LLM Course
Search documentation ⌘K
EN 3,566
0. SETUP
1. TRANSFORMER MODELS
2. USING 🤗 TRANSFORMERS
3. FINE-TUNING A PRETRAINED MODEL
4. SHARING MODELS AND TOKENIZERS
5. THE 🤗 DATASETS LIBRARY
6. THE 🤗 TOKENIZERS LIBRARY
7. CLASSICAL NLP TASKS
Introduction
Token classification
Fine-tuning a masked language model
Translation
Summarization
Training a causal language model from scratch
Question answering
Mastering LLMs
End-of-chapter quiz
8. HOW TO ASK FOR HELP
9. BUILDING AND SHARING DEMOS
10. CURATE HIGH-QUALITY DATASETS
11. FINE-TUNE LARGE LANGUAGE MODELS
12. BUILD REASONING MODELS NEW
COURSE EVENTS

Pytorch

Ask a question

# End-of-chapter quiz

Copy page

Let's test what you learned in this chapter!

### 1. Which of the following tasks can be framed as a token classification problem?

- [x] Find the grammatical components in a sentence.

**Correct!** Correct! We can then label each word as a noun, verb, etc.

- [ ] Find whether a sentence is grammatically correct or not.
- [x] Find the persons mentioned in a sentence.

**Correct!** Correct! We can label each word as person or not person.

- [ ] Find the chunk of words in a sentence that answers a question.

Submit **You got all the answers!**

### 2. What part of the preprocessing for token classification differs from the other preprocessing pipelines?

- [ ] There is no need to do anything; the texts are already tokenized.
- [x] The texts are given as words, so we only need to apply subword tokenization.

**Correct!** Correct! This is different from the usual preprocessing, where we need to apply the full tokenization pipeline. Can you think of another difference?

- [ ] We use `-100` to label the special tokens.
- [x] We need to make sure to truncate or pad the labels to the same size as the inputs, when applying truncation/padding.

**Correct!** Indeed! That's not the only difference, though.

Submit **You got all the answers!**

### 3. What problem arises when we tokenize the words in a token classification problem and want to label the tokens?

- [ ] The tokenizer adds special tokens and we have no labels for them.
- [x] Each word can produce several tokens, so we end up with more tokens than we have labels.

**Correct!** That is the main problem, and we need to align the original labels with the tokens.

- [ ] The added tokens have no labels, so there is no problem.

Submit **You got all the answers!**

### 4. What does "domain adaptation" mean?

- [ ] It's when we run a model on a dataset and get the predictions for each sample in that dataset.
- [ ] It's when we train a model on a dataset.
- [x] It's when we fine-tune a pretrained model on a new dataset, and it gives predictions that are more adapted to that dataset

**Correct!** Correct! The model adapted its knowledge to the new dataset.

- [ ] It's when we add misclassified samples to a dataset to make our model more robust.

Submit **You got all the answers!**

### 5. What are the labels in a masked language modeling problem?

- [x] Some of the tokens in the input sentence are randomly masked and the labels are the original input tokens.

**Correct!** That's it!

- [ ] Some of the tokens in the input sentence are randomly masked and the labels are the original input tokens, shifted to the left.
- [ ] Some of the tokens in the input sentence are randomly masked, and the label is whether the sentence is positive or negative.
- [ ] Some of the tokens in the two input sentences are randomly masked, and the label is whether the two sentences are similar or not.

Submit **You got all the answers!**

### 6. Which of these tasks can be seen as a sequence-to-sequence problem?

- [x] Writing short reviews of long documents

**Correct!** Yes, that's a summarization problem. Try another answer!

- [x] Answering questions about a document

**Correct!** This can be framed as a sequence-to-sequence problem. It's not the only right answer, though.

- [x] Translating a text in Chinese into English

**Correct!** That's definitely a sequence-to-sequence problem. Can you spot another one?

- [x] Fixing the messages sent by my nephew/friend so they're in proper English

**Correct!** That's a kind of translation problem, so definitely a sequence-to-sequence task. This isn't the only right answer, though!

Submit **You got all the answers!**

### 7. What is the proper way to preprocess the data for a sequence-to-sequence problem?

- [ ] The inputs and targets have to be sent together to the tokenizer with `inputs=...` and `targets=...`.
- [ ] The inputs and the targets both have to be preprocessed, in two separate calls to the tokenizer.
- [ ] As usual, we just have to tokenize the inputs.
- [x] The inputs have to be sent to the tokenizer, and the targets too, but under a special context manager.

**Correct!** That's correct, the tokenizer needs to be put into target mode by that context manager.

Submit **You got all the answers!**

### 8. Why is there a specific subclass of Trainer for sequence-to-sequence problems?

- [ ] Because sequence-to-sequence problems use a custom loss, to ignore the labels set to `-100`
- [x] Because sequence-to-sequence problems require a special evaluation loop

**Correct!** That's correct. Sequence-to-sequence models' predictions are often run using the `generate()` method.

- [ ] Because the targets are texts in sequence-to-sequence problems
- [ ] Because we use two models in sequence-to-sequence problems

Submit **You got all the answers!**

### 10. When should you pretrain a new model?

- [x] When there is no pretrained model available for your specific language

**Correct!** That's correct.

- [ ] When you have lots of data available, even if there is a pretrained model that could work on it
- [x] When you have concerns about the bias of the pretrained model you are using

**Correct!** That is true, but you have to make very sure the data you will use for training is really better.

- [ ] When the pretrained models available are just not good enough

Submit **You got all the answers!**

### 11. Why is it easy to pretrain a language model on lots and lots of texts?

- [ ] Because there are plenty of texts available on the internet
- [x] Because the pretraining objective does not require humans to label the data

**Correct!** That's correct, language modeling is a self-supervised problem.

- [ ] Because the 🤗 Transformers library only requires a few lines of code to start the training

Submit **You got all the answers!**

### 12. What are the main challenges when preprocessing data for a question answering task?

- [ ] You need to tokenize the inputs.
- [x] You need to deal with very long contexts, which give several training features that may or may not have the answer in them.

**Correct!** This is definitely one of the challenges.

- [ ] You need to tokenize the answers to the question as well as the inputs.
- [x] From the answer span in the text, you have to find the start and end token in the tokenized input.

**Correct!** That's one of the hard parts, yes!

Submit **You got all the answers!**

### 13. How is post-processing usually done in question answering?

- [ ] The model gives you the start and end positions of the answer, and you just have to decode the corresponding span of tokens.
- [ ] The model gives you the start and end positions of the answer for each feature created by one example, and you just have to decode the corresponding span of tokens in the one that has the best score.
- [x] The model gives you the start and end positions of the answer for each feature created by one example, and you just have to match them to the span in the context for the one that has the best score.

**Correct!** That's it in a nutshell!

- [ ] The model generates an answer, and you just have to decode it.

Submit **You got all the answers!**

Update on GitHub

← Mastering LLMs

Complete Chapter

End-of-chapter quiz
1. Which of the following tasks can be framed as a token classification problem?
2. What part of the preprocessing for token classification differs from the other preprocessing pipelines?
3. What problem arises when we tokenize the words in a token classification problem and want to label the tokens?
4. What does "domain adaptation" mean?
5. What are the labels in a masked language modeling problem?
6. Which of these tasks can be seen as a sequence-to-sequence problem?
7. What is the proper way to preprocess the data for a sequence-to-sequence problem?
8. Why is there a specific subclass of Trainer for sequence-to-sequence problems?
9. Why is it often unnecessary to specify a loss when calling compile() on a Transformer model?
10. When should you pretrain a new model?
11. Why is it easy to pretrain a language model on lots and lots of texts?
12. What are the main challenges when preprocessing data for a question answering task?
13. How is post-processing usually done in question answering?