1. Which of the following are true about pre-training in NLP? 1 / 1 point

- [x] It speeds training.

  Correct
  Correct.

- [x] It allows you to use information learned from a different task while working on a specific task.

  Correct
  Correct.

- [ ] It is not recommended because it takes a long time to pre-train a model.

- [x] It allows you to get better results.

  Correct
  Correct.

2. What is fine-tuning in NLP? 1 / 1 point

- (•) Fine tuning means taking existing weights of deeplearning model, and tweaking them a little bit to get a desired output, usually better results, on some specific task.
- ( ) Fine tuning means taking existing weights from a deeplearning model, let's say word embeddings, and then using those weights in another model as they are without changing them.
- ( ) Fine-tuning slows down your training.
- ( ) Fine tuning allows you to better prepare your data for training.

  Correct
  Correct.

3. Select all that apply for Masked Language Modeling. (MLM) 1 / 1 point

- [x] The cross entropy loss over V classes is used when predicting.

  Correct
  Correct.

- [x] The goal is to predict the masked token.

  Correct
  Correct.

- [x] The cross entropy loss over V classes is used when predicting.

Correct
Correct.

- [x] The goal is to predict the masked token.

Correct
Correct.

- [ ] There could only be one masked span in a sentence.
- [x] Choose 15% of the tokens at random: mask them 80% of the time, replace them with a random token 10% of the time, or keep as is 10% of the time.

Correct
Correct.

4. What does the BERT objective consist of? 1 / 1 point

- ○ It consists of a binary loss for next sentence prediction.
- ◉ It consists of the sum of a binary loss used for next sentence prediction and a cross entropy loss over V tokens used for the masked language modeling.
- ○ It consists of a cross entropy loss over V tokens used for the masked language modeling.
- ○ It consists of a triplet loss similar to the one you have seen used for siamese networks.

Correct
Correct.

5. Which of the following inputs could be used for the BERT model? 1 / 1 point

- ○ Question/Answer
- ○ Article/Summary
- ○ Hypothesis/Premise
- ◉ All of the above

Correct
Correct.

6. How does the prefix language model attention work in the T5 model? 1 / 1 point

- (●) It uses bidirectional attention for the inputs (i.e. X's) and causal attention mapping the outputs (Y's) at time t, to all the previous X's and outputs before timestep t.
- ( ) It uses an encoder decoder attention.
- ( ) It only uses causal attention through out.
- ( ) It uses bidirectional attention for the X's and the Y's.

Correct
Correct.

7. When training these latest NLP models, you end up training a model that can do many tasks. For example, you usually have data for sentiments, QA, chatbot, summarization, etc. The question now is how do you combine the datasets using temperature scaled mixing? 1 / 1 point

- ( ) You will sample in proportion to the size of each task's dataset.
- (●) You will adjust the "temperature" of the mixing rates. This temperature parameter allows you to weight certain examples more than others. When T = 1, this approach is equivalent to examples-proportional mixing and as T increases the proportions become closer to equal mixing.
- ( ) Each example in each batch is sampled uniformly at random from one of the datasets you train on.
- ( ) You will just use the data for the specific task you are training on.

Correct
Correct.

8. When doing fine-tuning, how do adapter layers work? 1 / 1 point

- (●) It allows you to add a new layer and then you only fine-tune the new layer you added.
- ( ) You freeze only the last layer, and then you gradually unfreeze each layer as you modify and fine-tune each layer starting from the end.
- ( ) You freeze half the layers, and then you gradually unfreeze each layer as you modify and fine-tune one at a time.
- ( ) You just take the pre-trained weights and start fine tuning on all of them in one go.

Correct
Correct.

7. When training these latest NLP models, you end up training a model that can do many tasks. For example, you usually have data for sentiments, QA, chatbot, summarization, etc. The question now is how do you combine the datasets using temperature scaled mixing? 1 / 1 point

- ○ You will sample in proportion to the size of each task's dataset.
- ◉ You will adjust the "temperature" of the mixing rates. This temperature parameter allows you to weight certain examples more than others. When T = 1, this approach is equivalent to examples-proportional mixing and as T increases the proportions become closer to equal mixing.
- ○ Each example in each batch is sampled uniformly at random from one of the datasets you train on.
- ○ You will just use the data for the specific task you are training on.

Correct
Correct.

8. When doing fine-tuning, how do adapter layers work? 1 / 1 point

- ◉ It allows you to add a new layer and then you only fine-tune the new layer you added.
- ○ You freeze only the last layer, and then you gradually unfreeze each layer as you modify and fine-tune each layer starting from the end.
- ○ You freeze half the layers, and then you gradually unfreeze each layer as you modify and fine-tune one at a time.
- ○ You just take the pre-trained weights and start fine tuning on all of them in one go.

Correct
Correct.

9. Which of the following is not evaluated using the GLUE benchmark? 1 / 1 point

- ○ Similarity
- ○ Paraphrase
- ○ Question duplicates
- ◉ Machine Translation

Correct
Correct.