LLM

Q1. What is tokenization, and why is it important in LLMs?

This step is crucial because LLMs do not understand raw text directly. Instead, they process sequences of numbers that represent these tokens.

Q2. What is LoRA and QLoRA?

LoRA and QLoRA are techniques designed to optimize the finetuning of Large Language Models (LLMs), focusing on reducing memory usage and enhancing efficiency without compromising performance in Natural Language Processing (NLP) tasks.

LoRA (Low-Rank Adaptation)

LoRA is a parameter-efficient fine-tuning method that introduces new trainable parameters to modify a model's behavior without increasing its overall size. It works by adding low-rank matrix adaptations to the model's existing layers, allowing for significant performance improvements while keeping resource consumption in check.

QLoRA (Quantized LoRA)

QLoRA builds on LoRA by incorporating quantization to further optimize memory usage. It uses techniques such as 4-bit Normal Float, Double Quantization, and Paged Optimizers to compress the model's parameters and improve computational efficiency. By reducing the precision of model weights (e.g., from 16-bit to 4-bit) while retaining most of the model's accuracy.

Q3. What is beam search, and how does it differ from greedy decoding?

Beam search is a search algorithm used during text generation to find the most likely sequence of words. Instead of choosing the single highest-probability word at each step (as greedy decoding does), beam search explores multiple possible sequences in parallel, maintaining a set of the top k candidates (beams). It balances between finding high-probability sequences and exploring alternative paths. This leads to more coherent and contextually appropriate outputs, especially in long-form text generation tasks.

Q4. Explain the concept of temperature in LLM text generation.

Temperature is a hyperparameter that controls the randomness of text generation by adjusting the probability distribution over possible next tokens. A low temperature (close to 0) makes the model highly deterministic, favoring the most probable tokens. Conversely, a high temperature (above 1) encourages more diversity by flattening the distribution, allowing less probable tokens to be selected.

Q5. What is masked language modeling, and how does it contribute to model pretraining?

- Masked language modeling (MLM) is a training objective where some tokens in the input are randomly masked, and the model is tasked with predicting them based on context. This forces the model to learn contextual relationships between words, enhancing its ability to understand language semantics. MLM is commonly used in models like BERT, which are pretrained using this objective to develop a deep understanding of language before fine-tuning on specific tasks.

**Q6. What are Sequence-to-Sequence Models?**

**Ans -** Sequence-to-Sequence (Seq2Seq) Models are a type of neural network architecture designed to transform one sequence of data into another sequence. These models are commonly used in tasks where the input and output have variable lengths, such as in machine translation, text summarization, and speech recognition.

**Q7. How do autoregressive models differ from masked models in LLM training?**

Autoregressive models, such as GPT, generate text one token at a time, with each token predicted based on the previously generated tokens. This sequential approach is ideal for tasks like text generation. Masked models, like BERT, predict randomly masked tokens within a sentence, leveraging both left and right context. Autoregressive models excel in generative tasks, while masked models are better suited for understanding and classification tasks.

**Q8. What role do embeddings play in LLMs, and how are they initialized?**

**Ans -** Embeddings are dense, continuous vector representations of tokens, capturing semantic and syntactic information. They map discrete tokens (words or subwords) into a high-dimensional space, making them suitable for input into neural networks. Embeddings are typically initialized randomly or with pretrained vectors like Word2Vec or GloVe. During training, these embeddings are finetuned to capture task-specific nuances, enhancing the model’s performance on various language tasks.

**Q9. What is next sentence prediction and how is useful in language modelling?**

**Ans -** Next Sentence Prediction (NSP) is a key technique used in language modeling, particularly in training large models like BERT (Bidirectional Encoder Representations from Transformers). NSP helps a model understand the relationship between two sentences, which is important for tasks like question answering, dialogue generation, and information retrieval.

**Q10. Explain the difference between top-k sampling and nucleus (top-p) sampling in LLMs.**

**Ans -** Top-k sampling restricts the model’s choices to the top k most probable tokens at each step, introducing controlled randomness. For example, setting k=10 means the model will only consider the 10 most likely tokens. Nucleus sampling, or top-p sampling, takes a more dynamic approach by selecting tokens whose cumulative probability exceeds a threshold p (e.g., 0.9). This allows for flexible candidate sets based on context, promoting both diversity and coherence in generated text.