

PRACTICE QUESTIONS

Oracle

OCI 2026 Generative AI Professional

SHARED BY [u/nightowl_oci2026](#) on aiexamcerts

EDITION Free Set 2

QUESTIONS 10

SHARED October 15, 2026

EXAM Oracle OCI 2026 Generative AI Professional

Shared by [u/nightowl_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. Why does the temperature setting matter when you call a text generation model?
- A It sets the maximum number of output tokens the service will bill for each single request
 - B It selects the OCI region and availability domain that host the model endpoint
 - C It fixes the number of dimensions in the embedding vectors that the model returns
 - D It controls sampling randomness, where lower values give more predictable output

Answer: D

Temperature rescales the probability distribution used to sample the next token, so low values favor the most likely tokens and high values add variety. Maximum output tokens is the setting that caps response length and cost, which makes it tempting, but it does not change randomness. Region and embedding dimensions are fixed by other configuration rather than by temperature.

2. Lowering the max output tokens setting makes the model's answers more factually accurate.
- A True
 - B False

Answer: B (False)

The max output tokens setting only limits how long a response can be, so a shorter answer can still be wrong or cut off mid-thought. Accuracy depends on the model, the prompt, and grounding context such as retrieved passages. The statement confuses a length limit with a quality control.

3. Which OCI feature lets an agent answer questions from a knowledge base built from your own documents?
- A Dedicated AI cluster hosting units sized for a single endpoint
 - B OCI Data Labeling annotation jobs for image and text datasets
 - C OCI Monitoring alarms that fire on endpoint latency thresholds
 - D The Generative AI Agents service with a knowledge base connected to your data sources

Answer: D

The Generative AI Agents service pairs a model with a knowledge base built from your data sources, so it can answer from your documents. Dedicated clusters host models but do not manage retrieval over documents, which makes them a tempting but incomplete answer. Data Labeling and Monitoring serve other purposes entirely.

4. Client applications call a hosted model through its inference _____ over HTTPS, using signed requests.

Answer: endpoint (also accepted: API endpoint)

An inference endpoint is the address that receives requests for a specific hosted model, and each call must be signed with your credentials. The blank names the address, not the bucket or the policy.

5. Which two statements about hybrid search are accurate? (Choose two.)

- A It combines keyword matching with vector similarity in a single ranked result list
- B It works only when the documents contain no numbers or product codes at all
- C It removes the need to split long documents into chunks before indexing them
- D It can improve results for queries that contain exact codes or product names
- E It is the same technique as temperature sampling applied to the search results

Answer: A, D

Hybrid search merges lexical matches with semantic vector matches, which helps when exact identifiers matter alongside meaning. Claims that it avoids chunking or fails on numbers are wrong. Temperature sampling is a generation setting, not a search method.

6. Which setting most directly caps how long a model's reply can be?

- A The temperature of the sampling step used for each generated token
- B The top-k cutoff that limits the candidate tokens considered at each step
- C The maximum output tokens limit set on the inference request itself
- D The embedding model used by the search step that finds relevant passages

Answer: C

The maximum output tokens limit stops generation at a set count, so the reply cannot exceed that length. Temperature and top-k change which tokens are picked, which can change phrasing as a side effect, so they are tempting but less direct. The embedding model does not control reply length.

7. A product team needs the model to return only a JSON object for downstream parsing. What is the best first step?

- A State the required JSON schema in the prompt and validate each reply
- B Increase the number of retrieved chunks to fifty for every request
- C Lower the frequency penalty to zero and hope the output format holds
- D Switch from a chat model to an embedding model for the formatting step

Answer: A

Stating the exact schema in the prompt and validating each reply gives the most reliable format control, and failed replies can be retried. Adding retrieved chunks changes the knowledge available, not the output structure, which makes it a tempting but wrong fix. Penalties and model type do not enforce a format.

8. Re-ranking retrieved chunks with a second scoring step can improve precision before the answer is generated.
- A True
 - B False

Answer: A (True)

A reranker scores the candidates more carefully than the first-pass vector search, so the top passages tend to be more relevant. The statement is true. The mistaken idea is that reranking replaces retrieval, but it works on the candidates that retrieval returns.

9. A fine-tuned model performs well on training prompts but poorly on new customer questions. What should the team do next?
- A Increase the number of training epochs until the training loss reaches zero
 - B Expand and diversify the held-out evaluation data, then review the training set
 - C Remove the validation set so that more examples are available for training
 - D Switch the endpoint to a higher temperature so the answers sound more confident

Answer: B

A gap between training prompts and new questions suggests overfitting or a narrow dataset, so diversifying the data and checking the evaluation set is the sensible next step. More epochs can deepen overfitting, which makes it tempting. Removing validation data hides the problem, and temperature does not fix a learned gap.

10. Which statement about deploying a fine-tuned model is most accurate?
- A The fine-tuned version permanently replaces the base model for every application at once
 - B A fine-tuned model can be used only from the notebook where it was trained
 - C Any change to the endpoint requires retraining the model from scratch on all data
 - D A candidate can be compared with the base model on the same test set first

Answer: D

Comparing the candidate and base models on the same evaluation set gives evidence before traffic is routed to the new version. Replacing the base model everywhere is risky and not required, which makes it tempting. Notebook-only use and retraining for every endpoint change are not requirements.