

PRACTICE QUESTIONS

NVIDIA

NCA-AMAD

NVIDIA-Certified Associate: AI Model and Application
Development

SHARED BY [u/Quick-Pixel-6684](#) on aiexamcerts

EDITION	Free Set 1
QUESTIONS	10
SHARED	October 18, 2026
EXAM	NVIDIA NCA-AMAD, NVIDIA-Certified Associate: AI Model and Application Development

Shared by [u/Quick-Pixel-6684](#) on aiexamcerts

10 practice questions for the NVIDIA NCA-AMAD exam, NVIDIA-Certified Associate: AI Model and Application Development, each followed by its answer and a short explanation. More free sets and the full practice exam: aiexamcerts.com/nvidia/nca-amad. Spotted a mistake? Email support@aiexamcerts.com with the exam code NCA-AMAD and the question number.

1. Which practice best supports auditability of a training dataset?

- A Keeping only the latest copy of the file and deleting older versions to save storage space
- B Deleting the ingestion logs after each run to save storage
- C Overwriting labels in place so the dataset always looks clean
- D Recording lineage linking each training example to its source and transformations

Answer: D

Lineage links each example to its origin and processing history, so an auditor can trace how the training set was built. Keeping only the latest copy, which is the tempting simplification, removes the history an audit needs.

2. A mixture-of-experts model activates only a subset of its _____ for each token it processes.

Answer: experts (also accepted: expert networks)

A router sends each token to a small number of experts, so the active parameter count per token is far below the total. This is why such models can be large in capacity while keeping per-token compute modest.

3. Which two techniques reduce the GPU memory needed to serve a large language model? (Choose two.)

- A Quantizing weights to FP8 or INT4, with an accuracy check against the target task
- B Increasing the batch size without limit to share the memory cost
- C Limiting the maximum context length to what the application actually needs
- D Keeping every request's KV cache in memory after the request completes
- E Adding more uncompressed examples to every prompt to improve quality

Answer: A, C

Lower-precision weights take fewer bytes each, and a shorter maximum context caps the KV cache. Unlimited batching and retained caches both raise memory use, which makes them tempting but harmful. Longer prompts add memory cost rather than reducing it.

4. Which long-term memory approach stores past user facts so that a later session can retrieve the relevant ones?
- A An external store that is queried at the start of each session for relevant facts
 - B The model's sampling temperature setting, which controls how random each token choice is
 - C A single prompt that is reset after every request
 - D Tokenization of the input text into subword units

Answer: A

An external memory store persists facts beyond one session and retrieves the relevant ones when needed. Resetting the prompt keeps only short-term context, which is the tempting but insufficient option. Temperature and tokenization do not store anything.

5. Which two statements about LoRA are correct? (Choose two.)
- A It updates every base-model weight in full on each training step
 - B It trains small low-rank adapter matrices while the base weights stay frozen
 - C It removes the need for any training data
 - D It produces small adapter files that can be swapped for different tasks
 - E It can only be applied to encoder-only models

Answer: B, D

LoRA freezes the base model and learns low-rank update matrices, which are small enough to store and swap per task. Full weight updates are what LoRA avoids, and it still needs task data. It applies to many transformer architectures, not only encoder-only ones.

6. Which two statements describe prefill and decode in LLM inference? (Choose two.)
- A Decode processes the whole prompt in one parallel pass
 - B Prefill processes the full input prompt and builds the KV cache
 - C Decode is always faster than prefill for each token
 - D Decode generates output tokens one at a time
 - E Prefill produces no cached state at all

Answer: B, D

Prefill reads the whole prompt in parallel and fills the KV cache, and decode then generates output tokens one step at a time. Assigning the parallel pass to decode is the tempting mix-up, and prefill does produce cached state.

7. A data flywheel uses production feedback to curate new training data and to evaluate successive model versions.
- A True
 - B False

Answer: A (True)

A data flywheel closes the loop between deployment and improvement, turning logged interactions and feedback into better data and better evaluations. The loop only helps when the curated data is checked for quality and privacy first.

8. Repeated copies of the same training example are removed by a step called _____.

Answer: deduplication (also accepted: de-duplication, dedupe)

Deduplication removes exact and near-duplicate records so that repeated examples do not overweight the model or inflate evaluation scores. Duplicates that cross the train and test split are a common cause of leakage.

9. Which metric is the time from sending a request until the first output token arrives?

- A Throughput in tokens per second averaged over a full day
- B Time to first token
- C Model file size on disk
- D Number of GPUs in the cluster

Answer: B

Time to first token measures how long a user waits before the response starts, which dominates perceived responsiveness in chat. Daily throughput measures capacity rather than the wait, which is the tempting but different metric.

10. Which NVIDIA application is used for robotics simulation and synthetic data generation for training physical AI?

- A cuDF
- B Triton Inference Server
- C Isaac Sim
- D Dask

Answer: C

Isaac Sim is NVIDIA's robotics simulation application, built on Omniverse, and it is used to generate labeled scenes and test robot behavior. Triton serves models, cuDF processes tabular data, and Dask schedules distributed work, so none of them simulates robots.