

PRACTICE QUESTIONS

NVIDIA

NCA-GENL

NVIDIA-Certified Associate: Generative AI LLMs

SHARED BY [u/Quiet-Jaguar-2305](#) on aiexamcerts

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

Shared by [u/Quiet-Jaguar-2305](#) on aiexamcerts

10 practice questions for the NVIDIA NCA-GENL exam, NVIDIA-Certified Associate: Generative AI LLMs, each followed by its answer and a short explanation. More free sets and the full practice exam:

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

1. What does the byte pair encoding (BPE) algorithm do when building a tokenizer vocabulary?

- A Compresses the trained model weights to 8-bit integers for serving
- B Repeatedly merges the most frequent adjacent pairs into new subword tokens
- C Ranks training documents by relevance to a query string
- D Splits audio waveforms into fixed-length frames

Answer: B

BPE starts from characters or bytes and iteratively merges the most frequent adjacent pair into a new token until the vocabulary reaches its target size. Compressing weights to 8-bit integers is tempting because it also shrinks the model, but that is quantization and it acts on weights, not on the vocabulary.

2. Because plain self-attention is _____ to the order of its inputs, transformers add positional information to the token embeddings.

Answer: permutation invariant (also accepted: permutation-invariant, order invariant, order-invariant)

Self-attention treats its input as a set, so reordering the tokens would produce the same set of outputs without positional information. Positional encodings, learned or sinusoidal, are added to the embeddings to restore word order.

3. Which NVIDIA product packages optimized LLM inference as containers that expose standard APIs?

- A NVIDIA NIM
- B NVIDIA Riva
- C NVIDIA Clara
- D RAPIDS cuML

Answer: A

NVIDIA NIM provides prebuilt, optimized inference microservices in containers with standard APIs, including an OpenAI-compatible interface for LLMs. Riva is the speech AI SDK, which is a common confusion with NIM because both are deployed as services.

4. NVIDIA Triton Inference Server can only serve TensorRT engines and cannot host PyTorch or ONNX Runtime models.

- A True
- B False

Answer: B (False)

Triton supports multiple backends, including TensorRT, PyTorch, ONNX Runtime and Python backends, all served from a model repository. The claim that it is limited to TensorRT is the wrong one.

5. Which two are standard hyperparameter search strategies? (Choose two.)

- A Grid search
- B Training on the test set until accuracy is high
- C Random search
- D Removing the validation split to save data
- E Editing the trained weights by hand

Answer: A, C

Grid search tries every combination on a fixed lattice, and random search samples the space, often finding good settings faster when only a few hyperparameters matter. Tuning against the test set is a form of leakage, which is why it is listed as a wrong approach.

6. When A/B testing a prompt change, which metric should be chosen as the primary decision metric?

- A Whichever metric looks best after the results are in and have been analyzed in detail
- B The number of tokens generated per response on average
- C A metric tied to the business goal, such as task success rate, fixed in advance
- D The GPU temperature recorded during the run on each node

Answer: C

Choosing the primary metric before the test avoids cherry-picking the result that flatters the change. Token count can move without any quality gain, which is why it is a weak primary metric.

7. Which pandas method removes duplicate rows from a DataFrame?

- A dropna
- B drop_duplicates
- C groupby
- D melt

Answer: B

drop_duplicates returns a DataFrame without repeated rows, optionally judged on a subset of columns. dropna removes rows with missing values, which is the common confusion since both names begin with drop.

8. Scaling a feature to zero mean and unit variance is called z-score _____.

Answer: standardization (also accepted: normalization, scaling)

Standardization subtracts the mean and divides by the standard deviation for each feature. Min-max scaling is a different transform that maps values into a fixed range, so it is not the same as z-score.

9. Which practice helps detect demographic bias in model outputs?

- A Training only on data from the largest demographic group
- B Removing the evaluation dataset to avoid overfitting
- C Evaluating quality and error rates separately across demographic subgroups
- D Raising sampling temperature to 2.0 for every test

Answer: C

Disaggregated evaluation compares performance across groups, which exposes gaps that an overall average hides. Training only on the majority group is a cause of bias, which is why it is a tempting but wrong fix.

10. What is NVIDIA NeMo Guardrails used for?

- A Training custom tokenizers from raw text corpora
- B Rendering 3D scenes for simulation
- C Compressing video files for streaming over low-bandwidth mobile networks
- D Defining rails that control LLM inputs, outputs and dialog flow

Answer: D

NeMo Guardrails lets developers define conversational rails in its configuration language, such as topic restrictions and output checks. Tokenizer training is a different task, which is why it is a plausible but incorrect option.