

PRACTICE QUESTIONS

# IBM watsonx Generative AI Engineer v1 - Associate

SHARED BY [u/eager\\_router\\_42](#) on aiexamcerts

|           |                                                   |
|-----------|---------------------------------------------------|
| EDITION   | Free Set 2                                        |
| QUESTIONS | 10                                                |
| SHARED    | October 15, 2026                                  |
| EXAM      | IBM watsonx Generative AI Engineer v1 - Associate |

---

Shared by [u/eager\\_router\\_42](#) 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](https://www.aiexamcerts.com/ibm/ibm-watsonx-generative-ai-engineer-v1-associate). Spotted a mistake? Email [support@aiexamcerts.com](mailto:support@aiexamcerts.com) with the exam name and the question number.

---

1. A prototype works well on one model, but the team must reduce cost for a high-volume task. What is the most sensible next step?

- A Switch to the largest model available without testing
- B Remove the evaluation set so the comparison is faster
- C Benchmark a smaller model on a fixed test set and compare quality and cost
- D Keep the current model unchanged, since cost can only be reduced by cutting traffic

**Answer: C**

A fixed evaluation set lets the team compare candidate models on the same inputs and decide whether the quality loss is acceptable. Choosing the largest model increases cost, which is the opposite of the goal. Removing evaluation or measurement makes the decision unsupported.

---

2. When a model is replaced, prompts that worked before can produce different formats or content, so the team should re-test them.

- A True
- B False

**Answer: A (True)**

Different models interpret the same prompt differently, so output format, tone and accuracy can shift. Re-testing against an evaluation set catches these changes before users do. This is why prompts are treated as part of the model dependency.

---

3. Which prompt structure best separates instructions, context and the user's question?

- A One long paragraph with no punctuation or line breaks
- B Only the user's question, with the instructions omitted so the model infers the task
- C A table of random words to improve creativity
- D Clear labeled sections for the instruction, the reference context and the question

**Answer: D**

Labeled sections make it easier for the model to follow the task and reduce confusion between instructions and supplied data. A single unbroken paragraph mixes these roles and invites errors. Omitting instructions removes the task definition, and random words do not improve reliability.

---

4. Prompt tuning learns a small set of continuous vectors called soft \_\_\_\_\_ while the base model's weights remain unchanged.

**Answer: prompts** (also accepted: prompt embeddings)

Soft prompts are learned embedding vectors prepended to the input, and they are trained for a task while the foundation model stays frozen. This is the defining feature that separates prompt tuning from fine-tuning. The blank asks for the thing that is learned.

5. Which TWO practices are sound when building a watsonx.ai application? (Choose two.)

- A Print the full prompt with the API key to the console
- B Hard-code the project identifier and never make it configurable
- C Keep API keys in environment variables or a secret store
- D Disable retries so that every failure is seen immediately
- E Log request metadata and errors without storing sensitive user content unnecessarily

**Answer: C, E**

Secret storage and minimal logging protect credentials and privacy. Printing keys and hard-coding identifiers create exposure and inflexibility. Disabling retries removes a basic resilience measure for transient errors.

6. Which parameter limits the length of the generated output rather than the input?

- A Context window
- B Top-k value, which limits the candidate tokens to a fixed count at each step
- C Max new tokens
- D Random seed

**Answer: C**

The maximum number of new tokens caps how much the model generates in a response. The context window limits the combined prompt and output. Top-k limits which tokens are sampled, and the random seed controls reproducibility of sampling.

7. What does top-p (nucleus) sampling do?

- A It keeps only the first p tokens of the prompt
- B It charges the account a percentage p of the tokens generated
- C It rounds all probabilities to the nearest tenth
- D It samples from the smallest set of tokens whose cumulative probability reaches the chosen threshold

**Answer: D**

Top-p restricts sampling to the most probable tokens that add up to the threshold, which trims unlikely choices. Truncating the prompt, pricing and rounding are not what it does. Confusing it with top-k, which keeps a fixed number of candidates, is a common error.

8. Increasing the number of retrieved passages always makes an answer more accurate, because the model simply ignores irrelevant text.

- A True
- B False

**Answer: B** (False)

Extra passages add noise and consume context, and irrelevant text can mislead the model. The number of passages is a tuning choice that should be evaluated. More is not automatically better.

9. In a RAG pipeline, what does the embedding step do?

- A Converts text chunks and the user's question into vectors so they can be compared by similarity
- B Encrypts the documents before they are stored
- C Translates the answer into another language
- D Formats the final answer as a table with a header row, sorted columns and a footer that lists every source

**Answer: A**

Embeddings map text to vectors, so semantically similar passages are found by vector similarity search. Encryption, translation and formatting are separate steps and do not create the vectors used for retrieval.

10. What does a LangChain PromptTemplate do?

- A It stores the model's weights in a file
- B It sends an email when a prompt fails
- C It defines a prompt with named variables that are filled in at run time
- D It converts prompts into audio files that the application plays back through the device speaker

**Answer: C**

A prompt template defines a reusable structure with placeholders that are filled with values for each call. It does not store weights, send email or produce audio. Templates keep prompts consistent across the application.