TOP 50 LLM

Interview Questions

Bhavishya Pandit

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

Ans - Tokenization is the process of splitting text into smaller units called tokens, which can be words, subwords, or even characters. For instance, the word “tokenization” might be broken down into smaller subwords like “token” and “ization.” This step is crucial because LLMs do not understand raw text directly. Instead, they process sequences of numbers that represent these tokens.

Effective tokenization allows models to handle various languages, manage rare words, and reduce the vocabulary size, which improves both efficiency and performance.

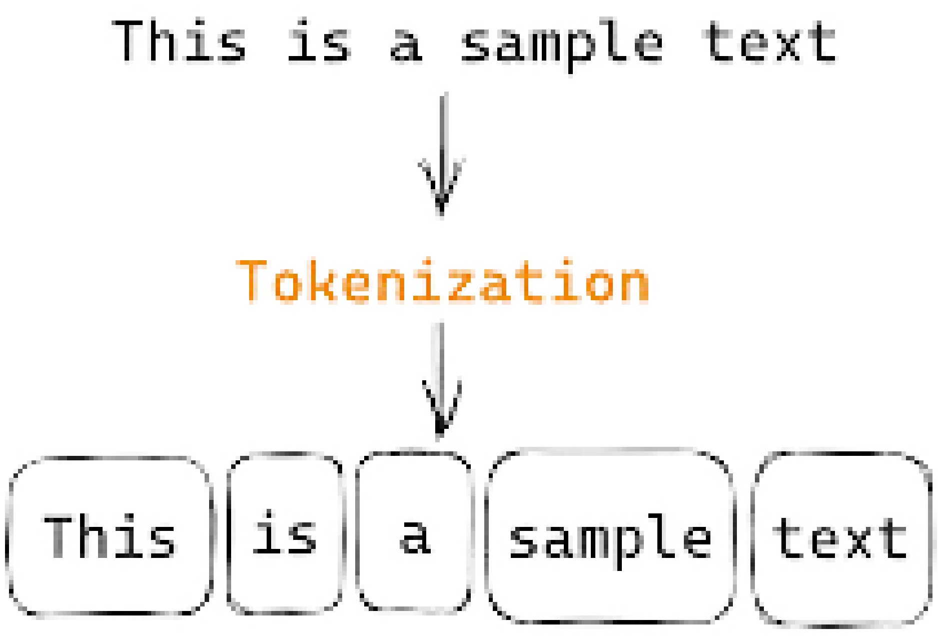
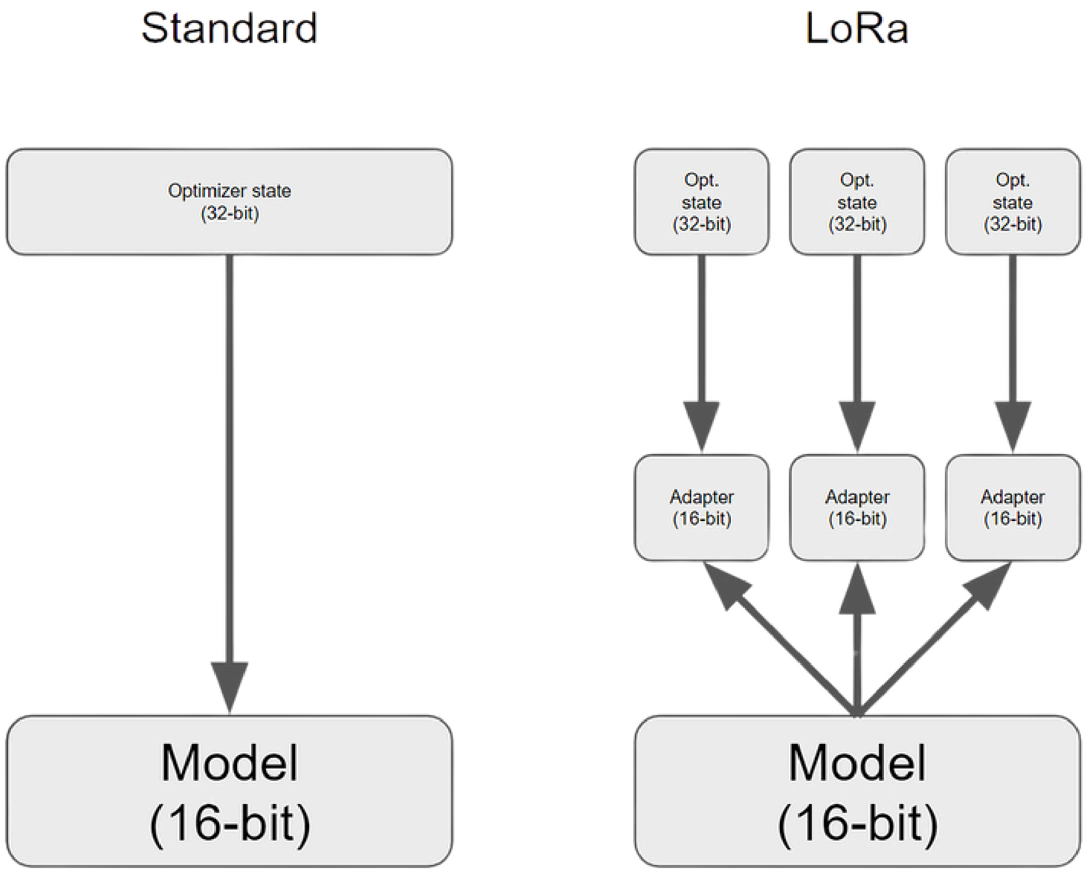


Image Source: Cognitive Class

Q2. What is LoRA and QLoRA?

Ans - LoRA and QLoRA are techniques designed to optimize the fine-tuning 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. By doing so, LoRA maintains the original parameter count, reducing the memory overhead typically associated with training large models. 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.