

PRACTICE QUESTIONS

NVIDIA

NCP-GENL

NVIDIA-Certified Professional: Generative AI LLMs

SHARED BY [u/Silver-Llama-7543](#) on aiexamcerts

EDITION Free Set 1

QUESTIONS 10

SHARED October 15, 2026

EXAM NVIDIA NCP-GENL, NVIDIA-Certified Professional: Generative AI LLMs

Shared by u/Silver-Llama-7543 on aiexamcerts

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

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

-
1. In a decoder-only transformer, what does the causal mask prevent?
- A Each token from attending to tokens that come after it in the sequence
 - B Each token from attending to any earlier position in the sequence, including its own
 - C The feed-forward layer from updating its weights
 - D The embedding table from being shared with the output projection

Answer: A

The causal mask sets attention scores for future positions to negative infinity before softmax, so generation depends only on the past. Attending to earlier tokens is allowed and required, which is why that option is a tempting misreading of the mask.

-
2. The Bradley-Terry model is commonly used to train a reward model from pairwise _____ data.

Answer: preference (also accepted: preference ranking, comparison, pairwise comparison)

Bradley-Terry turns pairwise comparisons into a probability that one response is preferred, and training maximizes that likelihood. Absolute scores are a different labeling scheme that this model does not use.

-
3. Which parallelism strategy splits individual weight matrices across GPUs within a layer?
- A Pipeline parallelism
 - B Tensor parallelism
 - C Data parallelism
 - D Sequence batching

Answer: B

Tensor parallelism shards matrix multiplications across devices and needs communication within each layer. Pipeline parallelism splits the model by layers rather than by matrices, which is the common confusion.

-
4. Gradient accumulation lets a team simulate a larger effective batch size by summing gradients over several micro-batches before an optimizer step.
- A True
 - B False

Answer: B (False)

Gradient accumulation does defer the optimizer step until several micro-batches have contributed gradients, which simulates a larger batch. The statement denies that effect, so it is false.

5. Which two statements about RLHF with PPO are accurate? (Choose two.)
- A The policy is never allowed to see the reward model's scores
 - B A KL penalty toward a reference policy limits how far the policy drifts from the SFT model
 - C PPO requires the reward model to be a classifier with exactly two outputs
 - D The reward model is trained separately from the policy being optimized
 - E RLHF removes the need for any human preference data

Answer: B, D

The KL term keeps the policy near the supervised model, and the reward model is usually trained first from preference data. Reward models can output scalar scores rather than two classes, and RLHF still depends on preference data, which makes those options wrong.

6. What does rotary position embedding (RoPE) do to the query and key vectors?
- A Adds a learned absolute index vector to every token ID before the first layer
 - B Rotates them by position-dependent angles so attention scores depend on relative offset
 - C Replaces attention with a convolution over the full sequence
 - D Stores the position of each token in an external database

Answer: B

RoPE rotates query and key pairs in each head by angles tied to their positions, so the dot product reflects relative distance. Adding learned absolute vectors is the older scheme that RoPE is often confused with.

7. What is the main role of a cross-encoder reranker in RAG?
- A Building the initial vector index for millions of documents with a cross-encoder pass
 - B Translating the query into SQL
 - C Scoring each query-passage pair jointly to reorder a shortlist more precisely
 - D Training the generator's weights

Answer: C

A cross-encoder reads the query and passage together, which gives sharper relevance scores for a small shortlist. Building the index is the job of the bi-encoder stage, which is the tempting confusion.

8. The memory that stores keys and values for previously processed tokens during decoding is called the _____ cache.

Answer: KV (also accepted: key-value, key value)

The KV cache lets each decode step reuse prior attention keys and values instead of recomputing them. Weight storage is separate, which is why it is not the answer here.

9. What is the main risk of evaluating a model on a benchmark that appears in its pretraining data?
- A Contamination can inflate scores, so the result overstates true capability
 - B The benchmark questions are automatically translated into other languages
 - C Perplexity becomes undefined
 - D Latency increases by a fixed amount

Answer: A

If the model has memorized benchmark items, its score reflects recall rather than generalization. Translation, perplexity and latency are not affected in the way the distractors describe.

10. Which metric matters most for interactive chat latency?
- A Total parameter count of the model and its tokenizer vocabulary size
 - B Time to first token together with inter-token latency
 - C Peak FLOPS of the cluster's interconnect switches
 - D Number of training epochs

Answer: B

Time to first token governs how quickly a user sees a response start, and inter-token latency governs how smoothly it streams. Parameter count and training epochs describe the model, not the experience of one request.