

PRACTICE QUESTIONS

ARTiBA

Chartered AI Engineering Professional

SHARED BY [u/Gentle-Switch-3920](#) on aiexamcerts

EDITION	Free Set 2
QUESTIONS	10
SHARED	October 15, 2026
EXAM	ARTiBA Chartered AI Engineering Professional

Shared by [u/Gentle-Switch-3920](#) 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. Which storage format is designed for analytical queries that read only a few columns across large datasets?
- A Apache Avro, a row-oriented format commonly used for streaming records between services as compact binary messages
 - B Plain JSON Lines, a text format that stores each record on its own line with all of its fields together
 - C Apache Parquet, a columnar format with per-column compression and statistics
 - D Protocol Buffers, a schema-based serialization format used for message exchange between microservices in a platform

Answer: C

Parquet stores data by column with compression and statistics, so queries read only the columns they need. Avro, JSON Lines, and Protocol Buffers store records row by row and suit messaging or interchange. Avro is the most tempting distractor because it is also a widely used data format, but it is row-oriented.

2. A microservice that performs model inference should keep per-user session state in its own memory so that replicas can handle returning users without a shared store.
- A True
 - B False

Answer: B (False)

Per-process memory breaks when requests for the same user land on different replicas, so session state belongs in a shared store or the client. Keeping it local makes scaling and failover unreliable. The statement is false.

3. Which pattern lets a serving system scale model replicas automatically based on request concurrency or resource utilization?
- A Horizontal autoscaling of a containerized model server, such as a Kubernetes HorizontalPodAutoscaler
 - B Vertical scaling by replacing the model file with a larger version that has more parameters and a bigger file size
 - C Scheduled restarts at fixed intervals so that memory leaks are cleared before peak traffic arrives each day
 - D Moving the model to a single larger server and disabling the load balancer to keep the configuration simple

Answer: A

A horizontal autoscaler adds or removes replicas as measured load changes, which matches variable inference demand. Vertical scaling and larger model files change capacity per instance rather than adapting it. Scheduled restarts are tempting because they clear memory, but they do not respond to demand.

4. A control that limits how much a single client can call an API within a time window is called rate _____.

Answer: limiting

Rate limiting caps the number of requests a client can make in a window, protecting the service from abuse and overload. It is commonly enforced at an API gateway. Authentication identifies the caller, while rate limiting controls how often the caller may act.

5. Which two practices help limit how much personal data a trained model can memorize? (Choose two.)
- A Training for many more epochs so that rare records are fully learned before the data is cleaned
 - B Removing direct identifiers and minimizing the fields used for training
 - C Storing the training data in an encrypted archive while keeping the decryption key in the same bucket
 - D Increasing the parameter count so that each individual record has less influence on the model's outputs
 - E Applying differential privacy with a defined privacy budget during training

Answer: B, E

Minimizing fields and removing identifiers reduce what the model can absorb, and differential privacy bounds how much any one record can shape the model. Training longer increases memorization, and encryption that keeps its key beside the archive protects nothing. Growing the model is tempting, but larger models tend to memorize more, not less.

6. Which practice most reduces the risk that a RAG system answers from outdated documents?
- A Increasing the number of retrieved passages to fifty so that the newest document is more likely to be included
 - B Fine-tuning the model once on the full document set and never updating the index again after the release
 - C Lowering the similarity threshold so that the system retrieves any passage that loosely matches the user's query
 - D Re-indexing source documents when they change and attaching source timestamps to the retrieved passages

Answer: D

A RAG index is only as current as its last update, so re-indexing on change and showing timestamps keeps answers and their sources current. Retrieving more passages or loosening thresholds adds noise without solving staleness. Fine-tuning once is tempting because it sounds thorough, but the model's weights would then go stale too.

7. Which statement about calibration is correct?
- A A well-calibrated classifier is one whose ROC curve reaches the top-left corner of the plot, which means that it ranks every positive above every negative example
 - B A well-calibrated classifier's predicted probability of 0.8 should correspond to about 80 percent of such predictions being positive
 - C Calibration measures how many features a model uses, so using fewer features always produces better calibration overall
 - D A calibrated model never misclassifies an example, because its predicted probabilities are pushed to exactly zero or one

Answer: B

Calibration asks whether predicted probabilities match observed frequencies, so a 0.8 prediction should be right about 80 percent of the time. Discrimination measured by ROC AUC is a different property, and calibration says nothing about feature counts. The zero-or-one option is tempting because confident predictions look precise, but calibration is about matching frequencies.

8. A model registry should store the approval status and the training data version for each model version, not only the model file itself.
- A True
 - B False

Answer: A (True)

A registry that holds only artifacts cannot show which data, code, and approval produced a version, which weakens audit and rollback. Recording lineage and approval status is standard practice. The statement is true.

9. Which design choice best supports a multi-model AI system in which different components evolve on different schedules?
- A Bundling all models into one container image that is rebuilt and redeployed together for every single change, so that the whole system always ships as one unit
 - B Sharing a single global variable among all components so that they always see exactly the same state
 - C Defining stable service interfaces with versioned contracts between components so that each can be deployed independently
 - D Routing all internal calls through one shared notebook that each team edits before every release goes out

Answer: C

Versioned interfaces let each component change on its own schedule as long as it honors its contract, which is the basis of independent deployability. Bundling everything forces joint releases, and shared global state couples components tightly. The single-image option is tempting because it simplifies packaging, but it removes the independence the design needs.

10. Which signal most directly indicates that a production model needs investigation before its business metrics are visible?
- A A statistically significant shift in the distribution of key input features compared with the training baseline
 - B A single failed health check on one replica during a routine node restart in the cluster that was resolved quickly
 - C A short increase in average response latency during a daily backup window on the database that runs overnight
 - D A slightly longer model training time than the previous run on the same hardware configuration last month

Answer: A

Input drift is an early warning that the model may be operating outside what it learned, often before labeled outcomes arrive. Transient health checks and latency during backups concern infrastructure, and training time is not a production signal. The health check option is tempting because it is an immediate alarm, but it reflects infrastructure rather than model behavior.