

PRACTICE QUESTIONS

ARTiBA

Chartered AI Engineering Professional

SHARED BY [u/quick_llama_96](#) on aiexamcerts

EDITION Free Set 1

QUESTIONS 10

SHARED October 15, 2026

EXAM ARTiBA Chartered AI Engineering Professional

Shared by [u/quick_llama_96](#) on aiexamcerts

10 practice questions for the ARTiBA Chartered AI Engineering Professional exam, each followed by its answer and a short explanation. More free sets and the full practice exam:

www.aiexamcerts.com/artiba/chartered-ai-engineering-professional. Spotted a mistake? Email support@aiexamcerts.com with the exam name and the question number.

1. A bank needs an explanation for each credit decision and must audit the model version used for every decision. Which design addition is most important?
 - A A larger ensemble of models that averages out individual errors so that no single decision is questioned later
 - B A dashboard that shows the average prediction for all customers grouped by postal code each week for review
 - C A caching layer that stores the last twenty-four hours of predictions so that repeated requests return faster
 - D A model registry that links each deployed version to its training data snapshot, metrics, and approval record

Answer: D

An audit needs to tie each decision to an exact model version and the evidence behind its release, which a registry with lineage provides. Ensembles and dashboards may improve performance or visibility, but they do not create an audit trail. The caching option is tempting because it records predictions, but it does not record how each model was built or approved.

2. Replacing direct identifiers with a keyed token so that records can still be linked without exposing the raw identifier is called _____.

Answer: pseudonymization (also accepted: pseudonymisation)

Pseudonymization replaces identifiers with tokens, keeping records linkable for analysis while reducing exposure. Because re-identification is still possible with the key, pseudonymized data remains personal data under GDPR. Full anonymization is a stronger and different standard.

3. Which technique reduces the memory needed to fine-tune a large language model by training small low-rank adapter matrices while keeping the base weights frozen?
- A Knowledge distillation, which trains a smaller student model to imitate the output distribution of a larger teacher model
 - B LoRA (low-rank adaptation)
 - C Quantization-aware pruning, which removes weights whose magnitudes fall below a fixed threshold after training completes
 - D Mixture-of-experts routing, which activates a different subset of feed-forward blocks for each input token processed

Answer: B

LoRA freezes the pretrained weights and learns small low-rank matrices, which sharply reduces trainable parameters and optimizer memory. Distillation, pruning, and mixture-of-experts are real techniques, but they solve other problems such as model size or routing. Distillation is the most tempting distractor because it also produces a smaller model.

4. A smaller epsilon value in differential privacy means weaker privacy protection for individuals in the dataset.
- A True
 - B False

Answer: B (False)

A smaller epsilon limits how much any one record can change the output, which means stronger privacy and typically lower accuracy. The statement reverses that relationship. The statement is false.

5. Which two actions help when a deep learning model's training loss keeps falling but its validation loss starts rising after a few epochs? (Choose two.)
- A Increase the batch size to double the number of examples used in each gradient update
 - B Remove the validation set and train longer on the full data to reduce the variance of the estimate
 - C Stop training early using the validation loss as the stopping criterion
 - D Add stronger regularization or data augmentation to the training pipeline
 - E Increase the depth of the network so that it can represent more complex patterns in the data

Answer: C, D

A rising validation loss alongside falling training loss is the classic overfitting signal, so early stopping and stronger regularization or augmentation address it directly. Larger batches, removing validation, and deeper networks all make overfitting more likely or hide it. Adding depth is tempting because it sounds like it adds capacity to learn, which is the problem here.

6. Which tool is most commonly used as a model registry in MLOps workflows to track model versions and their stage transitions?
- A Apache Spark, a distributed engine that partitions shuffle output across executors to reduce network traffic
 - B Great Expectations, a library that defines and runs data validation suites against incoming tables and files
 - C MLflow Model Registry
 - D Terraform, an infrastructure-as-code tool that provisions cloud resources from declarative configuration files

Answer: C

The MLflow Model Registry stores model versions, their metadata, and transitions between stages such as staging and production. Spark, Great Expectations, and Terraform all play roles in data and infrastructure work, but none is a model registry. Great Expectations is the most tempting distractor because it also tracks data quality expectations.

7. Which setting most directly controls how much randomness appears in an LLM's sampled output?
- A Temperature
 - B Context window length, which sets the maximum number of tokens the model can consider in one request at a time
 - C Embedding dimension, which sets the size of the vector that represents each token inside the network
 - D Batch size, which sets how many prompts are grouped together in each forward pass on the GPU

Answer: A

Temperature scales the probability distribution used for sampling, so lower values make output more deterministic and higher values make it more varied. Context window length limits input, embedding dimension sets internal representation size, and batch size affects throughput. The context window option is tempting because it is the setting people most often tune for long prompts.

8. Packaging a model together with its dependencies into a portable container image is one common step in model _____.

Answer: deployment

Containerizing a model captures its runtime and dependencies so the same artifact runs consistently across environments. This packaging step is part of getting a model into production. It is not a training step, and it does not change the model's weights.

9. A deployed fraud model's precision drops sharply, but its input distributions look unchanged. Which explanation should the team investigate first?
- A A rise in the number of requests per second, which increases load on the serving cluster without changing predictions
 - B Hardware throttling of the inference GPU, which affects only latency and never changes the quality of predictions
 - C Feature scaling applied twice, which would normally have caused a training failure rather than a production drop
 - D Concept drift, where the relationship between the inputs and fraud outcomes has changed

Answer: D

Stable inputs with falling precision point to a change in the input-to-outcome relationship, which is concept drift, such as new fraud patterns. Load and hardware throttling affect latency and throughput rather than prediction quality. Double scaling is tempting because it sounds like a preprocessing bug, but it would usually surface during training.

10. When tuning a model with k-fold cross-validation on time-ordered data, which change is needed to avoid leakage?
- A Shuffle the rows before each fold so that every time period appears in both the training and validation folds, which gives each fold a balanced view of the data
 - B Use time-based splits in which validation folds always come after the training data they are evaluated against
 - C Average the validation scores across folds and report only the single best fold as the final estimate
 - D Fit the scaler on the full dataset before splitting so that every fold uses the same normalization constants

Answer: B

Time-ordered data must not let validation rows precede training rows, or the model learns from the future. Shuffling mixes periods across folds and fitting preprocessing on all data leaks information, which makes both the shuffle and the global scaler tempting but wrong. Reporting only the best fold inflates the estimate.