Day 25: Advanced Fine-Tuning Techniques

- **Introduction**
  - We already performed some basic fine-tuning in our previous sessions
  - Pre-trained models from Transformers library --> fine-tuned them by training them on some external datasets
  - Let's understand fine-tuning a little better and explore some new techniques for the same
  - Also we'll try and understand some inner details and intricacies
- **Fine-tuning**
  - **Single line definition:**
    - "Taking a pre-trained model (trained on large-scale data) and adapting it for a more specific task using a smaller, specialized dataset"
    - Specialize in a niche domain or specific type of task, such as medical reports, legal documents, or creative writing
  - **Why?**
    - Domain-specific knowledge: GPT-3 can be adapted to specific industries or tasks (e.g., writing financial reports).
    - Improved performance: Fine-tuning on task-specific data typically improves performance on that particular task.
    - Customization: You can control the output of the model, tailoring it to fit your needs better.
- **Transfer Learning Techniques**
  - A core concept that enables fine-tuning
  - Using knowledge learned by a model from one task (pre-training) and applying it to another related task
  - Transfer learning for generative models --> build on the general language understanding of the model --> apply it to specific tasks.
  - **Two Key Methods in Transfer Learning:**
    - Feature extraction
      - Use the pre-trained model as the feature extractor
      - Freeze all the layers except the final few layers
      - Use the output as a feature for a new task
    - Full Fine-tuning
      - Update the weights of the entire model during training on the task-specific data
      - Smaller learning rate than in initial pre-training
- **Hyperparameter Tuning for Fine-Tuning**
  - To optimize a fine-tuned model's performance
  - Hyperparameters are values like learning rate, batch size, and the number of training epochs
  - **Key Hyperparameters for Fine-Tuning Generative Models:**
    - Learning Rate
      - Lower learning rate is often required during fine-tuning
      - Beacause we want to make subtle adjustments to the model without overwriting the knowledge it already learned
    - Batch Size
      - How much data is processed at once
      - Smaller batch size is often used for fine-tuning to avoid overwhelming the model
    - Epochs
      - Number of passes through the dataset during training
      - For fine-tuning, fewer epochs (often around 3-5) may be sufficient
    - Warmup Steps
      - Gradually increasing the learning rate at the start can stabilize training, especially for large models
    - Weight Decay
      - A regularization technique to prevent overfitting
- **Hands-on: Fine-tune davinci-002 on a specialized dataset**