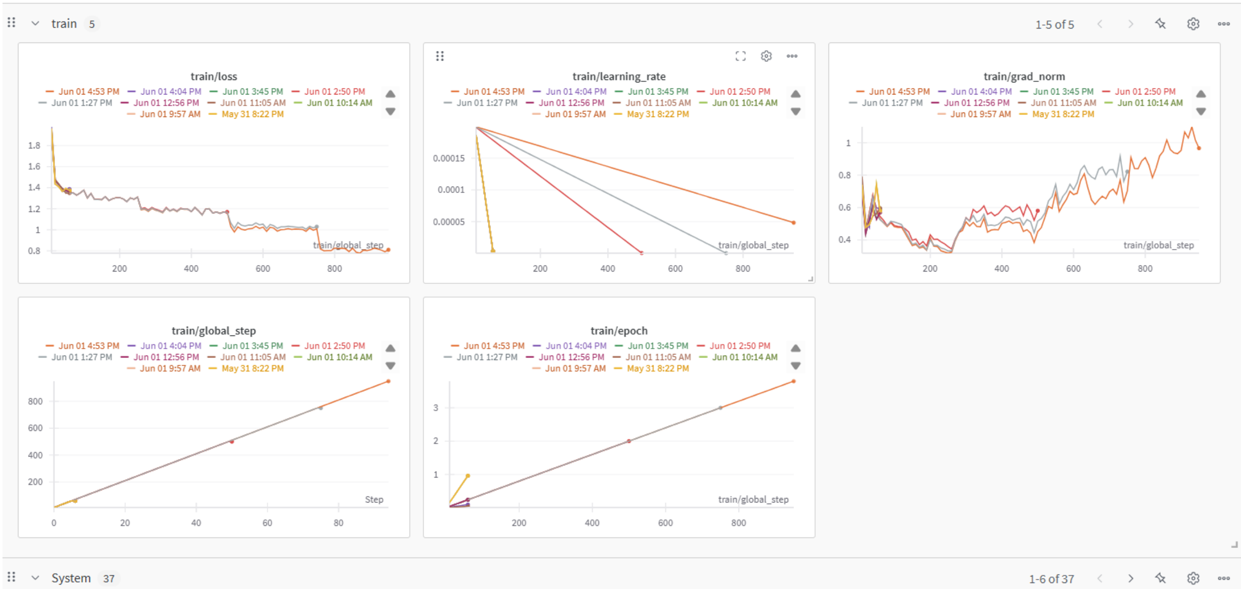


Figure 14 Training metrics of DeepSeek model across multiple runs, showing trends in loss, learning rate, gradient norm, and training progression

Figure 15 System resource utilization during DeepSeek model training, showing GPU power usage, memory consumption, and multiprocessor activity across multiple runs

Below figure _ is the snapshot of the post-training using max_step = 60.

Figure 16 Training loss log of the DeepSeek model over 60 steps using Unsloth with max_step set to 60 (1 Epoch)

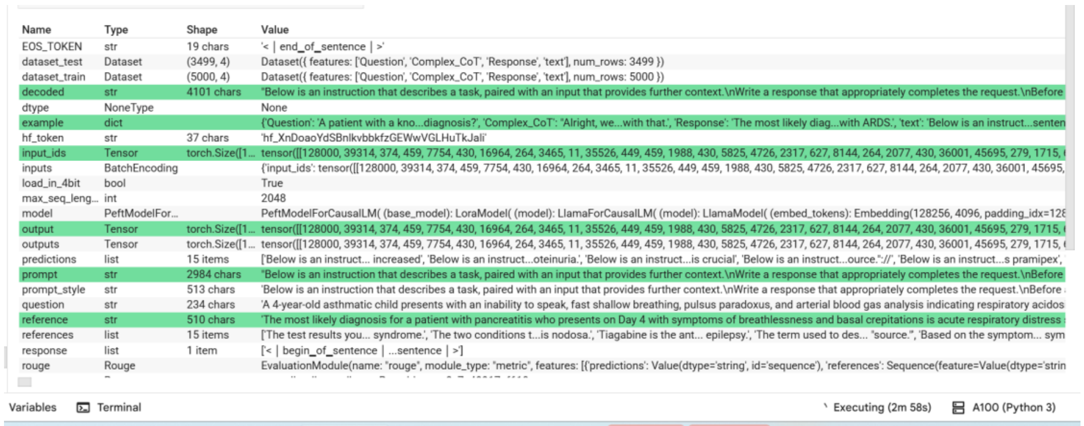


Figure 17 Model inference and evaluation pipeline showing tokenization, prompt formatting, and ROUGE score computation

Question:
A 4-year-old asthmatic child presents with an inability to speak, fast shallow breathing, pulsus paradoxus, and arterial blood gas analysis indicating respiratory acidosis. What is the next best step in the management of this patient?

Inferences:
 	
`<think>`
`Okay, so I'm trying to figure out the next best step in managing this 4-year-old asthmatic child. The child has several symptoms: inability to speak, fast shallow breathing, pulsus paradoxus, and respiratory acidosis on the blood gas. Let me break this down.`
 	
`First, the child can't speak, which could indicate a severe respiratory issue. Fast shallow breathing is typical in asthma, but combined with pulsus paradoxus and respiratory acidosis, this might be more than just a typical asthma attack.`
 	
`Pulsus paradoxus is a condition where the heart rate changes with each breath. It's often seen in conditions like severe asthma, where the increased effort to breathe can affect the heart. Respiratory acidosis means there's too much carbon dioxide in the blood, which can happen if the lungs aren't getting enough oxygen, leading the body to retain CO2.`

```
So, putting this together, the child's condition seems to be deteriorating.
The inability to speak suggests they can't breathe properly, which is a sign
of a severe respiratory emergency. Pulsus paradoxus and acidosis are
indicators that the child is in a life-threatening state.

In such cases, the next step would be to provide immediate care to stabilize
the child. The primary concern is ensuring they can breathe. Intubation is a
procedure where a tube is inserted into the trachea to help the child
breathe. This is usually done under anesthesia and requires a team of
specialists, like an anesthesiologist and pediatric intensivist.

Before intubation, other measures might have been tried, like increasing
oxygen, using a continuous positive airway pressure (CPAP) device, or
administering bronchodilators. However, given the severity indicated by
acidosis and pulsus paradoxus, intubation might be necessary to prevent
further complications, such as hypoxia or complete respiratory failure.

Therefore, the next best step is likely intubation to provide assisted
breathing and stabilize the child's condition.
</think>

The next best step in managing the 4-year-old asthmatic child with severe
respiratory distress, including pulsus paradoxus and respiratory acidosis,
is **intubation**. This procedure ensures the child receives assisted
breathing, stabilizing their condition and preventing complications from
severe respiratory failure.<｜end▁of▁sentence｜>
```

Table 4 DeepSeek model inference with clinical reasoning.

3. Mistral:

Figure 18 and Figure 19 illustrate the fine-tuning process and system configuration of the Mistral model. The model's output shown in Table 1 demonstrates structured decision-making aligned with medical rules.

Question: A 4-year-old asthmatic child presents with an inability to speak, fast shallow breathing, pulsus paradoxus, and arterial blood gas analysis indicating respiratory acidosis. What is the next best step in the management of this patient?
Inferences: <think> 1. The child's symptoms suggest a severe asthma exacerbation, which is causing respiratory distress. 2. Pulsus paradoxus, an increase in the systolic blood pressure by more than 10 mmHg during inspiration, is a sign of severe airway obstruction. 3. Respiratory acidosis, as indicated by the arterial blood gas analysis, is a result of CO2 retention due to impaired ventilation. 4. The inability to speak and fast shallow breathing are also indicative of respiratory distress.

Given these findings, the next best step in the management of this patient would be:

1. Immediate administration of high-dose systemic corticosteroids to reduce inflammation in the airways.
2. Initiate or increase the dose of bronchodilators, such as beta-2 agonists or anticholinergics, to relax the smooth muscles of the airways and improve airflow.
3. Provide oxygen therapy to increase the oxygen saturation and help alleviate the respiratory acidosis.
4. Consider intubation and mechanical ventilation if the patient's condition does not improve with initial treatments or if they develop signs of impending respiratory failure, such as cyanosis, severe hypoxemia, or decreasing level of consciousness.
5. Monitor the patient closely for any signs of complications, such as pneumothorax or hypotension, and adjust treatment accordingly.
6. Consult a pediatric pulmonologist or intensive care specialist for further management and potential admission to the pediatric intensive care unit (PICU).

This response is intended to provide a general approach to managing a severe asthma exacerbation in a child. The specific treatment plan may vary depending on the patient's individual circumstances and the judgment of the treating physician.</s>

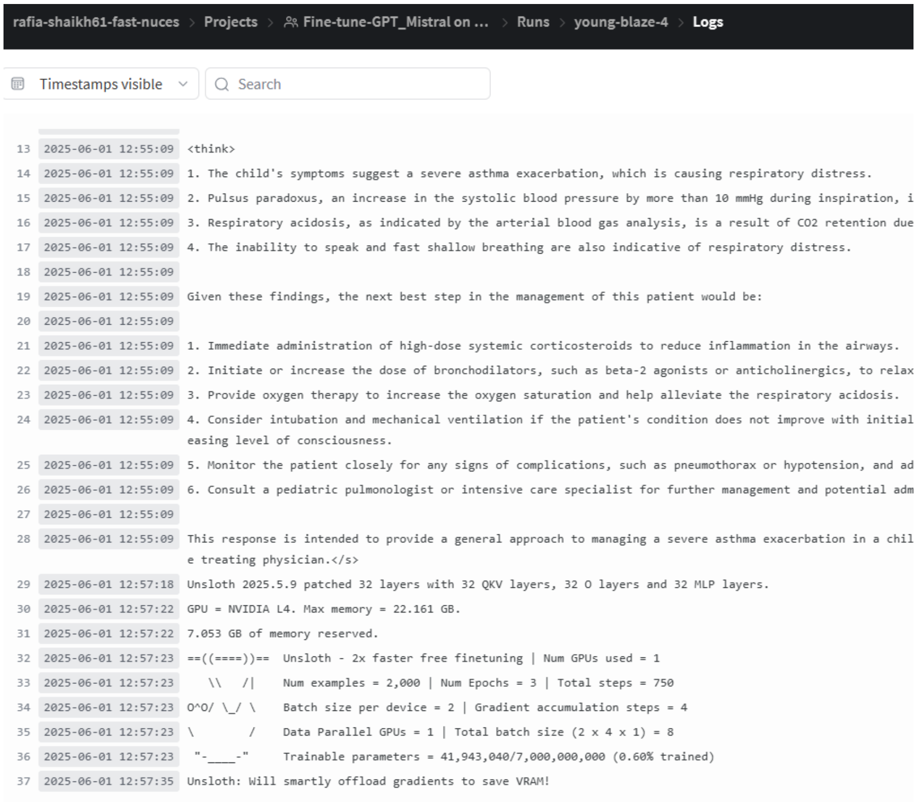
Table 5 Mistral model inference with clinical reasoning.

```
GPU = NVIDIA L4. Max memory = 22.161 GB.
7.053 GB of memory reserved.
```

```
trainer_stats = trainer.train()
```

255	0.900200
256	0.980900
257	0.895000
258	0.858200
259	0.915800
260	1.006700
261	0.892300
262	0.901900
263	0.887500
264	0.874600
265	0.947600
266	0.969400
267	0.811200

Figure 18 Training loss output during fine-tuning of the Mistral model



rafia-shaikh61-fast-nuces > Projects > Fine-tune-GPT_Mistral on ... > Runs > young-blaze-4 > Logs

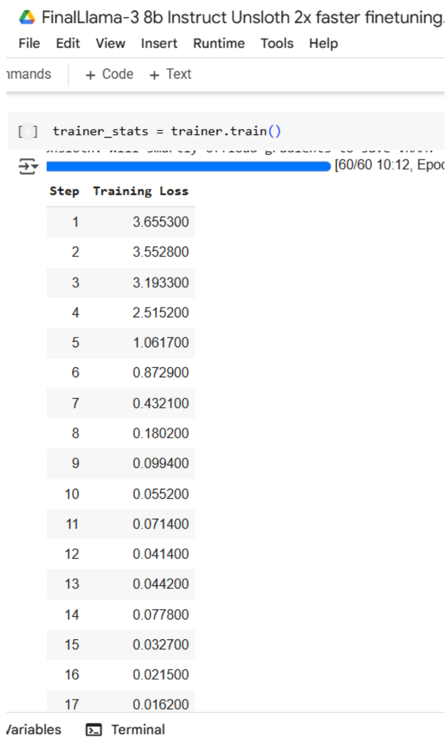
Timestamps visible   Search

```
2025-06-01 12:55:09 <think>
2025-06-01 12:55:09 1. The child's symptoms suggest a severe asthma exacerbation, which is causing respiratory distress.
2025-06-01 12:55:09 2. Pulsus paradoxus, an increase in the systolic blood pressure by more than 10 mmHg during inspiration, i
2025-06-01 12:55:09 3. Respiratory acidosis, as indicated by the arterial blood gas analysis, is a result of CO2 retention due
2025-06-01 12:55:09 4. The inability to speak and fast shallow breathing are also indicative of respiratory distress.
2025-06-01 12:55:09
2025-06-01 12:55:09 Given these findings, the next best step in the management of this patient would be:
2025-06-01 12:55:09
2025-06-01 12:55:09 1. Immediate administration of high-dose systemic corticosteroids to reduce inflammation in the airways.
2025-06-01 12:55:09 2. Initiate or increase the dose of bronchodilators, such as beta-2 agonists or anticholinergics, to relax
2025-06-01 12:55:09 3. Provide oxygen therapy to increase the oxygen saturation and help alleviate the respiratory acidosis.
2025-06-01 12:55:09 4. Consider intubation and mechanical ventilation if the patient's condition does not improve with initial
                    easing level of consciousness.
2025-06-01 12:55:09 5. Monitor the patient closely for any signs of complications, such as pneumothorax or hypotension, and ad
2025-06-01 12:55:09 6. Consult a pediatric pulmonologist or intensive care specialist for further management and potential adm
2025-06-01 12:55:09
2025-06-01 12:55:09 This response is intended to provide a general approach to managing a severe asthma exacerbation in a chil
                    e treating physician.</s>
2025-06-01 12:57:18 Unsloth 2025.5.9 patched 32 layers with 32 QKV layers, 32 O layers and 32 MLP layers.
2025-06-01 12:57:22 GPU = NVIDIA L4. Max memory = 22.161 GB.
2025-06-01 12:57:22 7.053 GB of memory reserved.
2025-06-01 12:57:23 ==((====))==  Unsloth - 2x faster free finetuning | Num GPUs used = 1
2025-06-01 12:57:23    \\   /|    Num examples = 2,000 | Num Epochs = 3 | Total steps = 750
2025-06-01 12:57:23 O^O/ \_/ \    Batch size per device = 2 | Gradient accumulation steps = 4
2025-06-01 12:57:23 \        /    Data Parallel GPUs = 1 | Total batch size (2 x 4 x 1) = 8
2025-06-01 12:57:23  "-____-"     Trainable parameters = 41,943,040/7,000,000,000 (0.60% trained)
2025-06-01 12:57:35 Unsloth: Will smartly offload gradients to save VRAM!
```

Figure 19 Log file showing Mistral model's inference output and system configuration during fine-tuning.

4. LLaMA 3:

The model failed to generate coherent responses. One likely reason is that the prompt did not follow the instruction format required by LLaMA-3 models, resulting in repetitive `<think>` outputs as shown in Figure 22. Despite this, the model showed a consistent decrease in training loss and stable learning behavior during fine-tuning as shown in Figure 20 and Figure 21.



Step	Training Loss
1	3.655300
2	3.552800
3	3.193300
4	2.515200
5	1.061700
6	0.872900
7	0.432100
8	0.180200
9	0.099400
10	0.055200
11	0.071400
12	0.041400
13	0.044200
14	0.077800
15	0.032700
16	0.021500
17	0.016200

Figure 20 Training loss output during fine-tuning of the LLaMA model

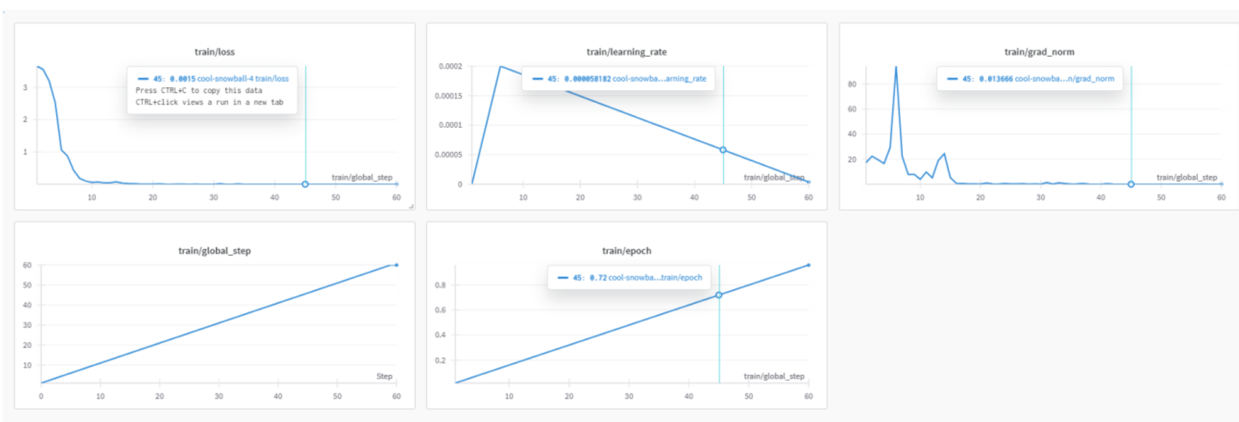
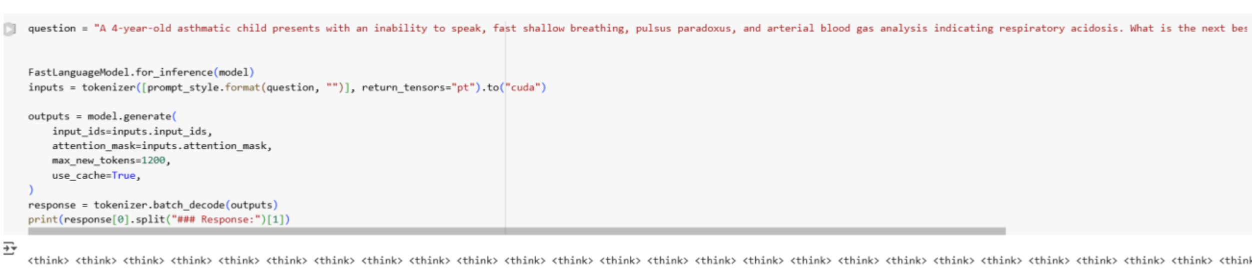


Figure 21 Training metrics of LLaMA-3 model showing loss reduction, learning rate decay, and gradient stability over 60 steps.



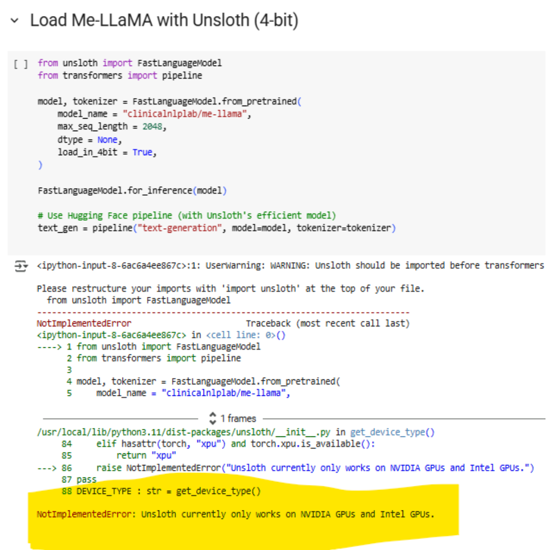
```
question = "A 4-year-old asthmatic child presents with an inability to speak, fast shallow breathing, pulsus paradoxus, and arterial blood gas analysis indicating respiratory acidosis. What is the next bes

FastLanguageModel.for_inference(model)
inputs = tokenizer([prompt_style.format(question, "")], return_tensors="pt").to("cuda")

outputs = model.generate(
    input_ids=inputs.input_ids,
    attention_mask=inputs.attention_mask,
    max_new_tokens=1200,
    use_cache=True,
)
response = tokenizer.batch_decode(outputs)
print(response[0].split("### Response:")[1])
```

```
<think> <think> <think> <think> <think> <think> <think> <think> <think> <think> <think> <think> <think> <think> <think> <think> <think> <think> <think> <think> <think> <think> <think> <think> <think> <think
```

Figure 22 Inference output with repetitive "<think>" tokens due to formatting issue.

5. Me-LLaMA

The last model I attempted to run was Me-LLaMA; however, at that time my GPU usage quota had reached its limit. As a result, I was unable to execute the model for inference or evaluation.


Load Me-LLaMA with Unsloth (4-bit)

```
from unsloth import FastLanguageModel
from transformers import pipeline

model, tokenizer = FastLanguageModel.from_pretrained(
    model_name = "clinicalnlplab/me-llama",
    max_seq_length = 2048,
    dtype = None,
    load_in_4bit = True,
)

FastLanguageModel.for_inference(model)

# Use Hugging Face pipeline (with Unsloth's efficient model)
text_gen = pipeline("text-generation", model=model, tokenizer=tokenizer)
```

```
<ipython-input-8-6ac6a4ee867c>:1: UserWarning: WARNING: Unsloth should be imported before transformers

Please restructure your imports with 'import unsloth' at the top of your file.
  from unsloth import FastLanguageModel
---------------------------------------------------------------------------
NotImplementedError                       Traceback (most recent call last)
<ipython-input-8-6ac6a4ee867c> in <cell line: 0>()
----> 1 from unsloth import FastLanguageModel
      2 from transformers import pipeline
      3
      4 model, tokenizer = FastLanguageModel.from_pretrained(
      5     model_name = "clinicalnlplab/me-llama",

1 frames
/usr/local/lib/python3.11/dist-packages/unsloth/__init__.py in get_device_type()
     84     elif hasattr(torch, "xpu") and torch.xpu.is_available():
     85         return "xpu"
---> 86     raise NotImplementedError("Unsloth currently only works on NVIDIA GPUs and Intel GPUs.")
     87 pass
     88 DEVICE_TYPE : str = get_device_type()

NotImplementedError: Unsloth currently only works on NVIDIA GPUs and Intel GPUs.
```

Figure 23 Error loading Me-LLaMA with Unsloth due to GPU compatibility issue

Figure 24 System-level GPU and memory utilization metrics during multiple training runs

Conclusion

In conclusion, domain-specific models like Me-LLaMA showed strong performance on complex medical reasoning tasks, while general-purpose models like DeepSeek also performed well when fine-tuned with proper prompts. BioBERT worked best for multiple-choice questions but failed on CoT-based reasoning. LLaMA-3 underperformed due to prompt mismatch. These findings suggest that model selection in medical AI applications should consider domain relevance and task type.

References

[1] J. Chen, Z. Cai, K. Ji, X. Wang, W. Liu, R. Wang, J. Hou, and B. Wang, "HuatuoGPT-o1, Towards Medical Complex Reasoning with LLMs," arXiv preprint arXiv:2412.18925, Dec. 2024. [Online]. Available: https://arxiv.org/abs/2412.18925

[2]X. Liu, J. Li, T. Bai, J. Gao, P. Zhang, and X. Zhao, "Research on Development of LLMs and Manual Comparison of Applications," 2024 10th International Conference on Big Data and Information Analytics (BigDIA), pp. 23–30, Oct. 2024, doi: 10.1109/bigdia63733.2024.10808821.

[3]F. Neha and D. Bhati, "A Survey of DeepSeek Models," Feb. 2025, doi: 10.36227/techrxiv.173896582.25938392/v1.

[4]R. Sapkota, S. Raza, and M. Karkee, "Comprehensive Analysis of Transparency and Accessibility of ChatGPT, DeepSeek, And other SoTA Large Language Models," Feb. 2025, doi: 10.20944/preprints202502.1608.v1.

[5]O. Aydin, E. Karaarslan, F. S. Erenay, and N. B. Džakula, "Generative AI in Academic Writing: A Comparison of DeepSeek, Qwen, ChatGPT, Gemini, Llama, Mistral, and Gemma ," Mar. 2025, doi: 10.36227/techrxiv.174137796.60885820/v1.

[6]L. Agarwal and A. Nasim, "Comparison and Analysis of Large Language Models (LLMs)," 2024, doi: 10.2139/ssrn.4939534.

[7]M. Cascella, J. Montomoli, V. Bellini, and E. Bignami, "Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios," Journal of Medical Systems, vol. 47, no. 1, Mar. 2023, doi: 10.1007/s10916-023-01925-4.

[8]U. Mumtaz, A. Ahmed, and S. Mumtaz, "LLMs-Healthcare: Current applications and challenges of large language models in various medical specialties," Artificial Intelligence in Health, vol. 1, no. 2, p. 16, Apr. 2024, doi: 10.36922/aih.2558.

[9]D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits, "What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams," Applied Sciences, vol. 11, no. 14, p. 6421, Jul. 2021, doi: 10.3390/app11146421

Annex: Code

1. DeepSeek:
   https://colab.research.google.com/drive/1DL0dd4Pw7hQfdvVTnS_D75_mCYLQ165a?usp=sharing
2. LLaMA:
   https://colab.research.google.com/drive/1otjN5E8RxldDtR4JrjjnMavxp6WNhKXi?usp=sharing
3. GPT (Mistral-based):
   https://colab.research.google.com/drive/1LtQMv2fJr1k78jjFhSegfqkioGzZAgGC?usp=sharing
4. BioBert:
   https://colab.research.google.com/drive/1cOxNlyyUME67lHo54f72RXPYwNPM20CR?authuser=2#scrollTo=Y-igtFoVRR4R
   https://colab.research.google.com/drive/1_AORRPsTAVhL8OD6XrHE1zVBtnUo1E1k?usp=sharing
   https://colab.research.google.com/drive/1JuJraOXluyqL_P3LCszEu7YQIBaVtpS1?usp=sharing
5. Me-LLaMA:
   https://colab.research.google.com/drive/1oB0-mKCLLC94pj2lWmDch_3tclwnzHYx?usp=sharing
6. Combined model's code file:
   https://colab.research.google.com/drive/1T7EanB17bxgjfnOow0vwmruzDO2KqfIG?usp=sharing