Summary and Highlights

Congratulations! You have completed this module. At this point in the course, you know the following:

- Direct Preference Optimization (or DPO) is a reinforcement learning technique designed to fine-tune models based on human preferences more directly and efficiently than traditional methods.
- DPO involves collecting data on human preferences by showing users different outputs from the model and asking them to choose the better one.
- DPO involves three models: the reward function, which uses an encoder model, the target decoder, and the reference model.
- In DPO, you can convert a complex problem into a simpler objective function that is more straightforward to optimize.
- Two main steps to fine-tuning a language model with DPO:
    - Data collection
    - Optimization
- Steps to fine-tune a language model with DPO and Hugging Face:
    - Step 1: Data preprocessing
        - Reformat
        - Define and apply the process function
        - Create the training and evaluation sets
    - Step 2: Create and configure the model and tokenizer
    - Step 3: Define training arguments and DPO trainer
    - Step 4: Plot the model's training loss
    - Step 5: Load the model
    - Step 6: Inferencing
- DPO leverages a closed-form optimal policy as a function of the reward to reformulate the problem
- Reward policy:

$$\pi_r(Y|X) = \frac{\pi_{ref}(Y|X) \exp\left(\frac{1}{\beta} r(X,Y)\right)}{Z(X)}$$

- Subtracting the reward model for two samples eliminates the need for the partition function

$$r(X, Y_w) - r(X, Y_l) = \beta \ln\left(\frac{\pi_r(Y_w|X)}{\pi_{ref}(Y_w|X)}\right) - \beta \ln\left(\frac{\pi_r(Y_l|X)}{\pi_{ref}(Y_l|X)}\right)$$

- Loss function:

$$-\sigma\left(\beta \ln\left(\frac{\pi_r(Y_w|X)}{\pi_{ref}(Y_w|X)}\right) - \beta \ln\left(\frac{\pi_r(Y_l|X)}{\pi_{ref}(Y_l|X)}\right)\right)$$