

Q1. What is NVIDIA Triton and what machine learning or AI problems is it designed to solve?

Answer:

Q2. Explain the core abstractions or APIs in NVIDIA Triton (models, tensors, pipelines, agents).

Answer:

Q3. How does NVIDIA Triton handle training: defining a model, specifying a loss, and running optimization?

Answer:

Q4. What is NVIDIA Triton's approach to data ingestion, preprocessing, and batching?

Answer:

Q5. How do you evaluate model performance in NVIDIA Triton (metrics, validation sets, cross-validation)?

Answer:

Q6. How does NVIDIA Triton support model deployment and serving in a production environment?

Answer:

Q7. What hardware acceleration does NVIDIA Triton support (GPU, TPU) and how do you configure it?

Answer:

Q8. How does NVIDIA Triton handle distributed or multi-GPU training?

Answer:

Q9. What are common debugging techniques for model training issues in NVIDIA Triton?

Answer:

Q10. How do you version and experiment-track models in a NVIDIA Triton-based ML workflow?

Answer:
