Started on	Tuesday, 4 February 2025, 12:43 AM
State	Finished
Completed on	Tuesday, 4 February 2025, 12:47 AM
Time taken	4 mins 15 secs
Grade	**5.00** out of 5.00 (**100**%)

Question **1**

Correct

Mark 1.00 out of 1.00

Which of the following methods can be employed to leverage a pre-trained model for performing a specific task?

1. Prompting
2. Updating the pre-trained weights via gradient descent using a specific dataset
3. Few-shot learning

○ only 1

○ only 2

○ both 1 & 3

● 1, 2, and 3 ✓

Your answer is correct.

The correct answer is:
1, 2, and 3

Question **2**

Correct

Mark 1.00 out of 1.00

LLMs might not be trained in a particular language data, still we can get our work done with some human supervision. Which of the following makes the model deal with unknown words during inference?

○ Encoder Architecture

○ Word embedding

○ Word embedding

○ Both Encoder-Decoder architecture

● Byte Pair Encoding ✓

Your answer is correct.

LLM uses Byte Pair Encoding (BPE) which can build a vocabulary from the text provided by the user, that helps in performing the tasks in other languages as well.

The correct answer is:
Byte Pair Encoding

Question **3**

Correct

Mark 1.00 out of 1.00

Match the following:

Column I	Column II
a. Prompt Engineering	i. Technique to reduce the size of deep neural networks by changing the precision of the weights and biases data structure
b. Low Rank Adaptation	ii. The process of structuring text that can be interpreted and understood by a generative AI model
c. Quantization	iii. Natural language text describing the task that an AI should perform
d. Prompt	iv. A technique that accelerates the fine-tuning of large models while consuming less memory

○ a-i, b-iv, c-ii, d-iii

○ a-i, b-iv, c-ii, d-iii

● a-ii, b-iv, c-i, d-iii ✓

○ a-ii, b-iii, c-i, d-iv

○ a-i, b-iii, c-ii, d-iv

Your answer is correct.

- Prompt: Natural language text describing the task that an AI should perform
- Prompt Engineering: The process of structuring text that can be interpreted and understood by a generative AI model
- Quantization: Technique to reduce the size of deep neural networks by changing the precision of the weights and biases data structure
- LoRA: A technique that accelerates the fine-tuning of large models while consuming less memory

The correct answer is:
a-ii, b-iv, c-i, d-iii

Question **4**

Correct

Mark 1.00 out of 1.00

Select the **False** statements w.r.t Quantization:

1. Model quantization is primarily used in reducing the model size and inference time.

2. The main disadvantage of quantizing a model is the potential loss in accuracy.

3. It can only be applied during the training phase of a model.

4. Quantized models are suitable for deployment on edge devices.

5. Post-training quantization typically requires retraining the model.

○ Only 2 and 3

● Only 3 and 5 ✓

○ Only 1 and 4

○ Only 3, 4, and 5

Your answer is correct.

False statements are:

● It can only be applied during the training phase of a model.

Reason: Quantization can be applied both during the training phase and after the model has been trained.

- Quantization-Aware Training (QAT) incorporates quantization into the training process. The model learns to be robust to the effects of quantization, allowing it to minimize accuracy loss when weights and activations are quantized.

- Post-Training Quantization (PTQ) is applied to a pre-trained model without retraining. The quantization process occurs after the model has already been trained.

● Post-training quantization typically requires retraining the model.

Reason: Post-training quantization (PTQ) does not require retraining the model. Instead, it is a technique that allows you to quantize a pre-trained model without going through the training process again.

The correct answer is:
Only 3 and 5

Question **5**

Correct

Mark 1.00 out of 1.00

Which of the following contains both Encoder and Decoder in the whole transformer architecture?

○ BERT

● T5 ✓

○ GPT

○ ChatGPT

Your answer is correct.

The correct answer is:
T5