

PRACTICE QUESTIONS

NVIDIA

NCA-AMAD

NVIDIA-Certified Associate: AI Model and Application
Development

SHARED BY [u/rapidjaguar2001](#) on aiexamcerts

EDITION	Free Set 2
QUESTIONS	10
SHARED	October 18, 2026
EXAM	NVIDIA NCA-AMAD, NVIDIA-Certified Associate: AI Model and Application Development

Shared by [u/rapidjaguar2001](#) 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. Strong accuracy on a public benchmark is usually enough to decide whether a model works for a specific business workflow.

- A True
- B False

Answer: B (False)

Public benchmarks rarely match a company's data, users or failure costs, so application-specific evaluations are needed. The benchmark score is a useful screening signal rather than the final decision.

2. Which model architecture is typically used to produce text embeddings for semantic search?

- A A diffusion denoiser used only for images
- B An autoregressive chat decoder that generates long replies
- C A convolutional image classifier with pooling layers
- D An encoder-style transformer that outputs dense vectors

Answer: D

Encoder-style transformers read a passage as a whole and output dense vectors that capture meaning, which suits similarity search. A chat decoder is tempting, but it is built to generate text rather than to produce comparable vectors.

3. Which fine-tuning method teaches an instruction-following style from labeled prompt and response pairs?

- A Supervised fine-tuning on prompt and response pairs
- B Prompt engineering with no training data at all, relying only on carefully written instructions
- C Rebuilding a vector index over the document store
- D Pruning weights until most of them are set to zero

Answer: A

Supervised fine-tuning updates the model on labeled examples of the desired prompt and response behavior. Prompt engineering changes behavior without training, which is the tempting but different approach, and pruning and index rebuilding do not teach a style.

4. Roughly how much GPU memory do the weights of a 7-billion-parameter model need in FP16?
- A About 3.5 GB for the full set of weights
 - B About 14 GB for the full set of weights
 - C About 0.7 GB for the full set of weights
 - D About 140 GB for the full set of weights

Answer: B

Seven billion parameters at 2 bytes each in FP16 is about 14 GB before activations and the KV cache are added. The 3.5 GB figure corresponds to 4-bit weights, which is the tempting mix-up with quantization.

5. Which practice most reduces the sim-to-real gap for a robot policy trained in simulation?
- A Training only on one fixed scene with no variation in lighting, textures, or other conditions
 - B Disabling sensor noise models so the simulated data is cleaner
 - C Domain randomization over lighting and dynamics, then validation on real sensors
 - D Lowering the render resolution of the training images only

Answer: C

Domain randomization exposes the policy to varied conditions so it does not overfit to one simulated appearance, and validation on real sensors confirms the transfer. Clean simulated sensors widen the gap, which is the tempting shortcut that backfires.

6. A technique that trains a smaller student model to imitate a larger teacher model's outputs is called model _____.

Answer: distillation

Distillation transfers behavior from a large teacher to a smaller student, which lowers serving cost and latency. The student learns from the teacher's outputs, not from the original training data alone.

7. Which pattern has one agent produce a plan of steps while separate executor agents or tools carry them out?
- A A single-prompt answer with no intermediate steps
 - B Random restarts of the same request until one succeeds
 - C Greedy decoding with no alternative paths considered
 - D Planner-executor

Answer: D

The planner-executor pattern separates deciding what to do from doing it, which helps with long multi-step tasks. Single-prompt answers and random restarts skip the planning structure, and greedy decoding is a sampling choice rather than an agent architecture.

8. Which technique speeds up decoding by letting a small draft model propose tokens that the larger model then verifies?
- A Speculative decoding
 - B Gradient checkpointing
 - C Label smoothing
 - D Dropout scheduling

Answer: A

Speculative decoding lets the target model verify several draft tokens in one pass, which saves time when the drafts are often accepted. Gradient checkpointing and dropout scheduling are training techniques, and label smoothing changes the loss, so none of them speeds up decoding.

9. Which model family is typically used for generating images from text prompts?
- A Linear regression fitted to pixel intensities
 - B Diffusion models
 - C K-means clustering of image colors
 - D Decision stumps applied to pixel rows

Answer: B

Diffusion models generate images by iteratively denoising, and text conditioning steers the result toward the prompt. The other options are classical methods that are not used as the core of modern text-to-image generation.

10. Which change most directly reduces the KV cache memory needed per request at a fixed batch size?
- A Increasing the maximum context length so long answers fit
 - B Adding more output heads to the attention layers
 - C Quantizing the KV cache or limiting the context length
 - D Increasing the number of concurrent users without any limit

Answer: C

The KV cache grows with context length and numeric precision, so quantizing it or capping the context shrinks the per-request footprint. Raising the context limit or the user count increases memory, which makes those options tempting but wrong.