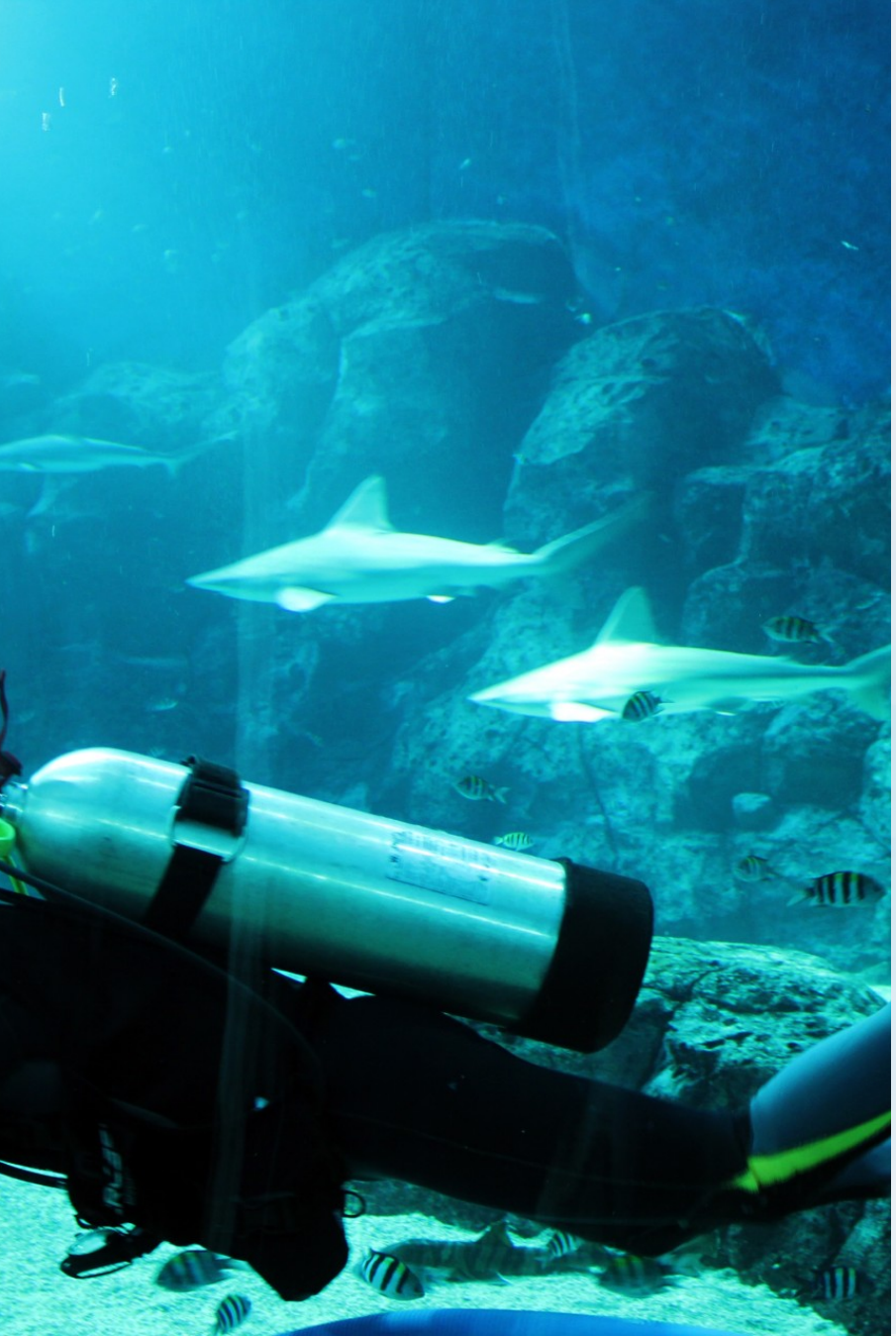


Deep Dive into Finetuning Models

CodeForward 2023 Conference

Link to Presentation:

https://bit.ly/deep-dive-finetuning-llms

by Kurt Niemi

About Your Presenter

Name: Kurt Niemi

Software Engineer / Machine Learning Engineer

Company: Sr Member Technical Staff @ Broadcom

Personal Email: kurtn718@gmail.com

LinkedIn: https://linkedin.com/in/kurtniemi

Link to this deck: https://bit.ly/deep-dive-finetuning-llms

Required Company Disclaimer: Everything in this presentation is my opinion only, and does not represent any official statement by my employer.

Agenda

Brief discussion of trying Prompting Techniques / RAG first

When should we fine-tune an LLM

When should we NOT fine-tune an LLM

What is fine-tuning?

Challenges in fine-tuning an LLM

Demo

Q&A

Why try Prompting Techniques first?

Maximizes Model Efficiency

Cost and Time Efficiency

Iterate more quickly

Reduced Dependency on External Resources

Developer/Engineering expertise not required - Anybody who knows the business domain and English can do prompting

RAG (Retrieval Augmented Generation) - Reasons to Use

Combining Retrieval with Generation for Enhanced Performance

Faster Deployment and Iteration

Allows one to understand baseline performance

Avoid overfitting LLM

When should we finetune an LLM

- For a specific task we want the best results
- Reduce latency and save money
- We want a smaller model - with less parameters - trained to imitate a larger model
- Because it's fun 😀 Or we want to learn how to further our expertise in Generative AI. 😀

When should we not finetune an LLM

Reducing hallucinations

Model will still hallucinate

Use RAG instead

Limited Data

Need labeled instruction response pairs

Need hundreds to thousands of examples

Low Quality Data

Garbage in - garbage out

Information changes frequently

Versions, time sensitive information

Would require frequent finetuning

Lack of Infrastructure or Money

to finetune the LLM

to host the LLM

Simple or previously trained on task

Creative writing

Summarization

Classification

Demo

Will show Jupyter notebook where we finetune a Mistral 7B Model using LudwigAI - and highlight some of the benefits along the way.

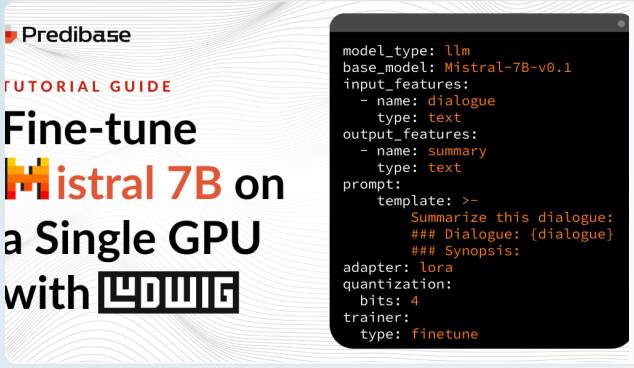
Additional Resources

Ludwig AI : https://ludwig.ai/latest/

Nvidia presentation 11-17-2023 on Tailoring LLMs to your Use case

https://www.nvidia.com/en-us/on-demand/session/llmdevday23-02

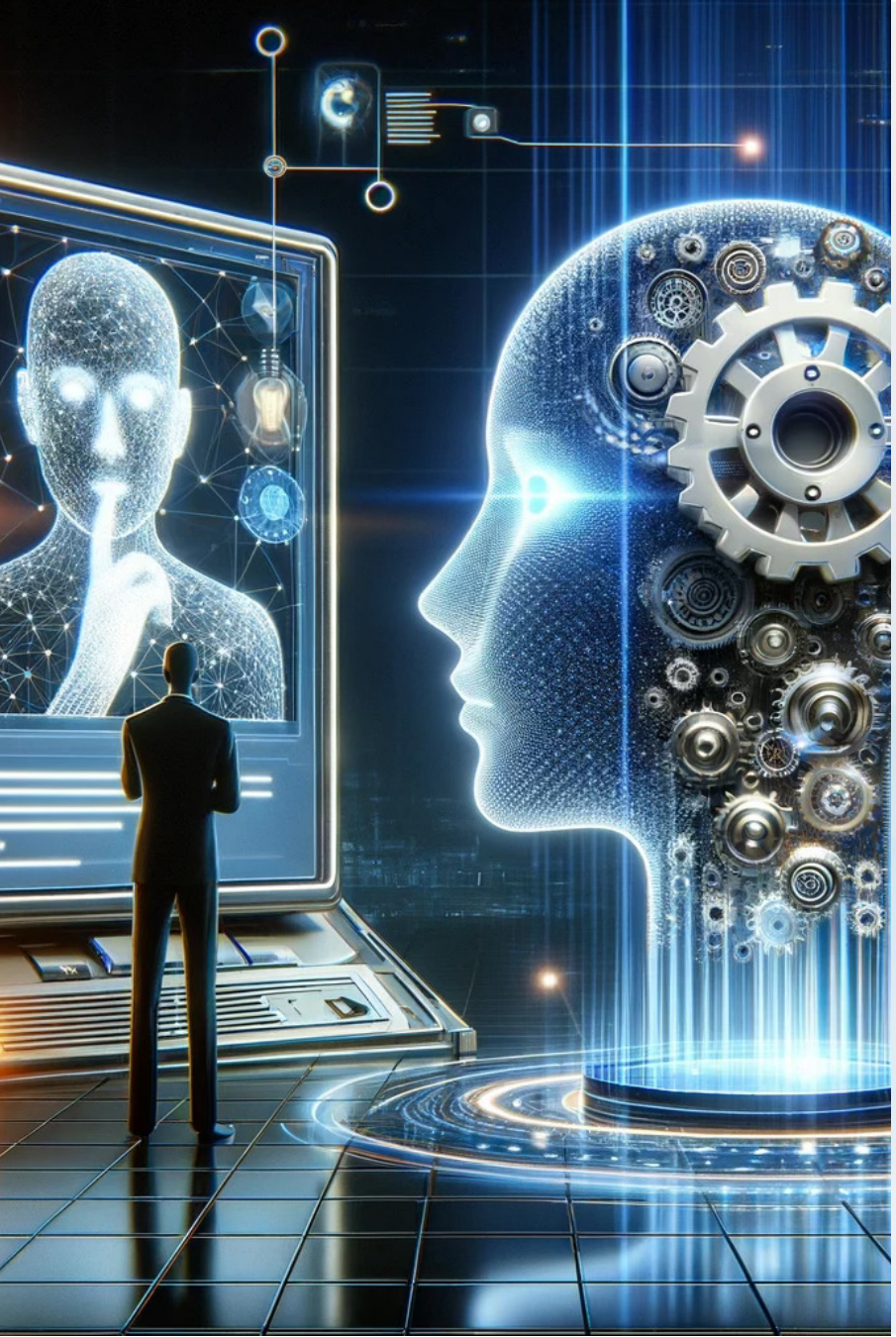
Finetuning Mistral 7B on a Single GPU with Ludwig



predibase

Fine-tuning Mistral 7B on a Single GPU with Ludwig -...

Learn how to fine-tune the recently released open-source LLM from Mistral AI (“Mistral 7B”) for a summarization task on single...



Q&A

Contact Info: Keep in touch!

LinkedIn: https://linkedin.com/in/kurtniemi

Personal Email: kurtn718@gmail.com

Link to this deck: https://bit.ly/deep-dive-finetuning-llms