

PRACTICE QUESTIONS

AI CERTs

AI+ Developer

SHARED BY [u/Serene-Crane-4635](#) on aiexamcerts

EDITION	Free Set 1
QUESTIONS	10
SHARED	October 15, 2026
EXAM	AI CERTs AI+ Developer

Shared by u/Serene-Crane-4635 on aiexamcerts

10 practice questions for the AI CERTs AI+ Developer exam, each followed by its answer and a short explanation. More free sets and the full practice exam: www.aiexamcerts.com/ai-certs/ai-developer. Spotted a mistake? Email support@aiexamcerts.com with the exam name and the question number.

1. A developer wants each project to have its own installed packages and versions. Which command creates an isolated environment?
- A pip freeze to lock versions
 - B python -m venv .venv
 - C python -m json.tool
 - D pip cache purge

Answer: B

The python -m venv command creates an isolated virtual environment with its own package set, which keeps dependency versions separate. pip freeze lists installed packages, json.tool formats JSON, and pip cache purge removes downloaded package files rather than creating environments.

2. A Python function that uses yield to produce values one at a time is called a _____.

Answer: generator

A generator function yields values lazily, which keeps memory use low when streaming large datasets or tokens. Calling it returns a generator object that can be iterated once. Generators are common when reading training data in batches.

3. Which sampling parameter controls how random the model's token choices are, with lower values producing more deterministic output?
- A max_tokens
 - B temperature
 - C stop sequence
 - D logit bias list

Answer: B

The temperature parameter scales the probability distribution over next tokens, so lower values make output more deterministic. max_tokens caps output length, a stop sequence ends generation at a marker, and a logit bias list adjusts token likelihoods for specific tokens.

4. The context window limits the total number of tokens the model can process in one request, including the prompt and the generated output.

- A True
- B False

Answer: A (True)

The context window is the total token budget for a call, so long prompts reduce the room left for the response. Exceeding it causes an error or truncation. Developers manage it by trimming history or retrieving only relevant text.

5. Which two techniques help a RAG system ground its answers in retrieved sources? (Choose two.)

- A Instructing the model to answer only from the provided context and say when the answer is missing
- B Removing the source identifiers from the retrieved chunks
- C Returning citations that link each claim to a source passage
- D Increasing temperature so the model explores more ideas
- E Disabling retrieval to reduce latency

Answer: A, C

Instructing the model to rely on the provided context and to admit gaps reduces unsupported claims, and citations let users check each claim. Removing source identifiers, raising temperature and disabling retrieval all weaken grounding.

6. Which prompting technique asks the model to show intermediate reasoning steps before the final answer on a multi-step problem?

- A Early stopping
- B Batch normalization
- C Chain-of-thought prompting
- D Cache warming

Answer: C

Chain-of-thought prompting elicits step-by-step reasoning, which often improves multi-step problem solving. Early stopping, batch normalization and cache warming are training or infrastructure techniques, not prompting techniques.

7. Which prompt element is most useful for making a model's output follow a fixed format that downstream code expects?

- A A lower frequency penalty on repeated words
- B A higher max_tokens value alone, without any format rules
- C A larger set of unrelated example topics
- D A clear output schema or example in the prompt

Answer: D

An explicit schema or example shows the exact fields and format, which makes parsing and validation easier. A lower frequency penalty only changes repetition, raising the token limit does not define the format, and unrelated example topics do not show the target structure.

8. A second model that re-scores retrieved passages for relevance before they are sent to the generator is called a _____.

Answer: reranker (also accepted: re-ranker, cross-encoder)

A reranker reorders the candidate passages with a more precise relevance model, so the generator receives the best evidence. It is used after a fast first-stage retrieval. Cross-encoders are a common type of reranker.

9. Which control best limits the cost and risk of a runaway agent that keeps calling tools?

- A Setting a maximum number of steps and a budget on tool calls per task
- B Removing the tool descriptions from the agent configuration entirely
- C Raising the model temperature so it stops repeating calls
- D Retrying every failed tool call without a limit

Answer: A

Step limits and call budgets stop an agent from looping indefinitely and bound the damage. Removing tool descriptions makes tool use impossible, a higher temperature does not stop loops, and unlimited retries make runaway cost worse.

10. Which split of a labeled dataset is used to make an unbiased final estimate of performance after all tuning decisions are made?

- A Training set
- B Test set
- C Augmentation set
- D Validation set

Answer: B

The test set is held out until the end and gives an unbiased estimate of performance after model and hyperparameter choices are final. The training set fits the model, augmentation adds variations, and the validation set guides tuning decisions, so it cannot give an unbiased final estimate.