

PRACTICE QUESTIONS

CertiProf

GAIPC

Generative AI Professional Certification

SHARED BY [u/Plucky-Falcon-4939](#) on aiexamcerts

EDITION Free Set 2

QUESTIONS 10

SHARED October 15, 2026

EXAM CertiProf GAIPC, Generative AI Professional Certification

Shared by [u/Plucky-Falcon-4939](#) 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. What does the term hallucination most precisely describe in generative AI?
 - A A model slowing down under heavy request volume at peak times
 - B A model refusing to answer when the prompt is written in another language
 - C A model producing fluent output that is not supported by facts or the provided sources
 - D A model reproducing its training data word for word in every single answer

Answer: C

Hallucination refers to confident, fluent output that is unsupported by facts or sources, which is why it is a quality and trust problem. Verbatim memorization is a tempting near miss, but it is a different phenomenon and occurs only occasionally. Latency and refusals are operational or policy behaviors rather than fabrication.

2. A generative model always returns the same output when given the same prompt, regardless of settings.
 - A True
 - B False

Answer: B (False)

This is false. Sampling settings such as temperature and the random seed affect output, and sampling by default introduces variation between runs.

3. Why do decoder-only models generate text one token at a time?
 - A Because the hardware cannot run multiple operations at the same moment
 - B Because the tokenizer only allows one new token for each prompt request
 - C Because the training data was split into single sentences for learning
 - D Because each new token depends on all tokens generated before it

Answer: D

Autoregressive generation conditions each new token on everything produced so far, so the next token can only be chosen after the previous one exists. Hardware parallelism is a tempting explanation, but it does not explain the dependency. Tokenizer rules and data splitting do not control generation order.

4. Showing the model a handful of input and output pairs in the prompt is a _____ learning technique.

Answer: few-shot (also accepted: few shot)

Few-shot prompting teaches the pattern through demonstrations without changing weights. Zero-shot prompting gives no examples, which is the tempting but different approach.

5. Which two statements about transformers and large language models are correct? (Choose two.)

- A Transformers cannot process sequences in parallel during training
- B Positional encodings help the model represent token order
- C The context window is the size of the model's weight files on disk
- D Decoder-only models are commonly used for chat-style text generation
- E Temperature changes the number of layers in the model

Answer: B, D

Positional encodings supply order information that attention does not provide on its own, and decoder-only models are the usual choice for chat generation. Transformers train on whole sequences in parallel, which is the reverse of the first statement. The context window is a token limit rather than a file size, and temperature is a sampling setting.

6. A prompt says 'Answer only with a category label from this list.' What control is being applied?

- A Constraint on the output format through instructions
- B Retrieval of external documents at query time from a vector index
- C Fine-tuning of the model weights on new labeled examples from the team
- D Temperature adjustment for more variety in the generated responses

Answer: A

The instruction restricts the answer format to an allowed set, which is an output constraint expressed in the prompt. Retrieval and fine-tuning change what the model knows or uses, which is the tempting mix-up. Temperature changes randomness rather than the allowed form of the answer.

7. In a RAG pipeline, what does the embedding step produce?

- A A billing record that tracks the cost of every retrieved passage
- B A final formatted answer for the end user with citations included
- C Vector representations of text chunks for similarity search
- D A list of guardrail rules written in plain English for the team

Answer: C

The embedding step converts chunks and queries into vectors so that the system can find semantically similar passages. The final answer comes from the generation step after retrieval, which is the tempting confusion. Guardrail rules and billing records are separate components.

8. Vector search ranks passages by exact keyword matches only.

- A True
- B False

Answer: B (False)

This is false. Vector search ranks by similarity between embeddings, so it can match passages that share meaning without sharing words.

9. Which risk is most associated with fine-tuning on a narrow dataset?

- A Permanent deletion of the prompts stored in the provider's audit logs
- B Catastrophic forgetting of general abilities
- C Automatic improvement in accuracy on every unrelated task type
- D Immediate loss of the tokenizer vocabulary across all deployed versions

Answer: B

Training narrowly can erode general capabilities the model had before, which is known as catastrophic forgetting. Improvement on unrelated tasks is the tempting but unlikely outcome, and it is the opposite of the risk. Tokenizer and log effects are not caused by ordinary fine-tuning.

10. Which dataset property most improves a fine-tuned model's results?

- A Records mixed from several unrelated tasks with no consistent format
- B Large volume of duplicated examples copied from one single source
- C High-quality, representative examples that match the target task
- D Examples scraped from the web without any review or labeling at all

Answer: C

Quality and representativeness matter more than sheer volume, since the model learns whatever the examples demonstrate. Large volumes of duplicates add bias without adding information, which is the tempting shortcut. Unreviewed scraped data and mixed formats teach inconsistent behavior.