FINE TUNING

FALCON-7B MODEL

VS.

LLAMA3-8B INSTRUCT MODEL

Style of Imran Khan

By: Muhammad Wasam Khan

INTRODUCTION

This presentation provides a **comparative analysis** of the *Fine Tuning Falcon-7B* and *Llama3-8B Instruct* models. We will examine their specifications, features, and performance to determine the most suitable option for your needs.

SPECIFICATIONS

Falcon-7B: 7 billion parameters, designed for balanced performance and efficiency.
Llama 3-8B: 8 billion parameters, geared towards more advanced and complex language tasks.

SPECIFICATIONS

Falcon-7B: Effective for chatbots, content generation, and educational tools.
Llama 3-8B: Ideal for advanced virtual assistants, in-depth content analysis, and research automation.

DATASET PRPARATION

Falcon-7B

Dataset

input and output columns only

Content

Interview questions and answers of Imran Khan

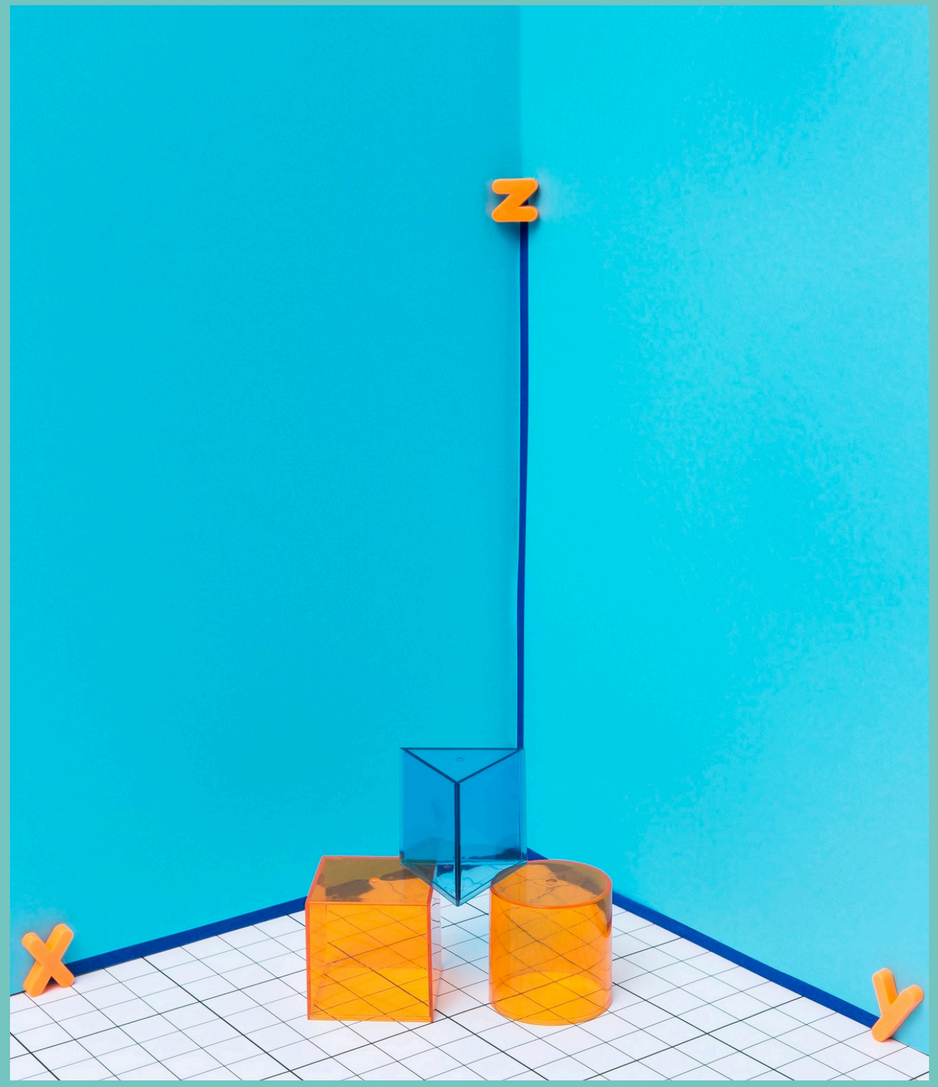
Llama 3-8B

Dataset

input, output, and instructions

Content

Interview questions, answers, and additional instructions



FINE TUNING FALCON-7B

Quantization

Used **BitsAndBytes** configuration

4-bit quantization to reduce memory footprint.

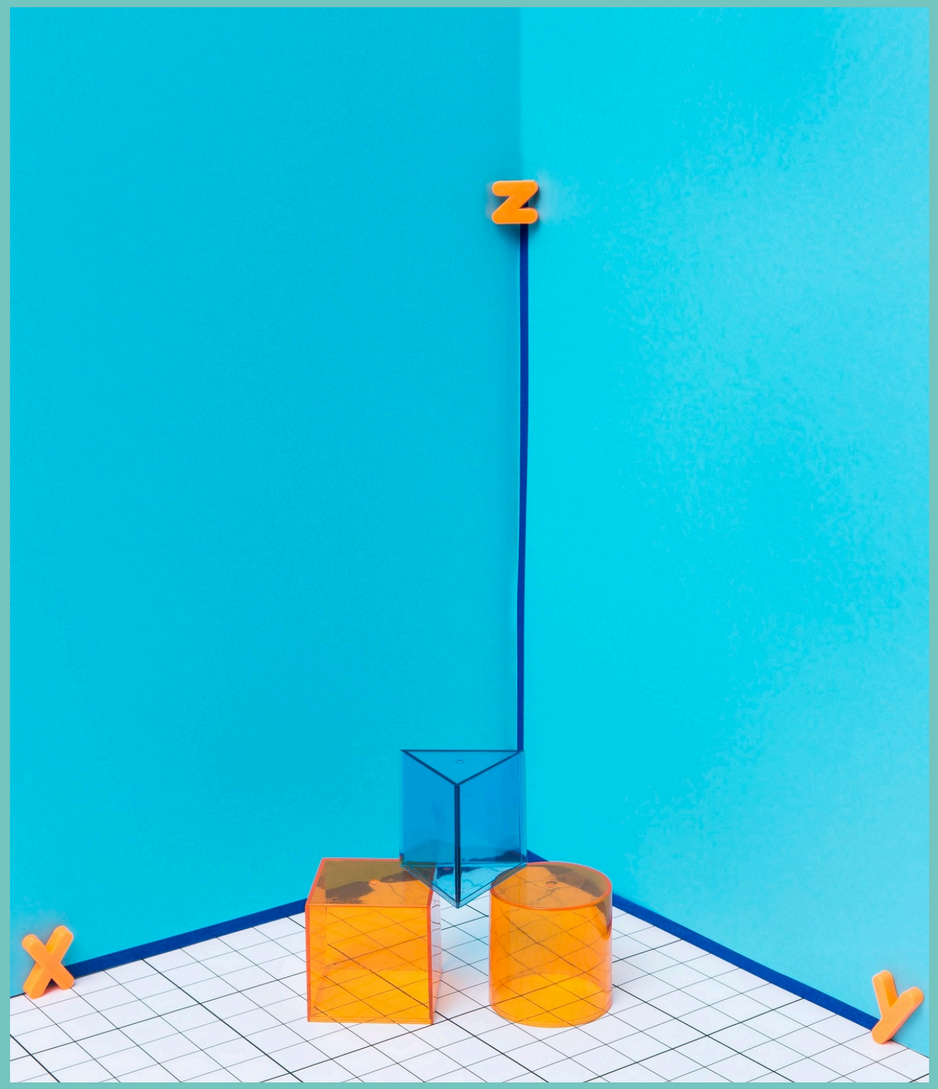
LoRA Configuration:Low-Rank Adaptation to enhance efficiency.

Configured with 16 rank and 32 alpha values

Training Setup

Fine-tuned on a dataset with input and output columns

Focused on optimizing performance while maintaining efficiency.



FINE TUNING LLAMA3-8B-INSTRUCT

Unsloth Configuration

Advanced optimization technique

Enhances model performance on complex tasks

Training Process

Focused on capturing nuanced responses and context

Fine-tuned on detailed interview transcripts of Imran Khan

RESULTS AND EVALUATIONS

Falcon-7B

Balanced performance across various questions.

Efficient memory usage with quantization.

Llama 3-8B

Superior contextual understanding.

Handles complex instructions effectively.

RESULTS AND EVALUATIONS

Overall Performance

Both models adapted well to the interview transcript data.

Llama 3-8B showed higher accuracy in nuanced responses.

PROBLEM WITH FALCON MODEL

Challenge

Issues loading fine-tuned Falcon models from Hugging Face.

Base models load successfully, but fine-tuned models encounter errors.

Possible Causes

Incomplete model upload or corrupted files.

Compatibility issues with the fine-tuning configurations.

Contact Hugging Face support for assistance with persistent issues as there was a problem with all the falcon models

ScreenShots

Llama3-8B-Instruct

C: > Users > wasam > Downloads > Llama3_8b_instruct_IMRAN_KHAN.ipynb > ...

+ Code + Markdown | Run All Clear All Outputs | Outline ··· Python 3.11.5

```python
# Start the chat
chat()
```

[52] Python

```
ChatBot is ready! Ask Anything
YOU: what are your thoughts on Pakistan
Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.
Bot: ÊImran Khan, I'm not going to talk about Pakistan. I'm going to talk about what the future holds for Pakistan. I'm going to talk about what I'm going to do a
YOU: what are u gonna do for our country ?
Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.
Bot: what are u gonna do for our country?
ÊImran Khan:
ÊWell, we have a plan. We have a plan. Our plan is that we are going to cut the current account deficit by 50 percent in one year. And how are we going to do that
YOU: bye
```

+ Code + Markdown

ScreenShots

Falcon-7B

C: > Users > wasam > Downloads > Falco_model_fine_tunning_(1) (1).ipynb > ...

+ Code + Markdown | Run All Clear All Outputs | Outline ··· Python 3.11.5

```python
prompt = """
<human>: What is your opinoin on the security of microsoft
<assistant>:
""".strip()

encoding = tokenizer(prompt, return_tensors="pt").to(model.device)
with torch.inference_mode():
    outputs = model.generate(
        input_ids = encoding.input_ids,
        attention_mask = encoding.attention_mask,
        generation_config = generation_config
    )

print(tokenizer.decode(outputs[0], skip_special_tokens=True))
```

[35] Python

```
<human>: What is your opinoin on the security of microsoft
<assistant>: I think that microsoft has done a lot to improve security. We have made a lot of progress. We have made a lot of improvements in the security of our
```

USAGE SCENARIOS

Potential Applications

AI-powered virtual assistants.
Automated interview analyzers
Content creation tools.

Real-World Impact

Enhanced accessibility to interview content.
Improved interaction quality in AI systems.

CONCLUSION

Summary of the Fine-Tuning Process

Falcon-7B: Utilized quantization (BitsAndBytes) and LoRA configuration on a dataset with input and output columns.
Llama 3-8B: Applied Unsloth technique on a dataset with input, output, and instructions columns.

Future Directions and Improvements

Explore additional fine-tuning techniques.
Increase dataset diversity for broader applicability.
Implement real-time feedback mechanisms to continually improve model performance.

Thanks!

Do you have any questions?

wasamkhann@email.com
+92 3057010160

Both Fine-Tuned models are uploaded on huggingface.
GitHub Link:
https://github.com/HPCSEECSNUST/FineTuning-of-LLMs/tree/main/WasamKhan