

PRACTICE QUESTIONS

NVIDIA

# NCA - GENM

NVIDIA-Certified Associate: Generative AI Multimodal

SHARED BY [u/vivid\\_robin\\_95](#) on aiexamcerts

|           |                                                                       |
|-----------|-----------------------------------------------------------------------|
| EDITION   | Free Set 1                                                            |
| QUESTIONS | 10                                                                    |
| SHARED    | October 15, 2026                                                      |
| EXAM      | NVIDIA NCA-GENM, NVIDIA-Certified Associate: Generative AI Multimodal |

---

Shared by [u/vivid\\_robin\\_95](#) on aiexamcerts

10 practice questions for the NVIDIA NCA-GENM exam, NVIDIA-Certified Associate: Generative AI Multimodal, each followed by its answer and a short explanation. More free sets and the full practice exam: [www.aiexamcerts.com/nvidia/nca-genm](http://www.aiexamcerts.com/nvidia/nca-genm). Spotted a mistake? Email [support@aiexamcerts.com](mailto:support@aiexamcerts.com) with the exam code NCA-GENM and the question number.

1. In a Vision Transformer (ViT), how is an input image first converted into a sequence of tokens?

- A Each pixel is treated as a separate word in a fixed vocabulary of 256 items, so the whole image becomes one long sentence
- B The image is split into fixed-size patches that are flattened and linearly projected
- C The image is converted to text by an OCR engine before encoding
- D The image is compressed to a single scalar brightness value

**Answer: B**

ViT divides the image into non-overlapping patches, flattens each one, and projects it into an embedding that acts as a token. Treating every pixel as a word would produce an impractically long sequence, which makes it a tempting but wrong design. OCR is only needed when text must be read from pixels, and a single brightness value discards almost all information.

2. The objective that pulls matching image and text embeddings together while pushing mismatched pairs apart is called \_\_\_\_\_ learning.

**Answer: contrastive**

Contrastive learning uses positive and negative pairs to shape the embedding space, as CLIP does with image-text pairs. Masked modeling is a different pretraining strategy that predicts hidden content, which makes it a plausible but wrong answer.

3. Which architecture is commonly used as the denoising backbone in latent diffusion models such as Stable Diffusion?

- A A U-Net, often with attention layers
- B A single-layer perceptron with no skip connections
- C A recurrent network with no spatial convolutions
- D A decision tree ensemble

**Answer: A**

U-Net-style networks with skip connections and attention are standard backbones for predicting noise in latent diffusion. A single-layer perceptron cannot model spatial structure at this scale, and recurrent networks are not the typical choice for image denoising. Decision trees are not used as the denoiser.

4. Diffusion models generate images by learning to predict and remove noise over a sequence of steps.

- A True
- B False

**Answer: A** (True)

The reverse process of a diffusion model is trained to predict the noise present at each step, and sampling repeatedly removes it. This is the core mechanism, so the statement is true.

5. Which two factors most influence the quality of a text-to-speech voice? (Choose two.)

- A The amount and quality of recorded training speech
- B The tokenizer vocabulary size of a downstream chatbot
- C The vocoder and acoustic model architecture
- D The frame rate of a demo video clip
- E The batch size of the production web server

**Answer: A, C**

Training speech quality and the architecture of the acoustic model and vocoder shape naturalness. A downstream chatbot vocabulary, a demo video frame rate and a web server batch size have no bearing on synthesized voice quality.

6. Which failure is called hallucination in a vision-language model?

- A The model's GPU temperature exceeds its rated limit
- B The model describes objects or details that are not present in the image
- C The tokenizer produces more tokens than the context window allows
- D The image file is saved with an incorrect color profile for the display device used by reviewers

**Answer: B**

Hallucination in VLMs means generated descriptions include objects, attributes or relations absent from the image. GPU temperature and file color profiles are operational issues, and token overflow is a context-length problem rather than a grounding failure.

7. In a LLaVA-style architecture, which component provides the language model with visual information?
- A An audio codec that encodes the image as a waveform
  - B A tokenizer that reads the raw image bytes as ASCII text so that each byte becomes a word
  - C A pretrained vision encoder whose outputs are projected into the LLM's input space
  - D A reward model that scores only the final answer

**Answer: C**

The architecture pairs a vision encoder with a projector so that image features become tokens the LLM can attend to. Treating image bytes as ASCII discards visual structure, and audio codecs and reward models do not provide visual input to the language model.

8. The standard evaluation metric for speech recognition that counts substitutions, deletions and insertions is word \_\_\_\_\_ rate.

**Answer: error**

WER is the ratio of edit operations to reference word count, with lower values meaning better accuracy. Character error rate is a related metric, but the standard term for speech recognition at word level is WER.

9. Why is FP8 or INT8 quantization often applied to a diffusion or language model before deployment?
- A It removes the need for a GPU at inference
  - B It guarantees identical outputs to the FP32 model
  - C It reduces memory footprint and can increase throughput, with accuracy checked afterward
  - D It eliminates the need for a text encoder because the quantized model reads prompts directly

**Answer: C**

Lower precision cuts memory and often raises throughput, but quality must be validated because rounding can change outputs. The claim of identical outputs is the tempting error, and quantization still requires a GPU or other accelerator. Text encoders are still needed for conditioning.

10. Which NVIDIA framework is used for training, customizing and fine-tuning generative AI models at scale?
- A NVIDIA Jetson Linux
  - B NVIDIA Metropolis
  - C NVIDIA Omniverse Kit
  - D NVIDIA NeMo

**Answer: D**

NeMo provides tooling for training and customizing large generative models, including multimodal and speech models. Metropolis targets vision AI applications, Omniverse Kit builds simulation applications, and Jetson Linux is an operating system for edge devices.