1

Which of the following is NOT a task typically associated with natural language processing?

- translation
- summarization
- finding entities
- **image segmentation**

2

A word embedding converts a word into what?

- an integer
- a floating point number
- a string
- **a vector**

3

Which of the following is NOT true about a self attention layer?

- The parameter count does not increase with the number of words in the input text
- **The parameter count does not increase if we use longer word embeddings**
- It can retain position information using positional encoding
- It can generate output of the same size as the input

4

Let $a[x_n, x_m]$ denote the contribution of the value v_m to the output $sa[x_n]$ of self-attention at position n. What of the following statements about $a[x_n, x_m]$ is FALSE?

- $a[x_n, x_m] \geq 0$
- $a[x_n, x_m] \leq 1$
- $\sum_{m=1}^{N} a[x_n, x_m] = 1$
- **$\sum_{n=1}^{N} a[x_n, x_m] = 1$**

5

Consider a multi-head self-attention layer with H heads and input size D x N. What size output does the first head compute?

- D x N
- **D/H x N**
- D x N/H
- D/H x N/H

6

Which of the following is NOT a component of a transformer layer?

- **convolution**
- self attention
- layer norm
- residual connections

7

Which of the following is a tokenization algorithm?

- positional encoding
- **byte pair encoding**
- one hot encoding
- multi hot encoding

8

What does the "T" in "GPT" stand for?

- **Transformer**
- Tensor
- Tensorflow
- Tape

9

What modification of regular self-attention is used in training an autoregressive language model like GPT3?

- (●) masked self-attention
- ( ) low rank self-attention
- ( ) sparse self-attention
- ( ) fast self-attention

10

What is a good working definition of a "large" language model (LLM)?

- (●) greater than 100B parameters
- ( ) greater than 100K parameters
- ( ) greater than 100M parameters
- ( ) greater than 100T parameters