

PRACTICE QUESTIONS

CertiProf

CPEFPC

Prompt Engineering Foundation

SHARED BY [u/humblepelican1987](#) on aiexamcerts

EDITION Free Set 1

QUESTIONS 10

SHARED October 15, 2026

EXAM CertiProf CPEFPC, Prompt Engineering Foundation

Shared by [u/humblepelican1987](#) on aiexamcerts

10 practice questions for the CertiProf CPEFPC exam, Prompt Engineering Foundation, each followed by its answer and a short explanation. More free sets and the full practice exam:

www.aiexamcerts.com/certiprof/cpefpc. Spotted a mistake? Email support@aiexamcerts.com with the exam code CPEFPC and the question number.

1. Which description of a token is most accurate for an LLM usage meter?
- A A paid license key that unlocks one API call for a single user session
 - B A unit of text, often a word fragment, that the model reads and produces
 - C A fixed-length vector stored with each user account in the database
 - D A signed certificate that proves who wrote the prompt and when it was sent

Answer: B

A token is the unit of text, often a whole word or a piece of one, that the model reads and generates. Pricing and context limits are usually counted in tokens, which makes the license key a tempting confusion. Vectors and certificates have their own roles and are not what a token is.

2. A large language model produces text by predicting the next _____, a small unit of text.

Answer: token

The model predicts the next token, a small unit of text that may be a whole word or part of one. The word character is a tempting alternative, but models operate on tokens. The blank asks for the unit that the model generates one step at a time.

3. Which prompt component provides the source material the model should use for its answer?
- A The temperature setting
 - B Context or input data
 - C The model's version number
 - D The max tokens limit

Answer: B

Context or input data is the material the model should work from, such as a document or a customer record. The temperature setting controls randomness and is a tempting but separate request parameter. The version number and max tokens limit do not provide source material.

4. It is helpful to specify the output format, such as a table or bullet list, when a prompt needs a particular layout.
- A True
 - B False

Answer: A (True)

Stating the format removes guesswork and makes the result easier to use or to parse. Layout instructions are one of the most reliable ways to control the shape of an answer. Leaving format unspecified often produces a reply that needs manual rework.

5. Which two statements about chain-of-thought prompting are accurate? (Choose two.)
- A It increases the number of output tokens the model produces
 - B It guarantees a correct final answer every time
 - C It removes the need for examples in every task
 - D It can improve results on multi-step reasoning tasks
 - E It works only in image models

Answer: A, D

Chain-of-thought can improve results on multi-step reasoning and it usually increases the number of output tokens. It does not guarantee correctness, which is the tempting overclaim. It does not replace examples in every task and it is not limited to image models.

6. What is few-shot prompting?
- A Limiting the reply to a few words so that the answer stays short and easy to read
 - B Including a small number of worked input and output examples before the new input
 - C Sending the prompt to a few different models at once to compare their answers
 - D Training the model on a few gigabytes of new data collected from the company's files

Answer: B

Few-shot prompting places a small number of worked input and output examples before the new input so the model copies the pattern. Limiting reply length is a tempting mix-up because the word few appears, but it is a different idea. Sending to several models and training on new data are different activities.

7. What does the max tokens setting control in a request?

- A The size of the model's training dataset
- B The number of users allowed in the session
- C The minimum confidence score for each word
- D The maximum length of the generated reply

Answer: D

The max tokens setting caps the length of the generated reply, which is useful for cost and for keeping answers short. The training dataset is fixed when the model is built, which makes it a tempting but unrelated number. User limits and confidence thresholds are not controlled by this setting.

8. The sampling setting that scales the probabilities before each token is chosen, so that higher values give more varied text, is called _____.

Answer: temperature

Temperature reshapes the probability distribution before sampling, so higher values give more variety. Top-p is a related setting that limits the candidate tokens, which is a tempting confusion. Temperature is the setting asked about.

9. What is grounding in prompting?

- A Basing the answer on supplied source material, and ideally citing it
- B Connecting the model to a physical robot that can move objects in the real world
- C Lowering the temperature to zero for every single request sent to the model
- D Encrypting the prompt before sending it so that the vendor cannot read the text

Answer: A

Grounding means the answer is based on supplied source material and, ideally, cites the passages that support it. Connecting to a robot is a physical integration and not what grounding means. Temperature and encryption are separate settings and controls.

10. A long conversation starts to ignore instructions given at the beginning. What is the most likely cause?

- A The model has permanently forgotten its system message in its weights after the update
- B The earlier content has been pushed out of, or diluted within, the context window
- C The temperature was set above 2 so that the replies are too random to follow
- D The user's account has been locked by the platform after too many requests

Answer: B

As a conversation grows, early instructions can fall out of the context window or become less influential among many tokens. Permanent forgetting in the weights is a tempting explanation, but the context is the cause here. A temperature setting or an account lock would produce different symptoms.