

PRACTICE QUESTIONS

Oracle

OCI 2026 Generative AI Professional

SHARED BY [u/crammer_oci2026](#) on aiexamcerts

EDITION Free Set 1

QUESTIONS 10

SHARED October 15, 2026

EXAM Oracle OCI 2026 Generative AI Professional

Shared by [u/crammer_oci2026](#) on aiexamcerts

10 practice questions for the Oracle OCI 2026 Generative AI Professional exam, each followed by its answer and a short explanation. More free sets and the full practice exam:

www.aiexamcerts.com/oracle/oci-2026-generative-ai-professional. Spotted a mistake? Email support@aiexamcerts.com with the exam name and the question number.

1. What is a token in OCI Generative AI text models?

- A A signed credential that authenticates each API call to the service endpoint
- B A numeric vector that stores the meaning of an entire document in a vector database
- C A sub-word unit that counts toward input and output limits
- D An IAM policy statement that grants a group permission to call a generative model

Answer: C

A token is the unit a model reads and generates, and the service measures input and output size in tokens. The signed credential option is tempting because every request is signed, but that role belongs to API keys, not tokens. The vector and IAM options describe embeddings and access control, which are separate concepts.

2. In a transformer model, the mechanism that lets every position attend to the other positions in the sequence is called _____ attention.

Answer: self (also accepted: self-attention, self attention)

Self-attention lets each token compute relevance scores against every other token in the same sequence, and it is the core building block of the transformer.

3. A team wants to host a fine-tuned version of a Cohere model with predictable throughput. Which OCI resource should they plan for?

- A A dedicated AI cluster with an endpoint created for that model
- B An on-demand call with a higher temperature setting for the requests
- C A bucket in Object Storage with versioning enabled for the artifacts
- D A vector store table with a larger embedding dimension for the index

Answer: A

Hosting a fine-tuned model requires a dedicated AI cluster with an endpoint on it, which reserves capacity for predictable throughput. On-demand calls use shared pretrained models, so they cannot host your fine-tuned weights. Object Storage and vector tables store data rather than serve a model.

4. OCI Generative AI on-demand inference lets you call pretrained base models without provisioning your own hosting capacity.
- A True
 - B False

Answer: A (True)

On-demand mode sends requests to shared pretrained models and is billed by usage, so no dedicated hosting is provisioned for you. The statement is true. The tempting error is assuming every call needs a dedicated cluster, which applies only to hosting custom or fine-tuned models.

5. Which two settings most affect whether the right passage is retrieved? (Choose two.)
- A The maximum output tokens used when the model writes the final answer
 - B The chunk size and overlap used when splitting documents for indexing
 - C The temperature used by the final answer generation step
 - D The number of documents returned per query by the retriever (top-k)
 - E The frequency penalty applied during generation of the response text

Answer: B, D

Chunk size and overlap decide what each retrievable unit contains, and the retrieval depth decides how many candidates reach the model. Output length and temperature influence the answer after retrieval, which makes them tempting but not retrieval settings. Frequency penalty acts during generation as well.

6. A model keeps repeating the same phrase in every paragraph. Which parameter is most directly used to discourage this?
- A Frequency penalty
 - B Embedding dimension of the vector index used during retrieval for the application
 - C Chunk overlap used when the source documents were split before indexing
 - D Distance metric used when comparing vectors in the search step of the pipeline

Answer: A

Frequency penalty lowers the probability of a token in proportion to how often it has already appeared, which reduces repetition. Embedding dimension and distance metric are retrieval settings, so they are tempting as vector terms but do not change generated text. Chunk overlap controls how documents are split before indexing.

7. Which field in a chat request sets the instructions that apply to the whole conversation?
- A The stop sequence list that ends generation when a chosen string appears
 - B The maximum output tokens value that caps the length of each reply
 - C The presence penalty value that discourages reuse of tokens already produced
 - D The system message with standing role, tone, and scope instructions

Answer: D

The system message holds standing instructions for role, tone, and scope across the conversation. Stop sequences and output limits control where and how long generation ends, which makes them tempting as conversation-level settings. Presence penalty affects repetition, not instructions.

8. A pattern that retrieves relevant passages first and then generates an answer from them is called _____ generation.

Answer: retrieval-augmented (also accepted: retrieval augmented, RAG)

Retrieval-augmented generation pairs a search step with the model so answers can be grounded in supplied passages. The term describes the combined pattern, not a single product.

9. A fine-tuning dataset contains many near-duplicate examples. What is the most likely consequence?
- A The model becomes better at many languages because the data is more diverse
 - B The endpoint's latency drops by half because the model has fewer patterns to learn
 - C The model may overfit to the repeated patterns and generalize poorly to new inputs
 - D The vector index dimension increases to store the extra duplicated examples

Answer: C

Duplicated examples over-weight a few patterns, so the model can memorize them and generalize poorly to new inputs. Better multilingual coverage needs different data, which makes it tempting. Latency and index dimension are not controlled by the training dataset.

10. Which resource is required to host a fine-tuned model endpoint for production calls in OCI?
- A An on-demand compartment policy that grants access but reserves no capacity
 - B A public Object Storage bucket that allows anonymous read access to its contents
 - C An OCI Functions application that calls the public base model on each request
 - D A dedicated AI cluster with the custom model deployed to an endpoint on it

Answer: D

Custom models are served from an endpoint on a dedicated AI cluster, which reserves the capacity they need. A bucket holds training data or artifacts, so it is tempting but does not serve the model. Functions and access policies do not provide hosting capacity for a custom model.