

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

Answer:

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

Answer:

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

Answer:

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

Answer:

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

Answer:

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

Answer:

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

Answer:

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

Answer:

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

Answer:

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

Answer:
