

PRACTICE QUESTIONS

CertiProf

GAIPC

Generative AI Professional Certification

SHARED BY [u/sleepy_ferret_35](#) on aiexamcerts

EDITION	Free Set 1
QUESTIONS	10
SHARED	October 15, 2026
EXAM	CertiProf GAIPC, Generative AI Professional Certification

Shared by [u/sleepy_ferret_35](#) on aiexamcerts

10 practice questions for the CertiProf GAIPC exam, Generative AI Professional Certification, each followed by its answer and a short explanation. More free sets and the full practice exam:

www.aiexamcerts.com/certiprof/gaipc. Spotted a mistake? Email support@aiexamcerts.com with the exam code GAIPC and the question number.

1. Which training approach teaches a language model general patterns from large unlabeled text before task adaptation?

- A Supervised fine-tuning on a small, carefully labeled task dataset
- B Self-supervised pretraining with next-token prediction over raw text
- C Reinforcement learning driven by hand-written rules for each domain
- D Clustering of documents according to their metadata and file fields

Answer: B

Self-supervised pretraining creates its own targets from raw text, such as predicting the next token, which allows training at very large scale. Supervised fine-tuning is the tempting choice, but it comes after pretraining and needs labeled data. Rule-driven reinforcement learning and metadata clustering do not describe how base language models learn.

2. The class of models that create new content such as text, images or code is called _____ AI.

Answer: generative

Generative AI refers to models that produce new content from patterns learned in training. Discriminative models, which classify or predict labels, are the tempting contrast but do not create content.

3. In an encoder-decoder transformer used for translation, what does the encoder primarily produce?

- A The final spoken audio for the translated sentence in the target voice
- B A list of translated tokens ready for direct output to the user
- C Contextual representations of the input sequence
- D A probability table for the next word in the output vocabulary

Answer: C

The encoder turns the source sequence into contextual representations that the decoder attends to while generating the translation. Producing finished tokens is the decoder's job, which is the tempting mix-up. Audio and probability tables are outside the translation encoder's role.

4. Self-attention allows each token representation to incorporate information from other tokens in the sequence.
- A True
 - B False

Answer: A (True)

This is true. Self-attention computes weighted combinations of other positions, which is how context from across the sequence reaches each token.

5. Which two practices improve retrieval quality in a RAG system? (Choose two.)
- A Indexing only the document titles and not their content
 - B Chunking documents at sensible boundaries with overlap where needed
 - C Retrieving a single random document for every query
 - D Adding metadata filters such as product version or date
 - E Disabling the reranker to reduce every query's latency

Answer: B, D

Sensible chunk boundaries keep related information together, and metadata filters narrow results to the right version or date. Indexing only titles discards the content that answers questions. Random retrieval and disabling reranking both reduce relevance, even though the latter may save time.

6. What is the main purpose of a system message in a chat-based model?
- A To choose the GPU type used for each inference request in the cluster
 - B To cap the number of tokens the model may return in its reply
 - C To set standing role, tone and constraints for the conversation
 - D To log every token the user typed in for billing and audit purposes

Answer: C

The system message sets persistent instructions such as role, tone and boundaries that apply across the whole conversation. A response token cap is a separate request parameter, which is the tempting confusion. Billing logs and hardware selection are handled by the platform, not the message.

7. Which output-control technique forces a model's reply to match a defined JSON structure?
- A Lowering the maximum context window size of the request for the call
 - B Raising the temperature to make the replies more creative and varied
 - C Increasing the number of few-shot examples in the prompt for consistency
 - D Structured output or JSON schema constrained decoding

Answer: D

Structured output modes and schema-constrained decoding restrict generation so the reply conforms to the requested shape, which makes it parseable. Examples can improve format adherence, but they do not guarantee it, which is the tempting middle ground. Temperature and context size do not enforce structure.

8. The component that converts text into vectors used for similarity search is the _____ model.

Answer: embedding

An embedding model maps text to vectors so that semantically similar passages are close together. A reranker scores candidate passages after retrieval, which is a different stage.

9. What does a LoRA adapter store?
- A The full list of prompts that each user has sent to the system so far
 - B A complete second copy of the base model weights for each customer
 - C A new tokenizer trained on the company's historical customer tickets
 - D Small low-rank matrices that modify selected layers

Answer: D

A LoRA adapter holds small low-rank matrices that are added to selected layers, so the base model stays unchanged and the adapter is small to store. A full copy of the model is the tempting but wasteful design LoRA avoids. Tokenizers and prompt logs are not what an adapter contains.

10. Why do teams build a golden dataset before fine-tuning?
- A A fixed reference for judging whether tuning helped
 - B It replaces the need for any evaluation after the training run ends
 - C It makes the training run complete in less than one minute of time
 - D It removes all personal data from the original sources before training

Answer: A

A golden dataset of approved inputs and outputs gives a stable benchmark, so the team can tell whether tuning improved the model. Skipping evaluation is the tempting shortcut that the golden set exists to prevent. Speed and data removal are separate concerns and are not solved by it.