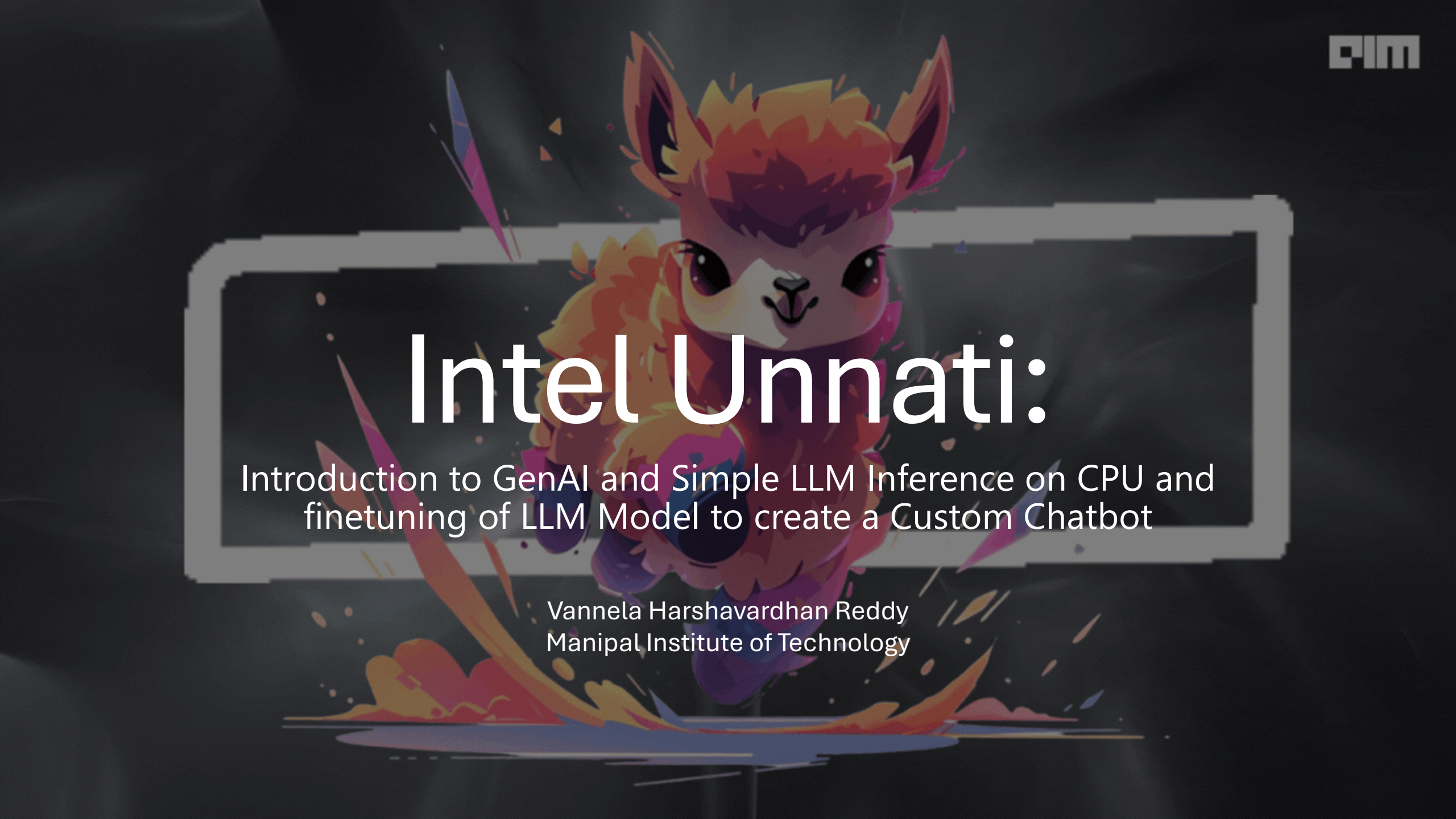


Intel Unnati:

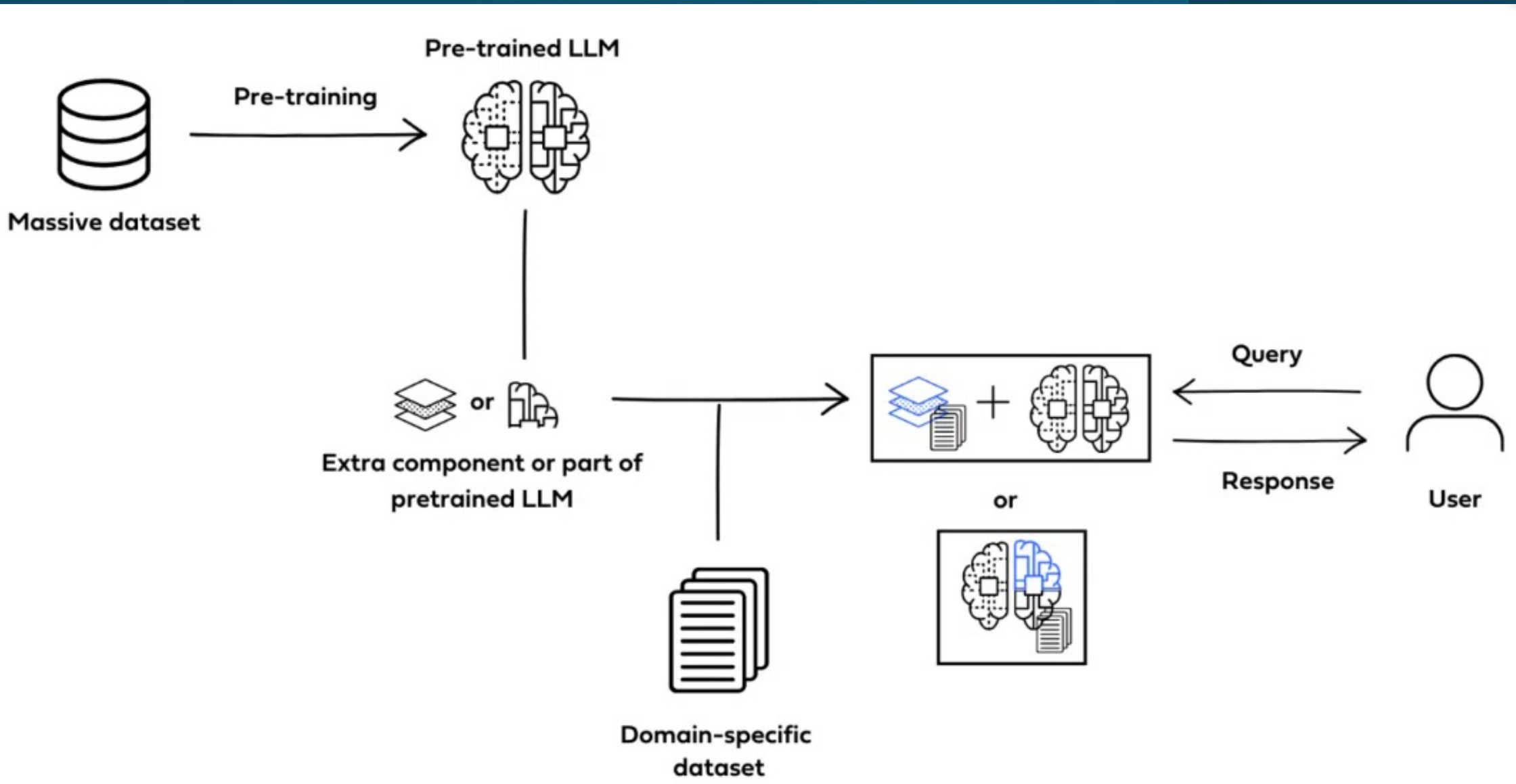
Introduction to GenAI and Simple LLM Inference on CPU and finetuning of LLM Model to create a Custom Chatbot

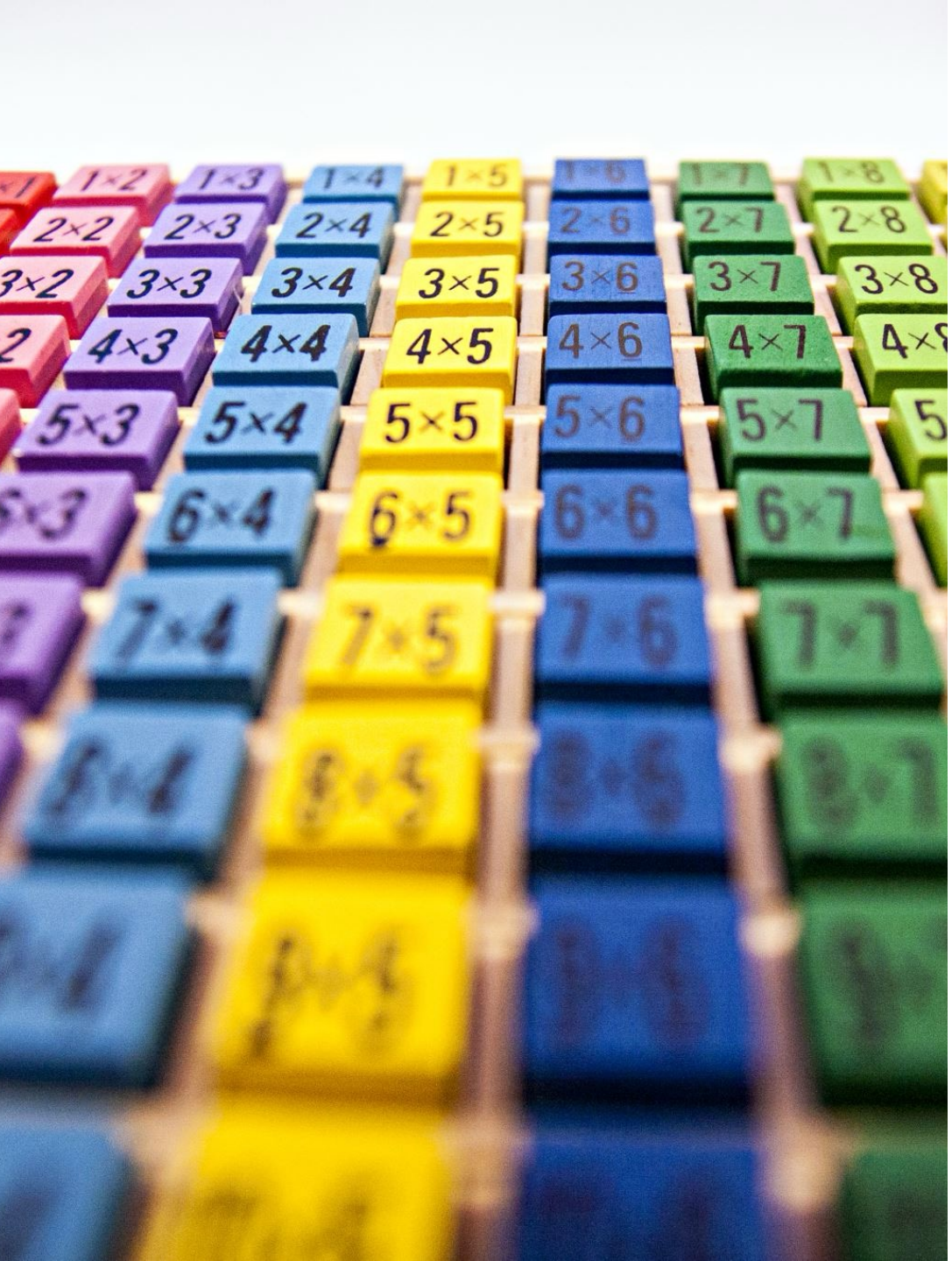
Vannela Harshavardhan Reddy
Manipal Institute of Technology

Introduction to Finetuning

Finetuning in machine learning involves further training a pretrained model on a new dataset to adapt its parameters for a specific task or domain:

- The pretrained model, initially trained on a general dataset like ImageNet or Wikipedia, learns relevant features.
- During finetuning, the model's weights adjust using the new dataset, which usually has fewer labeled examples.
- This process leverages the pretrained model's knowledge and tailors it to the new data's nuances.
- Initialization with pretrained weights and subsequent updates via backpropagation minimize task-specific loss.
- Finetuning is efficient, requiring less data and computation than training from scratch, and enhances model accuracy and efficiency in specialized tasks like computer vision and natural language processing.


Pre-trained LLM
Pre-training
Massive dataset
or
Extra component or part of pretrained LLM
Domain-specific dataset
or
Query
Response
User



Low Rank Adaptation of LLM's (LoRA)

- LoRA (Low Rank Adaptation of LLM's)** is an advanced technique in machine learning.
- It specifically targets pretrained language models (LLMs) by incorporating low-rank approximations.
- The primary aim is to enhance the efficiency and effectiveness of these models, especially in natural language processing (NLP) tasks.
- By applying low-rank adaptations, LoRA optimizes model performance while potentially reducing computational complexity.
- This approach is particularly valuable in environments where computational resources are limited or where efficiency in model inference is crucial.

Project Overview

- **Objective:** The objective of this project is to develop a Generative AI model capable of generating colour codes based on textual descriptions of colours. This project leverages the TinyLLama framework and utilizes LoRA (Low Rank Adaptation) for finetuning a pretrained model. The model aims to accurately predict RGB colour codes from natural language inputs describing colours.
- This will be a low-resource dependent application as TinyLlama does not require huge compute power and can be run locally on end devices as well.

Tech Used

PYTHON

HUGGING FACE

GOOGLE COLAB

Conclusion

- In conclusion, this project successfully developed a Generative AI model using TinyLLama and LoRA finetuning to generate color codes from textual descriptions.
- By leveraging advanced natural language processing techniques and optimization strategies, the model accurately predicts color codes, enhancing applications in digital design and creative industries.
- Moving forward, further refinements could focus on expanding the dataset diversity and exploring additional color code formats, ensuring broader applicability and precision in color generation tasks.