

PRACTICE QUESTIONS

SAP Certified - SAP Generative AI Developer

SHARED BY [u/Nimble-Comet-5940](#) on aiexamcerts

EDITION	Free Set 2
---------	------------

QUESTIONS	10
-----------	----

SHARED	October 15, 2026
--------	------------------

EXAM	SAP Certified - SAP Generative AI Developer
------	---

Shared by [u/Nimble-Comet-5940](#) on aiexamcerts

10 practice questions for the SAP Certified - SAP Generative AI Developer exam, each followed by its answer and a short explanation. More free sets and the full practice exam:

www.aiexamcerts.com/sap/sap-certified-sap-generative-ai-developer. Spotted a mistake? Email support@aiexamcerts.com with the exam name and the question number.

1. What is SAP AI Launchpad used for?

- A** A runtime that executes the inference containers for deployed models in the AI Core cluster of each tenant
- B** A vector database that stores embeddings for grounding and answers similarity queries from applications
- C** A management user interface for connecting to and operating one or more SAP AI Core instances
- D** A low-code tool for building SAP Fiori apps and their data bindings for business users

Answer: C

SAP AI Launchpad is the management UI that lets teams operate one or more AI Core instances, including their workflows and deployments. The runtime that executes containers is AI Core itself, which makes it a tempting but wrong description of the Launchpad. Vector storage belongs to SAP HANA Cloud and Fiori apps have their own tooling.

2. An application can call an AI Core inference endpoint with a named user's SAP ID password as its only credential.

- A** True
- B** False

Answer: B (False)

AI Core backend calls use an OAuth2 token obtained with the service key's client credentials. A named user's password is the tempting shortcut for interactive sign-in, but it is not the way application calls authenticate. The statement is false.

3. A team wants to switch from one LLM provider to another without rewriting the application. Which design choice supports this best?
- A Hard-code the provider's SDK calls in each function of the service so that each model is called directly from the code
 - B Embed the model weights directly in the CAP application package that is deployed to each environment
 - C Store the provider's API key in the browser so each user can switch between providers at any time
 - D Call the models through the Generative AI Hub's common interface and keep the model name in configuration

Answer: D

Using the Hub's common interface with the model name held in configuration means the switch is a configuration change. Hard-coding provider SDK calls is the tempting shortcut, and it forces code changes for every provider switch. Embedding weights or exposing keys in the browser are not sound designs for model access.

4. The maximum number of tokens a model can process in one request, prompt and output together, is its context _____.

Answer: window

The context window is the total token budget that a model accepts for one request. The token limit on output alone is a tempting reading, but the window covers both input and output.

5. Which two properties of vector retrieval have the largest effect on answer quality? (Choose two.)
- A Chunk size and overlap
 - B The number of worker lanes in the deployment
 - C The embedding model used for indexing and querying
 - D The font size of the answer text
 - E The number of browser tabs open during the session

Answer: A, C

Chunk size and overlap decide what each passage contains, and the embedding model decides how well meaning is matched. Worker lanes are a tempting capacity setting, but they affect throughput rather than relevance. Font size and browser tabs have no effect on retrieval.

6. Which statement about the orchestration service is correct?

- A It replaces the model provider and hosts the model weights itself inside the BTP subaccount
- B It composes modules such as templating, masking, filtering and grounding into one request
- C It works only with one hard-coded model version that the team selects during initial setup
- D It is required only when the application stores and searches embeddings for its search features

Answer: B

The orchestration service lets one request pass through several modules in a defined order, which keeps the pipeline in one place. Hosting weights itself is a tempting assumption, but orchestration calls provider models through the Hub. Model lock-in and an embeddings-only requirement are both incorrect.

7. An application needs a model to return JSON in a fixed structure. Which technique is the most direct?

- A Raise the temperature so the model varies its format
- B Remove all examples and instructions from the template
- C State the required output format in the prompt template and validate the reply before use
- D Send the same prompt twice in one request to double the output

Answer: C

Stating the output format in the template and validating the reply makes the structure explicit and catches deviations. Raising temperature increases variation, which makes it the tempting but wrong choice for structure. Removing instructions or repeating the prompt does not improve format compliance.

8. The same embedding model should be used to index documents and to embed the user's queries.

- A True
- B False

Answer: A (True)

Stored and query vectors must come from the same embedding model and dimension to be comparable. A different model for queries is a tempting optimization that silently breaks similarity scores. The statement is true.

9. A prompt works in testing but fails on long inputs. What is the most likely cause?
- A The SAP BTP subaccount is in a different region from the developers
 - B The service key has expired, which makes every call fail silently
 - C The model's name was written in uppercase in the configuration
 - D The input exceeds the context window, so content is truncated or the request is rejected

Answer: D

Failures that depend on input length point to the context window, where long prompts are cut short or refused. A regional mismatch is a tempting explanation for odd behavior, but it does not depend on input length. Expired keys cause authentication errors for all calls, and model name case is handled by the configuration.

10. Which instruction most reduces invented citations in a grounded answer?
- A Tell the model to answer only from the provided sources and to say when they do not contain the answer
 - B Tell the model to always produce at least five sources, even when the evidence is thin or missing from the context
 - C Ask the model to invent a plausible reference when it is unsure of the answer or of the source
 - D Raise the temperature so the model varies its source choices across repeated runs of the same question

Answer: A

Limiting answers to the supplied sources and allowing an explicit no-answer keeps the model from filling gaps with made-up references. Requiring five sources is a tempting rule, and it pushes the model to fabricate when the evidence is thin. Inventing references and raising temperature increase fabrication.