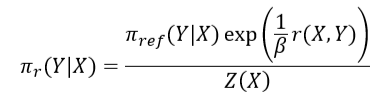
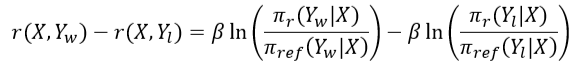
**Summary and Highlights**

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:



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



* Loss function:

A black square with a number of square and a number of square

AI-generated content may be incorrect.