

PRACTICE QUESTIONS

IBM watsonx Generative AI Engineer v1 - Associate

SHARED BY [u/mellow_maple_58](#) on aiexamcerts

EDITION	Free Set 1
QUESTIONS	10
SHARED	October 15, 2026
EXAM	IBM watsonx Generative AI Engineer v1 - Associate

Shared by [u/mellow_maple_58](#) on aiexamcerts

10 practice questions for the IBM watsonx Generative AI Engineer v1 - Associate exam, each followed by its answer and a short explanation. More free sets and the full practice exam:

www.aiexamcerts.com/ibm/ibm-watsonx-generative-ai-engineer-v1-associate. Spotted a mistake? Email support@aiexamcerts.com with the exam name and the question number.

1. Which factor matters most when choosing between a large and a small foundation model for a customer support summary?

- A The vendor's marketing description of the model on its public web page
- B The required output quality weighed against latency, cost and context length
- C The number of programming languages in which the vendor publishes code samples
- D Whether the model's name appears in the first row of the vendor's catalog page

Answer: B

Model choice is a trade-off: larger models often produce better quality but cost more and respond more slowly, while a context window that is too short can truncate input. Marketing copy, code sample languages and catalog position do not affect whether a model meets the task. Judging by the name alone is the tempting shortcut.

2. The number of tokens a model can consider in a single request is called its _____ window.

Answer: context

The context window is the total budget for the prompt plus the generated output in one call. Long documents may need chunking or retrieval because they exceed it. The blank asks for the term for that limit.

3. What does zero-shot prompting mean?

- A Asking the model to perform a task with exactly zero instructions
- B Training the model on zero examples before deployment
- C Asking the model to perform a task with instructions only, and no worked examples
- D Asking the model to answer in a single word regardless of the task

Answer: C

Zero-shot prompting relies on the instruction alone, so the model must generalize from its pretraining. Few-shot prompting adds examples to the prompt. A prompt with zero instructions gives the model no task, and limiting the answer to one word has nothing to do with the number of examples.

4. Adding a few well-chosen examples to a prompt can change the output format the model returns.

- A True
- B False

Answer: A (True)

Few-shot examples demonstrate the expected pattern, and models often mirror the format they see. This is one of the most reliable ways to control structure. Examples should match the target format exactly.

5. Which TWO practices improve the quality of a RAG system? (Choose two.)

- A Remove all headings from documents before indexing
- B Store source metadata such as document name and date with each chunk
- C Index a single very large file that contains everything
- D Skip testing because retrieval is deterministic
- E Evaluate retrieval and generation separately on a labeled question set

Answer: B, E

Metadata supports filtering and citations, and separate evaluation shows whether retrieval or generation is the weak point. Removing headings discards useful structure, and combining everything into one file hurts precision. Retrieval can still return wrong passages, so skipping tests is unsafe.

6. A team needs the same answer each time for the same prompt in an audit trail. Which setting helps most?

- A Greedy decoding with a low or zero temperature
- B High temperature sampling with a large top-p value
- C A high repetition penalty set to its maximum
- D A long list of stop sequences

Answer: A

Greedy decoding picks the most likely token at each step, and low temperature reduces variation, which makes output repeatable. Higher sampling settings introduce randomness that breaks repeatability. A repetition penalty changes word choice, and stop sequences only end generation at a marker.

7. What does a stop sequence do during generation?

- A It makes the model stop reading the prompt after the first sentence
- B It tells the model to end generation when a specified string appears in its output
- C It sets the minimum number of tokens to generate
- D It stops the model from ever using the letter A

Answer: B

A stop sequence is a string that halts generation when produced, which is useful for ending at a delimiter. It does not stop the model from reading the prompt, and it does not set a minimum length. Banning a letter is not what the parameter does.

8. The numeric representation of text that allows semantically similar passages to be found by comparison is called an _____.

Answer: embedding

An embedding maps text to a vector in a space where closeness reflects meaning, which is the basis of vector retrieval. The blank asks for the term for that representation. Chunks are the units of text, while embeddings are their vector forms.

9. Why should credentials for watsonx.ai be read from environment variables rather than hard-coded in a notebook?
- A So secrets are not exposed when the code is shared or committed to a repository
 - B So the model learns the credentials for future tasks and can reuse them across every project in the workspace without any further setup
 - C So the credentials are cached by the model server to speed up later calls
 - D So the notebook can display the API key in its output

Answer: A

Environment variables and secret stores keep keys out of source code and shared files. Speed and model learning are unrelated to credential storage. Printing a key in output exposes it and is the opposite of the goal.

10. In LangChain, what is the role of a text splitter in a RAG pipeline?
- A It generates the final answer from the retrieved passages
 - B It divides loaded documents into chunks before they are embedded and stored
 - C It deletes duplicate users from the database after each chain run
 - D It sets the color of the output in a notebook

Answer: B

A text splitter, such as a recursive character splitter, breaks documents into chunks of a chosen size with optional overlap. Answer generation is done by the model, and user deduplication and output color are unrelated to splitting text.