## Your grade: 85.71%

Your latest: **85.71%** • Your highest: **85.71%** • To pass you need at least 70%. We keep your highest score.

Next item

1. In large language models (LLMs), which of the following is true for the next token 't+1' distribution based on the current token 't'? 1 / 1 point

- Determine token 't+1' solely with respect to the specific token 't', without considering any prior context
- (selected) Condition token 't+1' on the current token 't' and the entire preceding sequence of tokens
- Random token 't+1' selection from the possible tokens, independent of 't'
- Uniform distribution of token 't+1' with a unique probability of getting selected

**Correct**
In large language models (LLMs), the prediction of the next token 't+1' depends on the current token 't,' allowing models to generate coherent and contextually appropriate responses.

2. Regarding the language models, how does a policy relate to the model's distribution for the possible outputs? 1 / 1 point

- A policy distribution represents all possible outputs based on the predefined rules.
- (selected) A policy for distributing tokens conditioned on a given set of query tokens.
- A policy is the same as the probability distribution, representing all possible output with the same probability.
- A policy is a deterministic set of output.

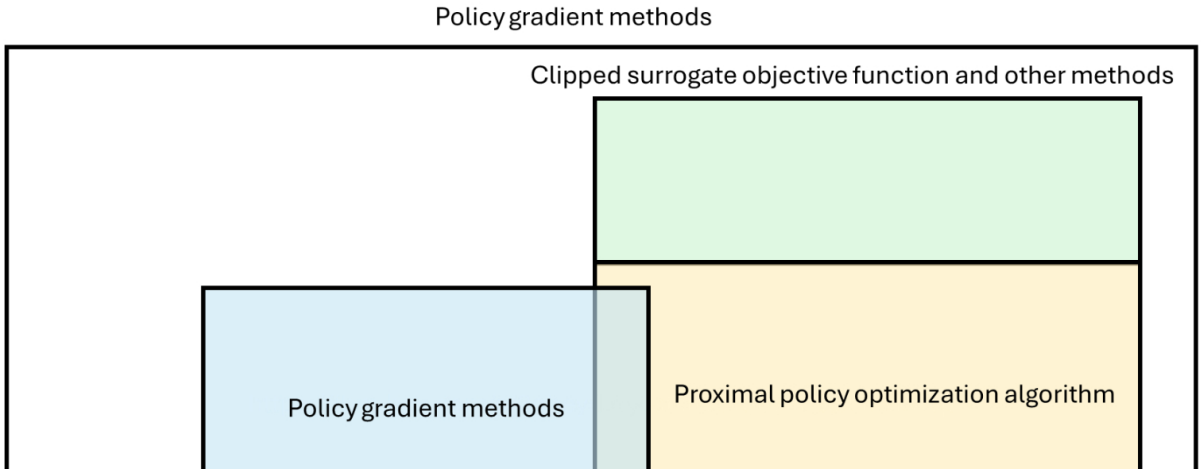
**Correct**
The policy usually dictates the model to select specific output from this distribution, balancing between selecting the most appropriate output.

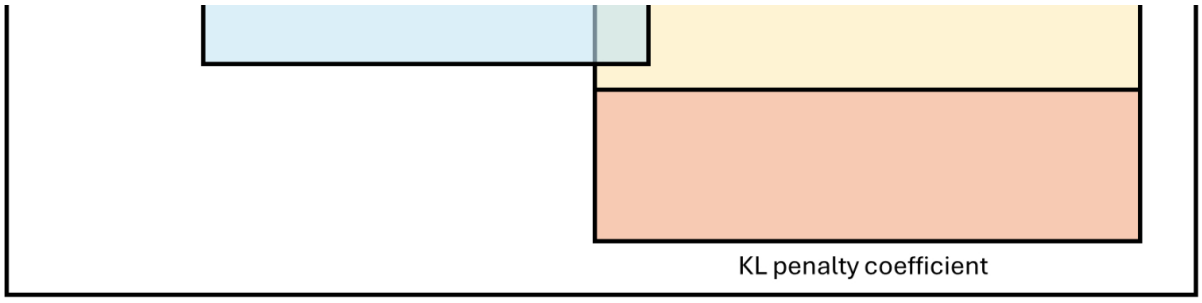
3. In reinforcement learning with human feedback (RLHF), which of the following is true to guide the model's learning process using the concept of expected rewards? 1 / 1 point

- The model randomly selects actions by considering expected rewards or human feedback
- The model selects outputs to maximize the sum of possible rewards for each token.
- (selected) The model selects outputs that maximize the expected reward and prioritizes actions so that the model receives positive human feedback.
- The model ignores human feedback and relies on a predefined reward function to determine the expected reward.

**Correct**
In RLHF, the model maximizes the expected rewards by adjusting its actions according to human feedback. The positive human feedback helps the model to identify the most likely actions and update the policy to increase the positive likelihood of selecting those actions.

4. In the Venn diagram below, what is the role of the KL penalty coefficient in the policy gradient methods? 0 / 1 point






- (●) Apply trust region constraint to limit the policy from deviating too far from the previous policy
- ( ) Adjusting the learning rate during training, ensuring that the policy update steps are appropriately sized
- ( ) Restricting the policy updates to a small region around the old policy
- ( ) Regulating the divergence between the old and the new policies

⊗ **Incorrect**
PPO ignores the trust region constraints and encourages clipped surrogate objective function to approximate the effects of the trust region.

5. Which of the following code snippets indicates the sentiment analysis pipeline's scoring function to evaluate the generated responses' quality or relevance? 1 / 1 point

(●)
```
score_1 = pipe_outputs[0][1]
score_2 = pipe_outputs[1][1]
print("Score for first response:", score_1)
print("Score for second response:", score_2)
```
```
Score for first response: {'label': 'POSITIVE', 'score': -2.0565342903137207}
Score for second response: {'label': 'POSITIVE', 'score': -2.7069687843322754}
```
```
rewards = [torch.tensor(output[1]["score"]) for output in pipe_outputs]
```
```
rewards:
[tensor(-2.0565), tensor(-2.7070)]
```

( )
```
from datasets import load_dataset
```
```
dataset_name = "imdb"
ds = load_dataset(dataset_name, split = "train")
```

( )
```
def tokenize(sample):
    sample["input_ids"] = tokenizer.encode(sample["review"])[: input_size()]
    sample["query"] = tokenizer.decode(sample["input_ids"])
    return sample
```
```
tokenize(ds[2]):

{'review': If only to avoid making this type of film in the future. This film is
interesting as an experiment but tells no cogent story. One might feel virtuous
for sitting thru it because it touches on so many IMPORTANT issues but it does
so without any discernable motive. The viewer comes away with no new perspectives
(unless one comes up with one while one's mind wanders, as it will invariably do
during this pointless film).One might better spend one's time staring out a window
at a tree growing.,
'label': tensor(0),
'input_ids': tensor([1532,  691,  284, 3368]),
```


- ○ 

```
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
```

**Correct**
This code snippet shows the model's confidence in the likelihood of generating positive responses in the sentiment analysis pipeline.

6. What would happen if you set the sentiment score to 0 in the given code snippet for the PPO trainer? 1 / 1 point

```
all_stats = []

change_score = 1
```

- ○ Setting the sentiment score to 0 loops the code for the PPO algorithm.
- ○ Setting the sentiment score to 0 rewards positive sentiment more.
- ◉ Setting the sentiment score to 0 increases the chances of generating higher rewards for negative sentiment.
- ○ Setting the sentiment score to 0 increases the PPO mean reward over time.

**Correct**
The model generates negative responses when the sentiment score is set to zero.

7. Which of the following statements correctly describes the computation of the expected rewards? 1 / 1 point

- ○ Expected reward is defined as the sum of the immediate rewards for the particular action, averaged over all possible states.
- ○ The expected reward is obtained from the best action evaluated by human feedback without considering future rewards.
- ○ The expected reward is defined as the reward obtained from all the possible actions in the given state, regardless of the action taken.
- ◉ Expected reward is the weighted sum of the future rewards, predicted by the reward model, given the current state and action.

**Correct**
In RLHF, the expected reward involves predicting future rewards based on the current state and action, weighted by their probabilities.