

PRACTICE QUESTIONS

Cisco

810-110 AITECH

Cisco AI Technical Practitioner

SHARED BY [u/silvercrane1985](#) on aiexamcerts

EDITION Free Set 2

QUESTIONS 10

SHARED October 15, 2026

EXAM Cisco 810-110 AITECH, Cisco AI Technical Practitioner

Shared by [u/silvercrane1985](#) on aiexamcerts

10 practice questions for the Cisco 810-110 AITECH exam, Cisco AI Technical Practitioner, each followed by its answer and a short explanation. More free sets and the full practice exam:

www.aiexamcerts.com/cisco/810-110-aitech. Spotted a mistake? Email support@aiexamcerts.com with the exam code 810-110 AITECH and the question number.

1. Which term describes a numeric vector that represents the meaning of a word, sentence, or document so that similar items sit close together?

- A Embedding
- B Checkpoint
- C Token limit
- D Gradient

Answer: A

An embedding is a dense vector representation that places semantically similar items near each other in vector space. A checkpoint is a saved set of model weights, which makes it the most tempting confusion. A token limit is a context budget, and a gradient is a training quantity.

2. Idempotent workflow steps can be retried safely without creating duplicate side effects.

- A True
- B False

Answer: A (True)

An idempotent step produces the same result when repeated, so retries after a timeout do not duplicate emails, records, or charges. This matters for automation that runs unattended. The statement is True.

3. Which technique gives the model a role and an audience, such as 'You are a network engineer explaining this to a junior analyst'?

- A Token budgeting for response length
- B Persona or role prompting
- C Retrieval indexing of document chunks
- D Batch scheduling of nightly jobs

Answer: B

Persona prompting sets a role and audience that shape vocabulary, depth, and focus. Token budgeting controls length and cost, which makes it the tempting but different technique. Retrieval indexing and batch scheduling are system components unrelated to the wording of the prompt.

4. The pattern in which an agent alternates reasoning steps with tool actions and observes the results is called ReAct, short for reasoning and _____.

Answer: acting (also accepted: action)

ReAct interleaves reasoning traces with actions, so each tool result informs the next step. Planning every step in advance without observing results is a different pattern. The name combines reasoning and acting.

5. Which two statements about large language model outputs are accurate? (Choose two.)

- A Outputs are probabilistic and can vary between runs
- B Outputs are verified facts because the model checks a fact database before answering
- C Outputs can contain confident errors, so important facts need verification
- D Outputs are identical for any prompt that contains the same words in a different order
- E Outputs cannot be longer than the prompt that produced them

Answer: A, C

Sampling makes outputs probabilistic, and models can state errors with confidence, so important claims need checking. The claim that outputs are verified is false because models do not consult a verified database by default, which makes it the most tempting error. The last two statements are wrong, since word order changes the input and outputs can be longer than prompts.

6. Which trigger suits a workflow that must react to a new support ticket within seconds?

- A A monthly batch job that reads all tickets in one pass
- B A manual export that a person runs each Friday
- C An event-driven webhook that fires on ticket creation
- D A nightly report emailed to a manager

Answer: C

An event-driven webhook reacts as soon as the ticket is created, which meets a seconds-level requirement. A monthly batch job is the most tempting cheap option, but it cannot respond in time. Manual exports and nightly reports are too slow for this need.

7. An agent repeatedly calls the same search tool with the same arguments and never finishes. What is the most appropriate guardrail?

- A Raising the temperature so the agent varies its search
- B Removing the search tool entirely so the agent has nothing left to call
- C Letting the loop continue until the user gives up
- D A step budget and loop detection that stops or escalates the run

Answer: D

Step budgets and repetition detection stop runaway loops and hand the task to a person or a fallback. Raising temperature may add variety but does not guarantee progress, which makes it a tempting but weak fix. Removing the tool breaks the task, and an unbounded loop wastes money.

8. Personal data sent to a third-party model provider should be covered by a data processing agreement that sets the permitted use and retention terms.

- A True
- B False

Answer: A (True)

Third-party processors need contractual terms that define permitted use, retention, and deletion, and a data processing agreement provides them. Without those terms, the organization cannot show it controlled the data. The statement is True.

9. Which control best prevents an LLM application from leaking one customer's data through retrieved documents?

- A Enforce access control on retrieval so users see only allowed documents
- B Increasing the number of retrieved documents for every user, every time
- C Removing access logs so retrieval is faster
- D Sharing one index across all users without filters

Answer: A

Filtering retrieval by the user's permissions prevents cross-customer exposure at the source. Retrieving more documents widens exposure, which makes it the most tempting but harmful option. Removing logs and sharing an unfiltered index both remove protection.

10. Which factor matters most when choosing between a hosted model and a self-hosted open-weight model?

- A Whichever option has the lowest list price per token, ignoring all other factors
- B Data sensitivity, cost at expected volume, and operating capacity
- C The number of programming languages the vendor's SDK supports, on its own
- D The total parameter count, on the assumption that a larger model always wins

Answer: B

Data sensitivity, cost at expected volume, and operating capacity drive the hosting decision. The lowest token price alone ignores data handling and the effort of running the model. SDK language coverage and parameter count do not decide where a model should run.