← Back Fine-Tuning Causal LLMs with Human Feedback and Direct Preference

Graded Assignment • 30 min

English ⌄ **Due** Jan 22, 11:59 PM IST

Your grade: 70%

Your latest: **70%** • Your highest: **70%** • To pass you need at least 70%. We keep your highest score.

Next item →

1. ______ controls the randomness of the distribution for the large language models (LLMs). 1 / 1 point

○ Min max tokens

○ Repetition penalty

○ Top K sampling

● Temperature

✓ **Correct**
Temperature plays an important role in controlling the randomness in LLMs. As temperature increases, the out would be more diverse and creative; however, the LLMs will deliver moderate and deterministic output as the temperature decreases.

2. Rollout is referred to as ___. 1 / 1 point

○ How to evaluate the performance of the policy by interaction with other departments

● How the model generates various responses for each query

○ How policy use the rewards collected during training

○ How to initiate policy before training

✓ **Correct**
Upon inserting any query, the model generates possible responses with queries to continue with the process.

3. Which of the following is the most suitable for ensuring the model learns to generate responses that align with human preferences while fine-tuning a pretrained large language model (LLM) using reinforcement learning with human feedback (RLHF)? 1 / 1 point

● Fine-tune the pretrained LLM on a reward model that has been trained on human feedback.

○ Allow the LLM to generate responses without human feedback and select the most preferable outputs using a Human.

○ Train the LLM from scratch using reinforcement learning signals.

○ Use supervised learning to train the LLM on human feedback.

✓ **Correct**
Review the Reinforcement Learning from Human Feedback (RLHF) video.

4. In proximal policy optimization (PPO), the log-derivative trick computes the gradient of the objective function with respect to the policy parameters from sampling. Which of the following expressions is correct to use in a log-derivative trick to maximize the PPO objective function from samples of the policy? 0 / 1 point

● $\hat{\theta} = \arg max_\theta \; [\sum_Y r(X,Y)\, \pi_\theta(Y|X)]$

○ $E\;[r_Y|\theta] = \; E_{Y\sim\pi_\theta(Y|X)}[r(X,Y)]$

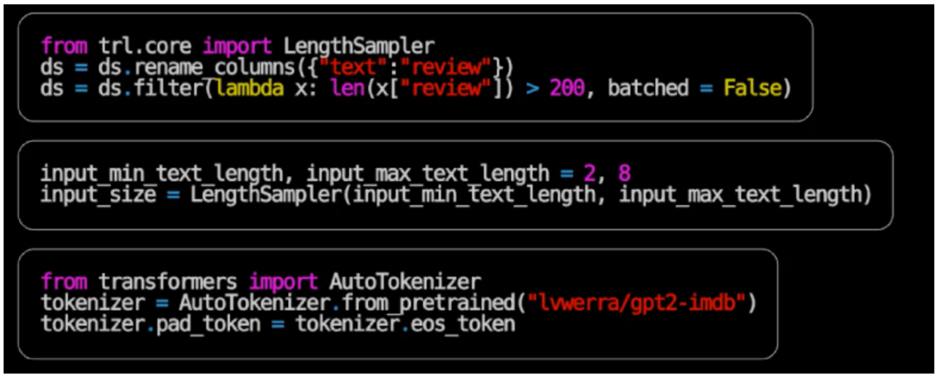
○ $E\;[r_Y|\theta] = \sum_Y r(X,Y)\, \pi_\theta(Y|X)$

○ $\nabla_\theta E[r_Y|\theta] \; = \; \sum_Y r(X,Y)\, \nabla_\theta \pi_\theta(Y|X)$

Incorrect
Review the Proximal Policy Optimization (PPO) video.

5. In the following code snippet, which helps to vary text lengths for data processing?

1 / 1 point



```
from trl.core import LengthSampler
ds = ds.rename_columns({"text":"review"})
ds = ds.filter(lambda x: len(x["review"]) > 200, batched = False)
```

```
input_min_text_length, input_max_text_length = 2, 8
input_size = LengthSampler(input_min_text_length, input_max_text_length)
```

```
from transformers import AutoTokenizer
tokenizer = AutoTokenizer.from_pretrained("lvwerra/gpt2-imdb")
tokenizer.pad_token = tokenizer.eos_token
```

○ tokenizer.pad_token

● LengthSampler

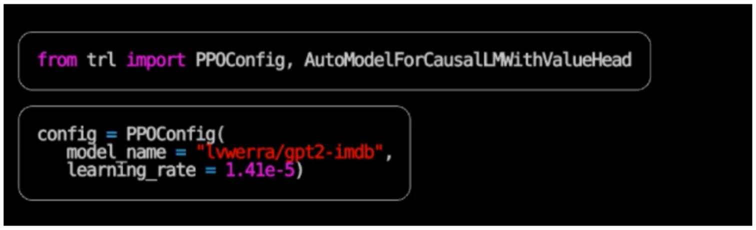
○ input_min_text_length

○ input_size

Correct
LengthSampler enhances model robustness and simulates realistic training conditions. It ensures efficient training by managing text input lengths and maintaining performance.

6. In the following code snippet, what is the use of the PPOConfig class?

1 / 1 point



```
from trl import PPOConfig, AutoModelForCausalLMWithValueHead
```

```
config = PPOConfig(
    model_name = "lvwerra/gpt2-imdb",
    learning_rate = 1.41e-5)
```

○ The PPOConfig class prepares data batches in a format suitable for the PPOTrainer.

● The PPOConfig class is useful for specifying the model and learning rate for PPO training.

○ The PPOConfig class configures the settings for PPO training.

○ The PPOConfig class trains the model using Kullback-Leibler (KL) divergence.

Correct
The PPOConfig class initializes the PPO configuration, and the model name specifies the model that needs to be fine-tuned.

7. What is the significance of β in the following equation?

0 / 1 point

$$\pi_*(X,Y) = \arg\max_{\pi}\left\{E_{X\sim D}\left[E_{Y\sim\pi_\theta(Y|X)}[r(X,Y)] - \beta D_{KL}[\pi_\theta(Y|X)\|\pi_{ref}(Y|X)]\right]\right\}$$

○ β parameter helps obtain state-of-the-art results through reward-free methods.

○ β parameter helps develop a reward model that uses actor-critic algorithms.

● β parameter helps ensure that the total probability is one for each query.

○ β parameter helps ensure that the total probability is one for each query.

○ β parameter helps control the deviation from the base reference policy.

Incorrect
Review the DPO: Partition Function video.

8. Why is it impractical to compute the partition function given by the following expression?

1 / 1 point

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

● The exponential growth in the number of terms makes the partition function increasingly difficult to compute as the length of sequences increases.

○ Multiplying the partition function by a scalar scales it by *c*, resulting in a new partition function that is difficult and impractical to calculate.

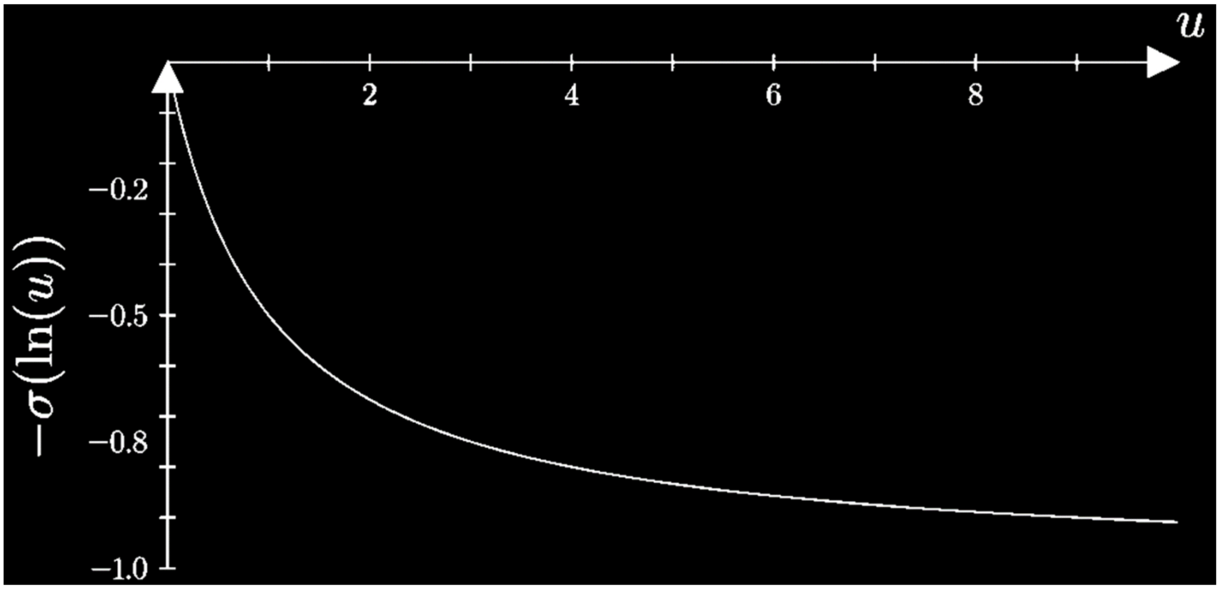
○ When expressing the partition function as a logarithm, the expression cannot be combined with other logarithmic functions, making it impossible to compute the partition function.

○ The optimal solution scales the partition function to the reward function, and with the beta parameter controlling the constant, it makes it difficult to compute the partition function.

Correct
For a sequence length of one, the partition function sums over all words in your vocabulary, *V*. Next, for a sequence length of two, Z(*X*) sums over all possible pairs of words in your vocabulary, creating a much larger set, V^2. Finally, generalizing for a sequence length of *T*, *Z*(*X*) sums over all possible sequences of length *T* in your vocabulary, V^T. This exponential growth in the number of terms makes the partition function increasingly difficult to compute as *T* increases.

9. Consider the following plot of the loss as a function of u, where u represents the ratio of probabilities of the positive sample and the negative sample.

0 / 1 point



Which of the following options is correct regarding the plot?

○ The model's loss is lower when $0 < u < 1$.

● The model's loss is higher when $u > 1$.

○ As u increases, the model gets better.

○ As the probability of the winning response decreases, the loss also decreases.

Incorrect
Review the From Optimal Policy to DPO video.

10. Identify which of the following codes is the first step to load, create and configure the model you are going to train for your task.

1 / 1 point

○ tokenizer = GPT2Tokenizer.from_pretrained(“gpt2”)

○ tokenizer.pad_token = tokenizer.eos_token

○ model_ref = AutoModelForCausalLM.from_pretrained(“gpt2”)

● model = AutoModelForCausalLM.from_pretrained(“gpt2”)

✓ **Correct**

The steps required to create and configure the model and tokenizer for your task begin by loading the decoder GPT-2 model. This is done using the “AutoModelForCausalLM” class from the Hugging Face Transformers library.