

PRACTICE QUESTIONS

NVIDIA

NCA-GENL

NVIDIA-Certified Associate: Generative AI LLMs

SHARED BY [u/Swift-Dolphin-9320](#) on aiexamcerts

EDITION	Free Set 2
QUESTIONS	10
SHARED	October 15, 2026
EXAM	NVIDIA NCA-GENL, NVIDIA-Certified Associate: Generative AI LLMs

Shared by [u/Swift-Dolphin-9320](#) 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. In LoRA fine-tuning, which parameters are actually trained?

- A All of the base model weights, with nothing frozen, so the full model is retrained each epoch
- B Only the tokenizer vocabulary entries
- C Small low-rank matrices added alongside frozen base weight matrices
- D Only the temperature used at sampling time

Answer: C

LoRA freezes the pretrained weight matrices and learns a low-rank update, usually two small matrices whose product is added to a frozen projection. This keeps the trainable parameter count small. Full fine-tuning is the distractor here, since it updates every base weight rather than a low-rank add-on.

2. Top-p (nucleus) sampling keeps a fixed number of highest-probability tokens, regardless of how much probability mass they cover.

- A True
- B False

Answer: B (False)

Top-p keeps the smallest set of tokens whose cumulative probability reaches p, so the number of candidates changes with the shape of the distribution. The fixed-count behavior described here belongs to top-k sampling instead.

3. Which tool compiles LLMs into optimized TensorRT engines for NVIDIA GPUs, with in-flight batching and paged KV cache support?

- A NeMo Guardrails
- B TensorRT-LLM
- C cuDF
- D Dask

Answer: B

TensorRT-LLM is NVIDIA's library for building optimized LLM engines, including features such as in-flight batching and paged KV cache. NeMo Guardrails is tempting because it is also an LLM tool, but it controls conversation rails rather than compiling engines.

4. The CUDA _____ is the NVIDIA software layer that lets developers write general-purpose parallel programs that run on GPUs.

Answer: platform (also accepted: toolkit, programming model)

CUDA is both a platform and a toolkit that includes compilers, libraries and a runtime for GPU computing. TensorRT is an inference optimizer built on top of that stack, so it is not the general programming layer.

5. Which two statements about a confusion matrix for a classifier are correct? (Choose two.)

- A It reports the GPU utilization of each layer
- B Rows and columns count actual versus predicted labels
- C It stores the optimizer state from training
- D Diagonal entries are the correct predictions for each class
- E It applies only to regression models

Answer: B, D

A confusion matrix tabulates actual against predicted classes, so the diagonal holds correct predictions and off-diagonal cells show mistakes. It is a classification tool, which is why the regression claim is wrong.

6. Which metric matters most when missing a true positive is very costly?

- A Precision
- B Specificity
- C Overall accuracy on a balanced sample
- D Recall

Answer: D

Recall is the share of actual positives that the model finds, so it is the metric to prioritize when false negatives are expensive. Precision is tempting, but it measures how many flagged items are correct and ignores the positives that were missed.

7. Which chart is best for showing the distribution of a single continuous variable?

- A Pie chart
- B Sankey diagram
- C Histogram
- D Word cloud

Answer: C

A histogram bins the values and shows their frequency and spread, which reveals skew and multiple modes. A pie chart shows parts of a whole for a few categories, which is why it is the tempting but poorer choice.

8. The Pearson correlation coefficient captures non-linear relationships between two numeric variables as well as linear ones.
- A True
 - B False

Answer: B (False)

Pearson measures linear association only, so a strong curved relationship can produce a coefficient near zero. Spearman rank correlation is the usual choice when the relationship is monotonic but not linear.

9. Which NIST framework organizes AI risk management into Govern, Map, Measure and Manage functions?
- A NIST AI Risk Management Framework (AI RMF 1.0)
 - B NIST SP 800-53 security and privacy control catalog for federal information systems
 - C ISO 9001 quality management standard
 - D PCI DSS payment card standard

Answer: A

The AI RMF 1.0 defines the four core functions Govern, Map, Measure and Manage for trustworthy AI risk. SP 800-53 is a security control catalog for information systems, which is a plausible NIST document but not the AI framework.

10. What does data provenance mean for a training dataset?
- A The speed at which data moves through the pipeline and storage tiers
 - B A record of where the data came from and how it was processed
 - C The key rotation schedule for the storage bucket
 - D The parameter count of the model trained on it

Answer: B

Provenance tracks sources, licenses, transformations and consent, which supports audits and licensing decisions. Pipeline throughput is an engineering metric, which is the tempting but unrelated meaning.