

PRACTICE QUESTIONS

NVIDIA

NCP-GENL

NVIDIA-Certified Professional: Generative AI LLMs

SHARED BY [u/solid_seal_65](#) on aiexamcerts

EDITION Free Set 2

QUESTIONS 10

SHARED October 15, 2026

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

Shared by [u/solid_seal_65](#) 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 QLoRA, what does double quantization refer to?

- A Running two GPUs at half precision in parallel
- B Training the adapters twice with different random seeds and averaging the results
- C Quantizing the quantization constants themselves to save additional memory
- D Storing both the base model and the merged model on disk

Answer: C

QLoRA quantizes the per-block scaling constants a second time, which recovers a small amount of memory on top of 4-bit weights. The other options describe hardware or workflow choices that are not part of the technique.

2. Hybrid retrieval that fuses BM25 keyword scores with dense vector scores cannot improve recall for queries containing exact identifiers.

- A True
- B False

Answer: B (False)

Keyword scoring matches exact tokens such as product codes, which dense embeddings can blur, so fusing both often improves recall for identifiers. The statement denies this benefit, so it is false.

3. Pipeline parallelism introduces which main inefficiency?

- A Every GPU must store a full copy of the optimizer state and the model's activations
- B Activations can never be recomputed
- C Pipeline bubbles, where some stages sit idle while others process micro-batches
- D Gradients cannot be averaged across replicas

Answer: C

Stages wait for upstream and downstream work, which creates idle bubbles that micro-batching reduces but does not remove. Full optimizer replication is a data parallel characteristic, which is why it is tempting but not the pipeline issue.

4. _____ is the label value commonly used in PyTorch cross-entropy loss to mark positions, such as padding, that should be excluded from the loss.

Answer: -100 (also accepted: ignore_index, -100 (ignore_index))

PyTorch's default ignore_index is -100, so labels set to that value contribute nothing to the loss. Zero is a valid class index, which is why it is a risky choice for masking.

5. Which two statements about ZeRO and FSDP are accurate? (Choose two.)
- A Parameter sharding adds communication to gather weights before each layer's compute
 - B They remove the need for any gradient communication
 - C They are usable only for inference, not for training
 - D Sharding optimizer states reduces per-GPU memory compared with full replication
 - E They make tensor parallelism unnecessary for every model size

Answer: A, D

Sharding states lowers the memory each GPU holds, and parameter sharding must gather weights before use, which adds traffic. Gradient communication still happens, and these methods are primarily training techniques, which makes the other options incorrect.

6. What is the purpose of the LoRA alpha hyperparameter?
- A It sets the number of attention heads in the adapted layers
 - B It sets the tokenizer vocabulary size
 - C It defines the length of the learning rate warmup
 - D It scales the low-rank update, so its effective contribution relates to alpha divided by the rank

Answer: D

The LoRA update is scaled by alpha divided by rank, so alpha controls the strength of the adaptation for a given rank. Head count and vocabulary size are architecture settings fixed by the base model.

7. What does speculative decoding use to speed up generation?
- A A lower temperature that skips sampling entirely and always returns the single most likely token
 - B A cached final answer shared by all users
 - C A small draft model proposes tokens and the target model verifies them in one parallel pass
 - D Removal of the attention layer

Answer: C

The draft model proposes several tokens, and the target model accepts or rejects them in one parallel pass, so accepted output matches the target's distribution. Lowering temperature changes the output and does not speed up the model, which is the tempting but different idea.

8. FP8 inference on recent NVIDIA data center GPUs such as Hopper can reduce memory and improve throughput compared with FP16 when accuracy is validated.

A True
B False

Answer: A (True)

FP8 halves weight and activation storage relative to FP16 and uses fast tensor core paths on supported hardware. Teams still need accuracy checks because the narrower format can change outputs.

9. Which change most directly reduces the memory needed for attention scores at long sequence lengths?

A Adding more transformer layers to the network
B Using a much larger learning rate during pretraining to reduce the total number of training steps needed
C FlashAttention, which computes attention in tiles without writing the full score matrix to HBM
D Storing all activations in FP64 precision

Answer: C

FlashAttention is an IO-aware exact attention kernel that avoids materializing the full n by n score matrix in high-bandwidth memory. Adding layers increases memory, which is why that option is a tempting but wrong fix.

10. Which chunking strategy best preserves meaning for technical documents with headings?

A Cutting every 64 characters regardless of content and sentence boundaries
B Using one chunk for the whole corpus
C Splitting only at the end of each file
D Splitting along headings and sections, with overlap where needed

Answer: D

Structure-aware chunks keep related text together so each retrieved passage is self-contained. Fixed character cuts can split sentences and tables, which is why they are the tempting but weaker option.