1. Interacting with Large Language Models (LLMs) differs from traditional machine learning models. Working with LLMs involves natural language input, known as a _____, resulting in output from the Large Language Model, known as the _____ . 1 / 1 point

Choose the answer that correctly fill in the blanks.

- ○ tunable request, completion
- ◉ prompt, completion
- ○ prompt, fine-tuned LLM
- ○ prediction request, prediction response

Correct
The input for working with LLMs is referred to as the prompt and the output from the LLM is referred to as the completion.

2. Large Language Models (LLMs) are capable of performing multiple tasks supporting a variety of use cases. Which of the following tasks supports the use case of converting code comments into executable code? 1 / 1 point

- ○ Information Retrieval
- ○ Invoke actions from text
- ◉ Translation
- ○ Text summarization

Correct
Translation focuses on converting languages, including coding languages so in this case the task focuses on translating code comments into executable code.

3. What is the *self-attention* that powers the transformer architecture? 1 / 1 point

- (●) A mechanism that allows a model to focus on different parts of the input sequence during computation.
- ( ) A technique used to improve the generalization capabilities of a model by training it on diverse datasets.
- ( ) The ability of the transformer to analyze its own performance and make adjustments accordingly.
- ( ) A measure of how well a model can understand and generate human-like language.

Correct
Self-attention is a key component in models like Transformers, where it enables the model to attend to different words in the input sequence to capture their relationships and dependencies.

4. Which of the following stages are part of the generative AI model lifecycle mentioned in the course? (Select all that apply) 1 / 1 point

- [x] Manipulating the model to align with specific project needs.

Correct
It is likely we will have to manipulate the model in some way to align it with the specific needs of the project.

- [ ] Performing regularization
- [x] Defining the problem and identifying relevant datasets.

Correct
It is crucial to define the problem being solved and identify relevant datasets instrumental to the project.

- [x] Selecting a candidate model and potentially pre-training a custom model.

Correct
Selecting a candidate model and potentially pre-training a custom model are important stages in the generative AI model lifecycle.

- [x] Deploying the model into the infrastructure and integrating it with the application.

Correct
Once we have a model performing to our needs, we can deploy it into the infrastructure and integrate it with the application.

5. "RNNs are better than Transformers for generative AI Tasks." 1 / 1 point

Is this true or false?

- ○ True
- ◉ False

Correct
While RNNs can be used for generative AI tasks, they struggle with compute and memory, making it hard to keep context in longer texts. The transformers architecture is more parallelizable and its dynamic attention mechanism helps to capture long-range dependencies in the input.

6. Which transformer-based model architecture has the objective of guessing a masked token based on the previous sequence of tokens by building bidirectional representations of the input sequence. 1 / 1 point

- ○ Sequence-to-sequence
- ○ Autoregressive
- ◉ Autoencoder

Correct
Autoencoder models are pre-trained using masked language modeling. They use randomly masked tokens in the input sequence and the pretraining objective is to predict the masked tokens to reconstruct the original sentence.

7. Which transformer-based model architecture is well-suited to the task of text translation? 1 / 1 point

- ○ Autoencoder
- ◉ Sequence-to-sequence
- ○ Autoregressive

Correct
Sequence-to-sequence models use both the encoder and decoders in the transformer-based architecture making them best suited for tasks such as translation, text summarization, and question answering. In the Transformers video, Mike explains it in more detail.

8. Do we always need to increase the model size to improve its performance? 1 / 1 point

- ○ True
- ◉ False

Correct
Recent trends show that we can build better LLMs without necessarily increasing model size year by year. Models like LLaMa and BloombergGPT have demonstrated the possibility of reducing model size while keeping great performance.

9. Scaling laws for pre-training large language models consider several aspects to maximize performance of a model within a set of constraints and available scaling choices. Select all alternatives that should be considered for scaling when performing model pre-training? 1 / 1 point

- ☑ Model size: Number of parameters

Correct
The size of the model in terms of number of parameters is a key scaling choice to consider with compute constraints because the number of parameters directly impacts the compute needs required during pre-training.

- [x] Dataset size: Number of tokens

Correct
The size of the pre-training data is an important factor to consider when scaling with compute constraints. This is because the size of the dataset directly affects the computational requirements during pre-training, and having a larger dataset generally leads to improved model performance.

- [x] Compute budget: Compute constraints

Correct
The compute budget plays a crucial role in scaling during pre-training. When faced with a limited compute budget, we may need to impose restrictions on either the model size or the dataset size.

- [ ] Batch size: Number of samples per iteration

10. "You can combine data parallelism with model parallelism to train LLMs." 1 / 1 point

Is this true or false?

- (•) True
- ( ) False

Correct
Combining data parallelism with pipeline parallelism is known as 2D parallelism.We can achieve 3D parallelism by combining data parallelism with both pipeline parallelism and tensor parallelism simultaneously.