Warm test

Q1:

Answer: Answer: D

Generative AI is a type of artificial intelligence (AI) that can create new content, such as text, images, audio, and video. It does this by learning from existing data and then using that knowledge to generate new and unique outputs.

Q2:

Answer: A

A generative AI model could be trained on a dataset of images of cats and then used to generate new images of cats. A discriminative AI model could be trained on a dataset of images of cats and dogs and then used to classify new images as either cats or dogs.

Q3:

Answer: E

A foundation model is a large AI model pre-trained on a vast quantity of data that is "designed to be adapted" (or fine-tuned) to a wide range of downstream tasks, such as sentiment analysis, image captioning, and object recognition.

Q4:

Answer: A, B, D

Q5:

Answer: C

Exit test

Q1:

Answer B.

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

Q2:

Answer: D

Q3
Answer: C
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.

Q4:

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

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

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

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

Q5

Answer:
False: 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.

Q6

Answer: A

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
Q7

Answer: B

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.

Q8
Answer:

A: 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.

B: 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.

D: 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.

Q9

Answer: D

Q10
Answer: A

Q11

Answer: C

BLEU focuses on precision and text translation while Rouge focuses on text summarization.

The evaluation metric that focuses on precision in matching the generated output to the reference text and is commonly used for text translation tasks is called "BLEU" (Bilingual Evaluation Understudy).

BLEU is a widely used metric for evaluating the quality of machine-translated text by comparing it to one or more reference translations. It measures the similarity between the generated translation and the reference translations based on n-gram overlap. BLEU focuses on precision by counting the number of n-grams (sequences of n words) in the generated translation that match n-grams in the reference translations.

While BLEU considers both precision and recall, it predominantly emphasizes precision because it penalizes the generated translation for including n-grams that do not appear in the reference translations. Therefore, higher BLEU scores indicate better precision in matching the generated output to the reference text.

Other evaluation metrics commonly used in text translation tasks include METEOR, TER (Translation Error Rate), and ROUGE (Recall-Oriented Understudy for Gisting Evaluation). However, among these, BLEU specifically emphasizes precision in matching the generated output to the reference text.

Q12

Answer: C

LoRA represents large weight matrices as two smaller, rank decomposition matrices, and trains those instead of the full weights. The product of these smaller matrices is then added to the original weights for inference.

Q13
Answer: A

A soft prompt refers to aa set of trainable tokens that are added to a prompt. Unlike the tokens that represent language, these tokens can take on any value within the embedding space. The token values may not be interpretable by humans, but are located in the embedding space close to words related to the language prompt or task to be completed.

Q14
Answer: False

Prompt Tuning focuses on optimizing the prompts given to the model using trainable tokens that don't correspond directly to human language. The number of tokens you choose to train, however, would be a hyperparameter of your training process.

Q15
Answer: True

By training a smaller number parameters, whether through selecting a subset of model layers to train, adding new, small components to the model architecture, or through the inclusion of soft prompts, the amount of memory needed for training is reduced compared to full fine-tuning.