

PRACTICE QUESTIONS

NVIDIA

NCA-AIIO

AI Infrastructure and Operations Associate

SHARED BY [u/Clever-Llama-6582](#) on aiexamcerts

EDITION Free Set 2

QUESTIONS 10

SHARED October 15, 2026

EXAM NVIDIA NCA-AIIO, AI Infrastructure and Operations Associate

Shared by u/Clever-Llama-6582 on aiexamcerts

10 practice questions for the NVIDIA NCA-AIIO exam, AI Infrastructure and Operations Associate, each followed by its answer and a short explanation. More free sets and the full practice exam:

www.aiexamcerts.com/nvidia/nca-aiio. Spotted a mistake? Email support@aiexamcerts.com with the exam code NCA-AIIO and the question number.

1. A team reports that its model scores 99 percent accuracy on the training set and 61 percent on held-out data. Which condition does this pattern most likely indicate?

A Vanishing learning rate
B Underfitting
C Data leakage into the test set only
D Overfitting

Answer: D

A large gap between training and validation performance is the classic sign of overfitting, where the model memorizes training examples. Underfitting shows poor performance on both sets, which is not the case here. Leakage would usually inflate the validation score, not depress it.

2. Serving a model at lower numeric precision can change its output quality, so accuracy should be re-measured after quantization.

A True
B False

Answer: A (True)

Quantization introduces rounding error that can shift predictions, which is why teams re-evaluate. The claim that output is always identical is the tempting error, so the statement as written is true.

3. Which statement correctly describes a transformer's self-attention mechanism?

A Each token is replaced by its nearest cluster centroid
B Each token is processed only with the tokens immediately to its left in a fixed window
C Each token computes weighted relevance scores against every token in the sequence
D Each token is converted into a single fixed hash value

Answer: C

Self-attention computes query-key similarities so every token can weigh information from all others in the sequence. A fixed left-only window describes a restricted or causal local pattern rather than full self-attention. Centroid replacement is closer to vector quantization, and hashing discards the context that attention relies on.

4. A technique that stops training when validation performance stops improving is called _____ stopping.

Answer: early

Early stopping halts training once validation loss or accuracy plateaus or worsens, reducing overfitting. Checkpointing saves state but does not decide when to stop, which makes it a tempting wrong answer.

5. Which two practices help control the cost of running GPU inference in production? (Choose two.)
- A Autoscaling replicas based on request queue depth and latency
 - B Keeping every GPU at maximum clock with no idle policy
 - C Batching requests where latency targets allow
 - D Serving a 70B model for simple classification tasks
 - E Pinning every request to a dedicated GPU regardless of load

Answer: A, C

Autoscaling matches capacity to demand and batching raises throughput per GPU when latency budgets permit. Running a large model for trivial tasks, keeping hardware at full power while idle and dedicating a GPU to every request all raise cost.

6. What is the role of the learning rate during gradient descent training?
- A It sets how large a step the optimizer takes when updating weights
 - B It sets how many examples are processed in each batch
 - C It sets how many layers the network contains and how wide each layer is
 - D It sets the proportion of data held out for validation

Answer: A

The learning rate scales each gradient-based update, so it controls how far weights move on each step. Batch size is tempting because it also affects training dynamics, but it is a separate hyperparameter. Layer count is an architecture choice, and the validation split is a data partitioning decision.

7. A company wants to serve a trained model to thousands of requests per second with dynamic batching, multiple model versions and standard HTTP and gRPC endpoints. Which NVIDIA component is designed for this?
- A Triton Inference Server
 - B NeMo training container
 - C Base Command Manager
 - D cuBLAS library

Answer: A

Triton Inference Server supports dynamic batching, concurrent model execution and multiple model repositories behind standard protocols. NeMo is a framework for building and training models, which makes it a tempting but wrong choice for serving. Base Command Manager manages cluster provisioning, and cuBLAS provides linear algebra kernels.

8. NVLink is the standard interconnect used between separate racks in a data center.
- A True
 - B False

Answer: B (False)

NVLink is designed for high-bandwidth links inside a node and across tightly coupled NVSwitch systems, while racks are usually linked with InfiniBand or Ethernet. Because it is not the standard inter-rack fabric, the statement is false.

9. Which network technology is widely used in AI training clusters for low-latency, high-throughput node-to-node traffic and supports RDMA?
- A Gigabit Ethernet over Cat5e copper
 - B Fibre Channel storage area network
 - C InfiniBand
 - D USB 3.0 network adapters

Answer: C

InfiniBand offers very low latency and RDMA support, which is why many AI clusters use it for collective traffic between nodes. Gigabit Ethernet, Fibre Channel SANs and USB adapters do not provide RDMA-class latency and bandwidth for training traffic, so each is a tempting but unsuitable choice.

10. Which product family from NVIDIA is a complete, pre-integrated DGX-based AI system designed to scale to large, multi-rack training clusters?
- A DGX SuperPOD
 - B Jetson Nano developer kit
 - C GeForce gaming card
 - D Tesla-branded P100 workstation

Answer: A

DGX SuperPOD is an NVIDIA reference architecture built from DGX systems and networking for large-scale AI clusters. Jetson is an edge AI module family and is not designed for data center training clusters. GeForce is a consumer graphics line, and the P100 is an older data center GPU rather than a scalable cluster system.