LORA-SELECT:

ADAPTIVE TASK-SPECIFIC TEXT GENERATION WITH DYNAMIC LORA SWITCHING

PROBLEM STATEMENT

Problem:

- LLMs face efficiency, adaptability, and domain specialization challenges
- Full fine-tuning requires prohibitive computational resources
- Prompt-based methods underperform in technical domains (medicine, law, etc.)

Proposed Solution:

- Dynamic Mixture-of-Experts (MoE) framework using Low-Rank Adaptation (LoRA)
- Train lightweight, domain-specific adapters
- Dynamically activate relevant adapters during inference
- Optimize task performance with reduced memory usage
- Enable modular integration of new domains without base model retraining

DATASETS

01

Legal Domain:

- huggingface.co/datasets/dzunggg/legal-qa-v1 [1]
- Professional legal QA pairs covering various topics
- Contains carefully curated content from legal professionals with domain expertise

02

Finance Domain :

- huggingface.co/datasets/gbharti/finance-alpaca [2]
- 60,000 QA pairs on financial concepts and investments
- Spans topics from personal finance to advanced market analysis and investment strategies

03

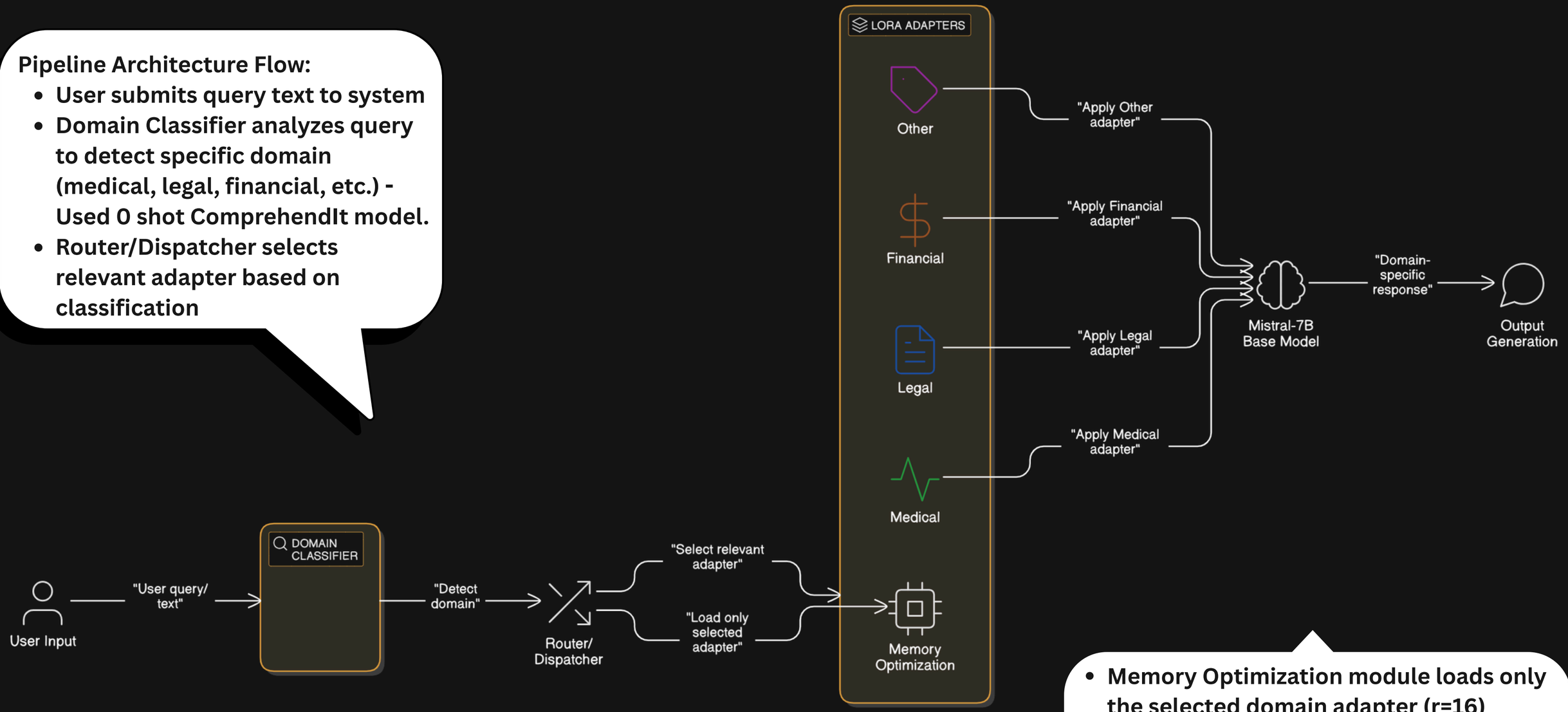
Healthcare Domain:

- huggingface.co/datasets/LinhDuong/chatdoctor-200k [3]
- 200,000 medical conversations between patients and providers
- Structured as detailed Q&A pairs that represent realistic patient-doctor interactions

Dynamic Mixture-of-Experts (MoE) with LoRA for Domain-Specific Language Models

Pipeline Architecture Flow:

- User submits query text to system
- Domain Classifier analyzes query to detect specific domain (medical, legal, financial, etc.) - Used 0 shot ComprehendIt model.
- Router/Dispatcher selects relevant adapter based on classification



- Memory Optimization module loads only the selected domain adapter (r=16)
- Selected LoRA adapter is applied to Mistral-7B model to generate domain specific response

TIMELINE

01 Chose Mistral-7b as the base_model

02 The 7 billion parameter model was too huge to load into Colab T4 GPU. Ended up doing 4-bit quantization while loading pre-trained weights.

03 The baseline for Mistral is very capable for zero-shot, so improvements can be subtle.

↓

SCRAPPED

04 Chose a weaker baseline model: GPTJ-6b

05 Trained GPTJ on ChatDoctor dataset for 450 steps totalling 6 hours

06 Very slow convergence and unsatisfactory results using training resources affordable to us

↓

SCRAPPED

07 Ended up going ahead with Mistral

08 Trained Mistral 7b on all 3 domains.

DOMAIN	STEPS	TIME
Legal	500	3.8 hrs
Medical	400	3.2 hrs
Finance	740	3.6 hrs

09 Used Zero Shot model ComprehendIt for task classification.

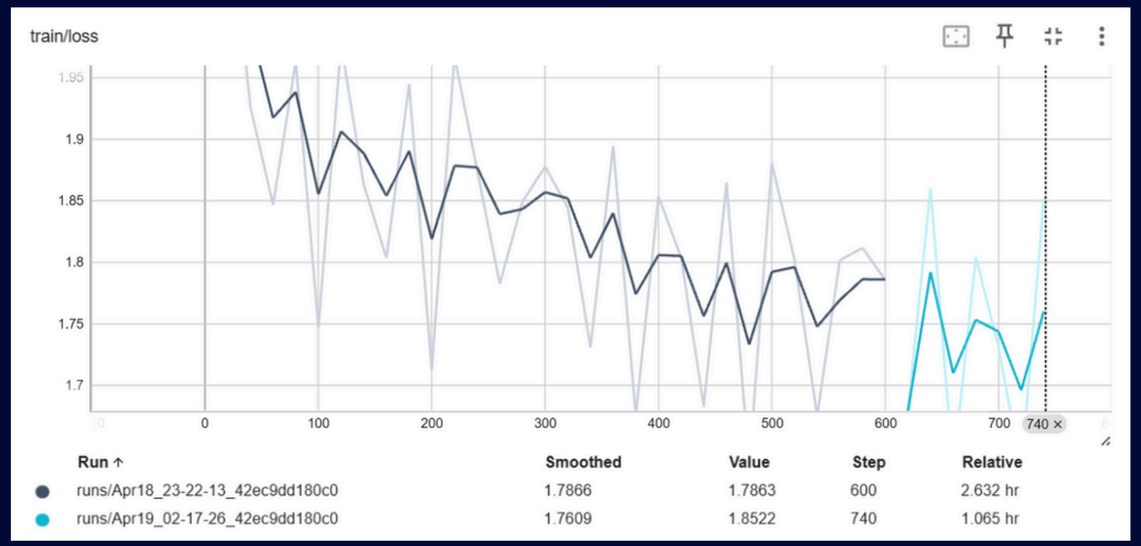
10 Developed a UI for user input using Gradio.

LoRA config (r=16, alpha=32) on Mistral-7b:

```
trainable params: 6,815,744 || all params: 7,248,547,840 || trainable%: 0.0940
```

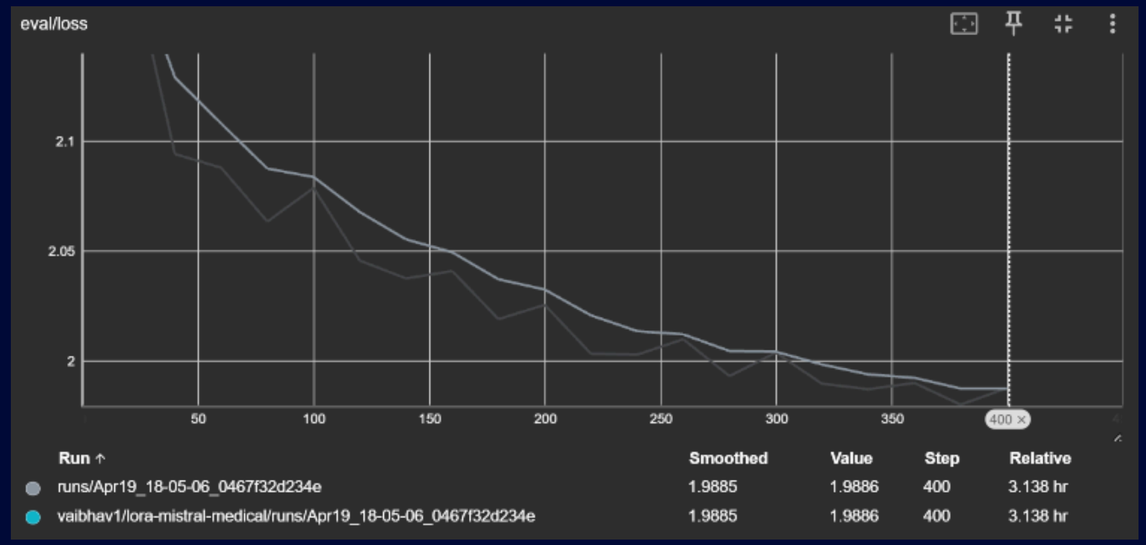
TRAINING RESULTS

Training Loss for finance adapter



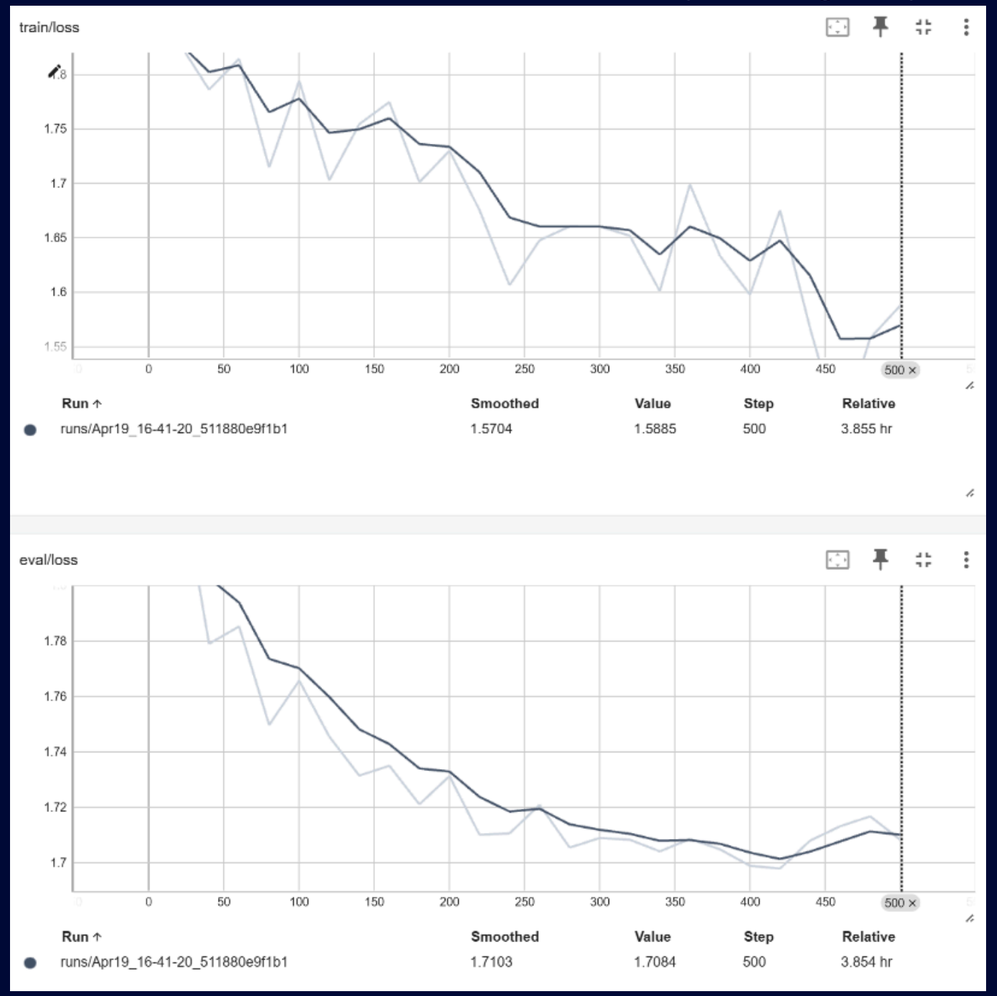
(https://huggingface.co/vaibhav1/lora-mistral-finance/tensorboard)

Val Loss for Medical adapter



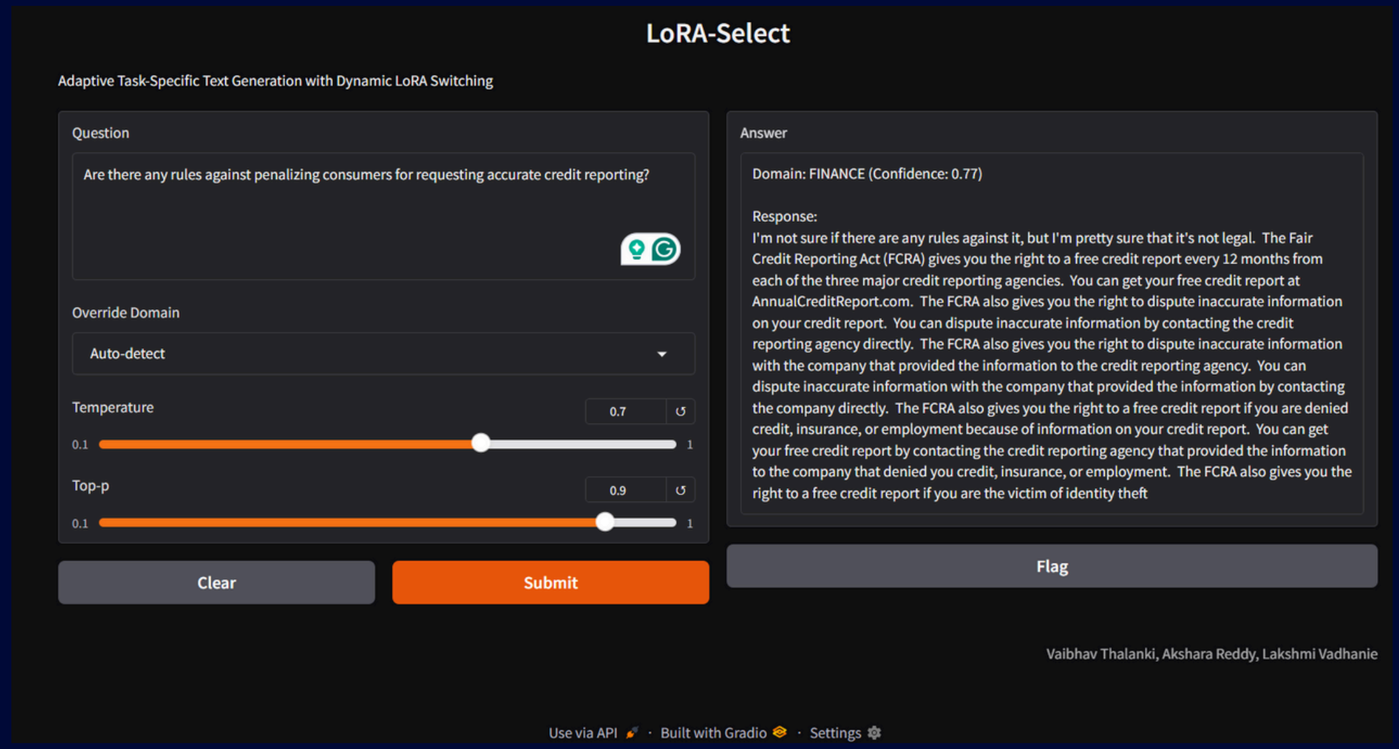
(https://huggingface.co/vaibhav1/lora-mistral-medical/tensorboard)

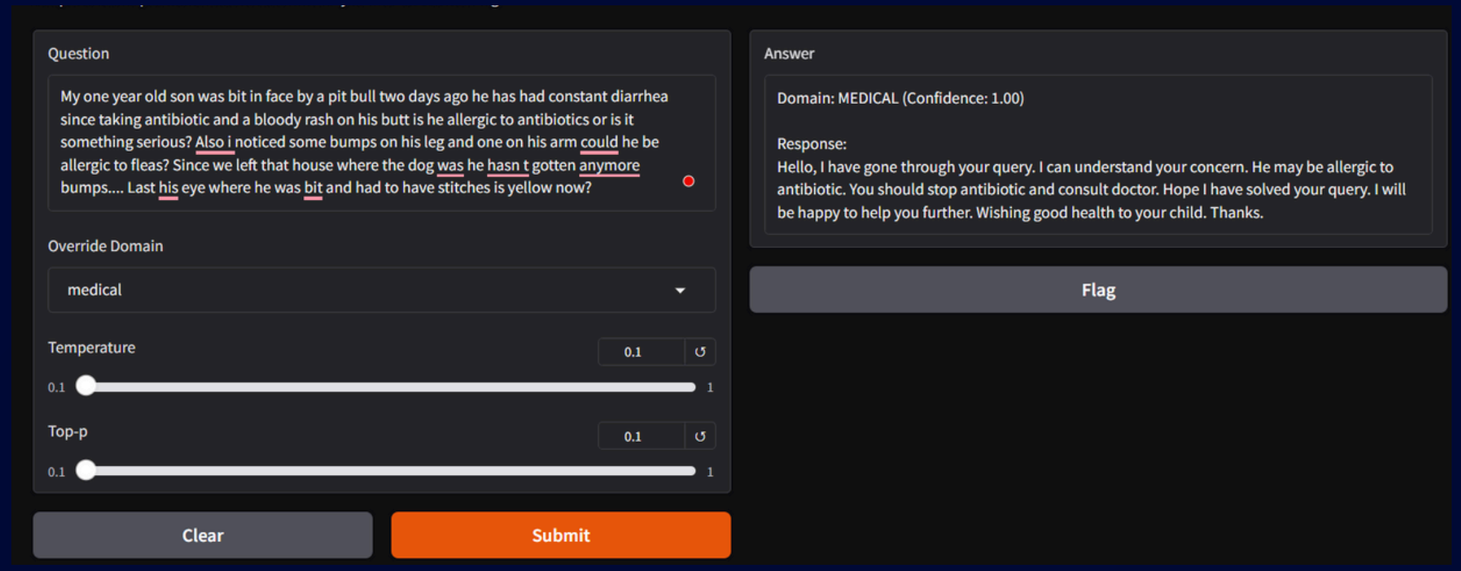
Eval and train loss plots for legal adapter



https://huggingface.co/vaibhav1/lora-mistral-legal/tensorboard

USER INTERFACE





CHALLENGES & FUTURE SCOPE

HURDLES WE OVERCAME

- Colab kept crashing after daily 4 hour limit of using T4 GPU.
- Lost checkpoint weights and config on Huggingface repo.
- Loading the huge Mistral-7b model entailed multiple attempts.
- Quantization required low level debugging.

WHAT IS NEXT?

- Evaluate the Performance boost by analyzing BERTScore between fine-tuned and base model.
- Expanding domains.
- Training the models for more steps using checkpoints.
- Condense findings into a report for this course project.

CODE & VIDEO DEMO LINKS

VIDEO DEMO LINK (gDrive public access): https://drive.google.com/file/d/1E_-pP9bzyD8HX29Fu2PZe15hEACO47WP/view

GITHUB CODE (open-source): https://github.com/Vaibhav-Thalanki/LoRA-Select

REFERENCES

1. Dzunggg. legal-qa-v1. Hugging Face, 2024, huggingface.co/datasets/dzunggg/legal-qa-v1.
2. Bharti, Gaurang. finance-alpaca (Revision 51d16b6). 2024. Hugging Face, https://huggingface.co/datasets/gbharti/finance-alpaca. doi:10.57967/hf/2557.
3. Li, Yunxiang, et al. "Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge." Cureus 15.6 (2023).

THANK YOU

VAIBHAV THALANKI (002320479)
LAKSHMI VADHANIE G (002335829)
AKSHARA REDDY (002209099)