IBM AI Engineering Professional Certificate

Quiz Answers 2026

102 solved MCQs with explanations, organised by course: Machine Learning with Python, Keras, TensorFlow, PyTorch, Generative AI and LLMs, RAG and LangChain.

Free PDF from mcqstop.com

Updated 16 November 2026. Online version with new questions: mcqstop.com/ibm-ai-engineering-professional-certificate-answers/

These are practice questions covering the course topics, written for revision. They are not copied from IBM's graded assessments.

Course 1: Machine Learning with Python

1. In machine learning, what is the difference between supervised and unsupervised learning?

A. Supervised uses labelled data; unsupervised finds patterns in unlabelled data
B. Supervised is faster
C. Unsupervised needs more labels
D. There is no difference

Answer: A. Supervised uses labelled data; unsupervised finds patterns in unlabelled data
Supervised models learn a mapping from inputs to known outputs; unsupervised models discover structure such as clusters.

2. Which of the following is a regression problem?

A. Predicting whether an email is spam
B. Predicting the CO2 emission of a car from its engine size
C. Grouping customers into segments
D. Detecting objects in an image

Answer: B. Predicting the CO2 emission of a car from its engine size
Regression predicts a continuous value; the course uses the CO2 emissions dataset for this.

3. Which of the following is a classification problem?

A. Predicting house prices
B. Predicting whether a tumour is benign or malignant
C. Forecasting temperature
D. Estimating salary

Answer: B. Predicting whether a tumour is benign or malignant
Classification predicts a discrete category.

4. In simple linear regression, the model fits:

A. A curve through all points
B. A straight line that minimises the sum of squared errors
C. A decision boundary
D. Clusters

Answer: B. A straight line that minimises the sum of squared errors
Ordinary least squares finds the line with the lowest mean squared error.

5. Which metric is commonly used to evaluate a regression model?

A. Accuracy
B. Mean Squared Error (MSE)
C. F1 score
D. Jaccard index

Answer: B. Mean Squared Error (MSE)
MSE, RMSE, MAE and R-squared are regression metrics.

6. What does R-squared measure?

A. Classification accuracy
B. The proportion of variance in the target explained by the model
C. The number of features
D. The learning rate

Answer: B. The proportion of variance in the target explained by the model
R-squared of 1 means a perfect fit; 0 means no better than the mean.

7. What is the purpose of splitting data into training and test sets?

A. To make the dataset smaller
B. To estimate how the model performs on unseen data
C. To speed up training
D. To remove outliers

Answer: B. To estimate how the model performs on unseen data
Out-of-sample evaluation shows how well the model generalises.

8. Which algorithm classifies a point based on the majority class of its nearest neighbours?

A. Logistic regression
B. K-Nearest Neighbours (KNN)
C. K-means
D. Linear regression

Answer: B. K-Nearest Neighbours (KNN)
KNN is a lazy, instance-based classifier.

9. In KNN, what happens if K is too small?

A. The model underfits
B. The model becomes sensitive to noise and overfits
C. Training takes longer
D. Accuracy always improves

Answer: B. The model becomes sensitive to noise and overfits
A small K captures noise; a very large K over-smooths.

10. A decision tree chooses the attribute to split on by:

A. Random selection
B. Maximising information gain (reducing entropy)
C. Alphabetical order
D. Minimising the number of nodes

Answer: B. Maximising information gain (reducing entropy)
Entropy measures impurity; the best split lowers it most.

11. Logistic regression is used for:

A. Predicting continuous values
B. Binary classification, outputting a probability
C. Clustering
D. Dimensionality reduction

Answer: B. Binary classification, outputting a probability
It applies the sigmoid function to a linear combination of features.

12. The sigmoid function outputs values between:

A. -1 and 1
B. 0 and 1
C. 0 and infinity
D. -infinity and infinity

Answer: B. 0 and 1
The output can be read as a probability.

13. What does a Support Vector Machine (SVM) try to find?

A. The shortest tree
B. The hyperplane with the maximum margin between classes
C. The nearest neighbours
D. The centroid of each cluster

Answer: B. The hyperplane with the maximum margin between classes
SVM maximises the distance between the hyperplane and the closest points (support vectors).

14. What is the 'kernel trick' in SVM?

A. Removing features
B. Mapping data into a higher-dimensional space to make it linearly separable
C. Speeding up training
D. Reducing the dataset

Answer: B. Mapping data into a higher-dimensional space to make it linearly separable
Common kernels: linear, polynomial, RBF, sigmoid.

15. K-means is an example of:

A. Supervised classification
B. Unsupervised clustering
C. Regression
D. Reinforcement learning

Answer: B. Unsupervised clustering
K-means partitions data into K clusters around centroids.

16. In K-means, the algorithm stops when:

A. Accuracy reaches 100%
B. Centroids no longer move significantly
C. K equals the number of points
D. The tree is fully grown

Answer: B. Centroids no longer move significantly
It iterates assign-and-update until convergence.

17. Which metric evaluates a classification model by comparing predicted and actual labels in a table?

A. R-squared
B. Confusion matrix
C. MSE
D. Silhouette score

Answer: B. Confusion matrix
From it you compute precision, recall and F1.

18. Precision is defined as:

A. TP / (TP + FN)
B. TP / (TP + FP)
C. TN / (TN + FP)
D. (TP + TN) / total

Answer: B. TP / (TP + FP)
Precision is the share of predicted positives that are correct.

19. Recall is defined as:

A. TP / (TP + FP)
B. TP / (TP + FN)
C. TN / (TN + FN)
D. FP / (FP + TN)

Answer: B. TP / (TP + FN)
Recall is the share of actual positives that were found.

20. Which Python library is mainly used for classical machine learning in the course?

A. TensorFlow
B. scikit-learn
C. PyTorch
D. Keras

Answer: B. scikit-learn
scikit-learn provides regression, classification, clustering and evaluation tools.

21. What is overfitting?

A. The model is too simple
B. The model fits training data too closely and performs poorly on new data
C. The model trains too fast
D. The data is too large

Answer: B. The model fits training data too closely and performs poorly on new data
High training accuracy with low test accuracy signals overfitting.

22. Which technique helps evaluate a model more reliably than a single train/test split?

A. Using all data for training
B. K-fold cross-validation
C. Removing the test set
D. Increasing K in KNN

Answer: B. K-fold cross-validation
Each fold serves once as the test set.

Course 2: Introduction to Deep Learning and Neural Networks with Keras

23. What is an artificial neuron's output computed from?

A. The sum of inputs only
B. A weighted sum of inputs plus a bias, passed through an activation function
C. A random number
D. The number of layers

Answer: B. A weighted sum of inputs plus a bias, passed through an activation function
$z = w.x + b$, then $a = f(z)$.

24. What is forward propagation?

A. Updating weights
B. Passing input data through the network layer by layer to produce an output
C. Shuffling data
D. Removing neurons

Answer: B. Passing input data through the network layer by layer to produce an output
Each layer transforms the previous layer's output.

25. What is backpropagation?

A. Moving data backwards to the user
B. Computing gradients of the loss with respect to weights and propagating them backward to update weights
C. Deleting layers
D. Normalising inputs

Answer: B. Computing gradients of the loss with respect to weights and propagating them backward to update weights
It applies the chain rule from the output layer back to the input.

26. What does gradient descent do?

A. Increases the loss
B. Iteratively adjusts weights in the direction that reduces the loss
C. Chooses the activation function
D. Adds layers

Answer: B. Iteratively adjusts weights in the direction that reduces the loss
The learning rate controls the step size.

27. What is the 'vanishing gradient' problem?

A. Gradients become too large
B. Gradients become very small in early layers, slowing or stopping learning
C. Weights are deleted
D. The dataset is too small

Answer: B. Gradients become very small in early layers, slowing or stopping learning
It occurs with sigmoid/tanh activations in deep networks.

28. Which activation function helps overcome the vanishing gradient problem?

A. Sigmoid
B. ReLU
C. Tanh
D. Step function

Answer: B. ReLU
ReLU outputs max(0, x) and does not saturate for positive inputs.

29. The softmax function is typically used in:

A. Hidden layers
B. The output layer for multi-class classification
C. The input layer
D. Regression outputs

Answer: B. The output layer for multi-class classification
It converts scores into probabilities that sum to 1.

30. In Keras, which class is used to build a model layer by layer?

A. Sequential
B. Functional
C. Pipeline
D. Layerwise

Answer: A. Sequential
model = Sequential(); model.add(Dense(...)).

31. Which Keras layer is a fully connected layer?

A. Conv2D
B. Dense
C. MaxPooling2D
D. LSTM

Answer: B. Dense
Dense connects every input to every output neuron.

32. For a regression model in Keras, a common loss function is:

A. categorical_crossentropy
B. mean_squared_error
C. binary_crossentropy
D. hinge

Answer: B. mean_squared_error
MSE is standard for continuous targets.

33. For multi-class classification in Keras, a common loss function is:

A. mean_squared_error
B. categorical_crossentropy
C. mean_absolute_error
D. huber

Answer: B. categorical_crossentropy
Used with a softmax output layer.

34. Which optimiser is most commonly used as a default in Keras?

A. SGD
B. Adam
C. Adagrad
D. Momentum

Answer: B. Adam
Adam adapts the learning rate for each parameter.

35. What is an 'epoch'?

A. One weight update
B. One complete pass through the entire training dataset
C. One layer
D. One batch

Answer: B. One complete pass through the entire training dataset
Training usually runs for many epochs.

36. What is a 'shallow' vs 'deep' neural network?

A. Shallow has one hidden layer; deep has many
B. Shallow uses images; deep uses text
C. Shallow is slower
D. No difference

Answer: A. Shallow has one hidden layer; deep has many
Depth lets networks learn hierarchical features.

37. What is a Convolutional Neural Network (CNN) best suited for?

A. Tabular data
B. Image data
C. Time series only
D. Text only

Answer: B. Image data
Convolution and pooling layers extract spatial features.

38. What does a pooling layer do in a CNN?

A. Adds parameters
B. Reduces spatial dimensions and computation
C. Increases image size
D. Normalises colours

Answer: B. Reduces spatial dimensions and computation
Max pooling keeps the largest value in each window.

39. What is a Recurrent Neural Network (RNN) designed for?

A. Images
B. Sequential data such as text and time series
C. Tabular data
D. Clustering

Answer: B. Sequential data such as text and time series
RNNs keep a hidden state across time steps.

40. What is an LSTM?

A. A type of CNN
B. A recurrent network with gates that handles long-term dependencies
C. An optimiser
D. A loss function

Answer: B. A recurrent network with gates that handles long-term dependencies
Long Short-Term Memory networks mitigate vanishing gradients in sequences.

41. What is an autoencoder?

A. A supervised classifier
B. An unsupervised network that learns to compress and reconstruct its input
C. An optimiser
D. A pooling layer

Answer: B. An unsupervised network that learns to compress and reconstruct its input
Used for dimensionality reduction and denoising.

42. What does 'dropout' do?

A. Deletes the dataset
B. Randomly disables neurons during training to reduce overfitting
C. Speeds up inference
D. Increases learning rate

Answer: B. Randomly disables neurons during training to reduce overfitting
A regularisation technique.

Course 3: Deep Learning with Keras and TensorFlow

43. In TensorFlow 2, which mode is enabled by default?

A. Graph mode
B. Eager execution
C. Lazy mode
D. Static mode

Answer: B. Eager execution
Eager execution evaluates operations immediately.

44. What is a tensor?

A. A type of neuron
B. A multi-dimensional array
C. A loss function
D. A dataset

Answer: B. A multi-dimensional array
Scalars, vectors and matrices are tensors of rank 0, 1 and 2.

45. What is transfer learning in deep learning?

A. Training from scratch
B. Reusing a pre-trained model's weights for a new but related task
C. Moving data between GPUs
D. Copying datasets

Answer: B. Reusing a pre-trained model's weights for a new but related task
Common with ImageNet-trained models such as VGG16 and ResNet.

46. When fine-tuning a pre-trained model, 'freezing' layers means:

A. Deleting them
B. Keeping their weights fixed during training
C. Doubling them
D. Randomising them

Answer: B. Keeping their weights fixed during training
Usually early layers are frozen and later layers trained.

47. What is data augmentation?

A. Buying more data
B. Creating modified versions of training data, such as rotated or flipped images
C. Removing outliers
D. Compressing data

Answer: B. Creating modified versions of training data, such as rotated or flipped images
It improves generalisation.

48. What does a Keras 'callback' such as EarlyStopping do?

A. Starts training early
B. Stops training when a monitored metric stops improving
C. Skips layers
D. Deletes checkpoints

Answer: B. Stops training when a monitored metric stops improving
Prevents overfitting and saves time.

49. What is the Keras Functional API used for?

A. Only simple stacks
B. Models with multiple inputs/outputs or non-linear topology
C. Data loading
D. Plotting

Answer: B. Models with multiple inputs/outputs or non-linear topology
More flexible than Sequential.

50. What is batch normalisation?

A. Normalising labels
B. Normalising layer inputs per mini-batch to stabilise and speed training
C. Shuffling batches
D. Reducing batch size

Answer: B. Normalising layer inputs per mini-batch to stabilise and speed training
It also has a mild regularising effect.

51. What is a Transformer's key mechanism?

A. Convolution
B. Self-attention
C. Pooling
D. Recurrence

Answer: B. Self-attention
Self-attention relates every token to every other token.

52. What is a GAN?

A. A regression model
B. Two networks, a generator and a discriminator, trained in competition
C. A clustering algorithm
D. An optimiser

Answer: B. Two networks, a generator and a discriminator, trained in competition
Generative Adversarial Networks produce realistic synthetic data.

53. What is a diffusion model?

A. A classifier
B. A generative model that learns to reverse a gradual noising process
C. A pooling method
D. A tokeniser

Answer: B. A generative model that learns to reverse a gradual noising process
Used for image generation.

54. What is 'reinforcement learning' with Keras/TensorFlow typically used for in the course?

A. Image classification
B. Training agents (e.g. Q-learning) to maximise reward
C. Clustering
D. Translation

Answer: B. Training agents (e.g. Q-learning) to maximise reward
Deep Q-Networks combine Q-learning with neural networks.

Courses 4-5: Neural Networks and Deep Learning with PyTorch

55. In PyTorch, what is a tensor's 'requires_grad' attribute for?

A. Printing the tensor
B. Telling autograd to track operations for gradient computation
C. Changing its shape
D. Moving it to GPU

Answer: B. Telling autograd to track operations for gradient computation
Set requires_grad=True for parameters you want to optimise.

56. Which method computes gradients in PyTorch?

A. .forward()
B. .backward()
C. .step()
D. .zero_grad()

Answer: B. .backward()
loss.backward() populates .grad on tensors.

57. What does optimizer.step() do?

A. Computes the loss
B. Updates parameters using the computed gradients
C. Resets gradients
D. Loads data

Answer: B. Updates parameters using the computed gradients
Called after loss.backward().

58. Why call optimizer.zero_grad() each iteration?

A. To delete the model
B. Because gradients accumulate by default
C. To increase speed
D. To shuffle data

Answer: B. Because gradients accumulate by default
Without it, gradients from previous batches add up.

59. Which PyTorch class do you subclass to define a custom model?

A. torch.Tensor
B. torch.nn.Module
C. torch.optim.SGD
D. torch.utils.Data

Answer: B. torch.nn.Module
You define layers in __init__ and the computation in forward().

60. What do Dataset and DataLoader do in PyTorch?

A. Define the model
B. Dataset wraps the data; DataLoader batches and shuffles it
C. Compute gradients
D. Plot results

Answer: B. Dataset wraps the data; DataLoader batches and shuffles it
DataLoader(dataset, batch_size=..., shuffle=True).

61. Which loss function is used for multi-class classification in PyTorch?

A. nn.MSELoss
B. nn.CrossEntropyLoss
C. nn.L1Loss
D. nn.BCELoss

Answer: B. nn.CrossEntropyLoss
It combines LogSoftmax and NLLLoss.

62. Which loss is used for linear regression in PyTorch?

A. nn.CrossEntropyLoss
B. nn.MSELoss
C. nn.NLLLoss
D. nn.BCELoss

Answer: B. nn.MSELoss
Mean squared error.

63. What is mini-batch gradient descent?

A. Using the whole dataset per update
B. Updating weights using small subsets of the data
C. Using one sample per update
D. No updates

Answer: B. Updating weights using small subsets of the data
A balance between batch and stochastic gradient descent.

64. What does torch.nn.Linear(in_features, out_features) create?

A. A convolution layer
B. A fully connected (linear) layer
C. An activation
D. A loss

Answer: B. A fully connected (linear) layer
It applies y = xA^T + b.

65. What is the purpose of a validation set?

A. Final testing
B. Tuning hyperparameters and monitoring overfitting during training
C. Training
D. Deployment

Answer: B. Tuning hyperparameters and monitoring overfitting during training
The test set is kept for the final unbiased estimate.

66. How do you move a PyTorch model to a GPU?

A. model.gpu()
B. model.to('cuda')
C. model.fast()
D. model.device()

Answer: B. model.to('cuda')
Tensors must be moved to the same device.

67. What is 'momentum' in optimisation?

A. A type of loss
B. A term that accelerates gradient descent by accumulating past gradients
C. A regulariser
D. A data loader

Answer: B. A term that accelerates gradient descent by accumulating past gradients
It helps escape shallow minima and smooths updates.

68. What is the softmax regression model used for?

A. Regression
B. Multi-class classification
C. Clustering
D. Generation

Answer: B. Multi-class classification
It generalises logistic regression to many classes.

69. In a CNN in PyTorch, nn.Conv2d's 'kernel_size' sets:

A. The number of classes
B. The size of the filter window
C. The learning rate
D. The batch size

Answer: B. The size of the filter window
Stride and padding control the output size.

70. What does model.eval() do?

A. Trains the model
B. Sets layers like dropout and batch norm to evaluation mode
C. Deletes gradients
D. Saves the model

Answer: B. Sets layers like dropout and batch norm to evaluation mode
Use torch.no_grad() as well during inference.

Courses 7-11: Generative AI, LLMs and Fine-Tuning Transformers

71. What is a large language model (LLM)?

A. A database
B. A transformer-based model trained on huge text corpora to understand and generate language
C. A search index
D. A tokenizer

Answer: B. A transformer-based model trained on huge text corpora to understand and generate language
Examples: GPT, Llama, IBM Granite.

72. What is tokenization?

A. Encrypting text
B. Splitting text into tokens (words or sub-words) that a model can process
C. Translating text
D. Summarising text

Answer: B. Splitting text into tokens (words or sub-words) that a model can process
Methods include word-based, character-based and sub-word (BPE, WordPiece).

73. What is an embedding?

A. A compressed file
B. A dense vector that represents the meaning of a token or text
C. A type of loss
D. A layer of dropout

Answer: B. A dense vector that represents the meaning of a token or text
Similar meanings have nearby vectors.

74. Which architecture uses only a decoder and is used for text generation?

A. BERT
B. GPT
C. Word2Vec
D. CNN

Answer: B. GPT
GPT models are autoregressive decoders.

75. Which architecture uses only an encoder and suits understanding tasks?

A. GPT
B. BERT
C. GAN
D. RNN

Answer: B. BERT
BERT is trained with masked language modelling.

76. What is positional encoding in a transformer?

A. A loss function
B. Information added to embeddings so the model knows token order
C. A tokeniser
D. An optimiser

Answer: B. Information added to embeddings so the model knows token order
Self-attention alone is order-agnostic.

77. What is self-attention?

A. Paying attention to the user
B. A mechanism that weighs the relevance of every token to every other token in a sequence
C. A pooling method
D. A regulariser

Answer: B. A mechanism that weighs the relevance of every token to every other token in a sequence
It captures long-range dependencies.

78. What is fine-tuning an LLM?

A. Training from scratch
B. Further training a pre-trained model on task-specific data
C. Shrinking a model
D. Tokenising data

Answer: B. Further training a pre-trained model on task-specific data
Adapts general knowledge to a domain.

79. What is PEFT?

A. A dataset
B. Parameter-Efficient Fine-Tuning: updating a small number of parameters instead of the whole model
C. A tokenizer
D. A metric

Answer: B. Parameter-Efficient Fine-Tuning: updating a small number of parameters instead of the whole model
It cuts compute and storage costs.

80. What is LoRA?

A. A loss function
B. Low-Rank Adaptation: adding small trainable low-rank matrices to frozen weights
C. A dataset format
D. A prompt style

Answer: B. Low-Rank Adaptation: adding small trainable low-rank matrices to frozen weights
A popular PEFT method.

81. What is QLoRA?

A. LoRA with quantised (e.g. 4-bit) base model weights to reduce memory
B. A query language
C. A tokenizer
D. A new architecture

Answer: A. LoRA with quantised (e.g. 4-bit) base model weights to reduce memory
It enables fine-tuning large models on modest hardware.

82. What is quantization?

A. Adding layers
B. Reducing the numerical precision of model weights to shrink size and speed inference
C. Tokenising
D. Clustering

Answer: B. Reducing the numerical precision of model weights to shrink size and speed inference
For example from 32-bit floats to 8-bit or 4-bit.

83. What is instruction tuning?

A. Tuning hyperparameters
B. Fine-tuning a model on instruction-response pairs so it follows prompts better
C. Pruning
D. Quantising

Answer: B. Fine-tuning a model on instruction-response pairs so it follows prompts better
It improves zero-shot task following.

84. What is RLHF?

A. A tokenizer
B. Reinforcement Learning from Human Feedback
C. A dataset
D. A pooling layer

Answer: B. Reinforcement Learning from Human Feedback
Human preference signals train a reward model that guides the LLM.

85. Which library is widely used to load pre-trained transformers?

A. scikit-learn
B. Hugging Face Transformers
C. OpenCV
D. Matplotlib

Answer: B. Hugging Face Transformers
It provides models, tokenizers and pipelines.

86. What is in-context learning?

A. Retraining the model
B. The model adapting to a task from examples given in the prompt without weight updates
C. Fine-tuning
D. Quantising

Answer: B. The model adapting to a task from examples given in the prompt without weight updates
Few-shot prompting is in-context learning.

87. What is perplexity?

A. A tokeniser
B. A measure of how well a language model predicts a sample; lower is better
C. A layer
D. An optimiser

Answer: B. A measure of how well a language model predicts a sample; lower is better
It is the exponential of the average negative log-likelihood.

88. What does 'temperature' control during generation?

A. Hardware heat
B. The randomness of the sampled output
C. Model size
D. Batch size

Answer: B. The randomness of the sampled output
Lower values make output more deterministic.

Courses 12-13: RAG, LangChain and AI Agents

89. What does RAG stand for?

A. Random Answer Generator
B. Retrieval-Augmented Generation
C. Recurrent Attention Graph
D. Rapid AI Gateway

Answer: B. Retrieval-Augmented Generation
RAG retrieves relevant documents and passes them to the LLM as context.

90. Why use RAG?

A. To make models smaller
B. To ground answers in up-to-date or private data and reduce hallucination
C. To speed up training
D. To replace tokenisation

Answer: B. To ground answers in up-to-date or private data and reduce hallucination
The model does not need retraining when the knowledge changes.

91. What is a vector database used for in RAG?

A. Storing images
B. Storing embeddings and retrieving the most similar chunks to a query
C. Training models
D. Tokenising text

Answer: B. Storing embeddings and retrieving the most similar chunks to a query
Examples: Chroma, FAISS, Milvus.

92. What is 'chunking' in a RAG pipeline?

A. Deleting documents
B. Splitting documents into smaller pieces before embedding
C. Merging answers
D. Compressing models

Answer: B. Splitting documents into smaller pieces before embedding
Chunk size and overlap affect retrieval quality.

93. What is LangChain?

A. A blockchain
B. A framework for building LLM applications by chaining prompts, models, retrievers and tools
C. A tokenizer
D. A GPU

Answer: B. A framework for building LLM applications by chaining prompts, models, retrievers and tools
It provides chains, agents, memory and document loaders.

94. In LangChain, what is a 'chain'?

A. A dataset
B. A sequence of calls (prompt, model, parser) linked together
C. A GPU cluster
D. A loss function

Answer: B. A sequence of calls (prompt, model, parser) linked together
LCEL uses the pipe operator to compose them.

95. What is an AI agent?

A. A human operator
B. An LLM-driven system that decides which tools or actions to use to reach a goal
C. A dataset
D. A tokenizer

Answer: B. An LLM-driven system that decides which tools or actions to use to reach a goal
Agents plan, call tools and observe results in a loop.

96. What is a 'prompt template'?

A. A fixed answer
B. A reusable prompt with placeholders filled at run time
C. A model checkpoint
D. A tokenizer file

Answer: B. A reusable prompt with placeholders filled at run time
It keeps prompts consistent across calls.

97. What does a 'retriever' do in RAG?

A. Generates text
B. Finds the most relevant documents for a query
C. Trains the model
D. Evaluates accuracy

Answer: B. Finds the most relevant documents for a query
Usually via similarity search over embeddings.

98. Which similarity measure is common for comparing embeddings?

A. Hamming distance
B. Cosine similarity
C. Jaccard index
D. Edit distance

Answer: B. Cosine similarity
It compares the angle between vectors.

99. What is 'memory' in a LangChain application?

A. GPU RAM
B. Stored conversation history that gives the model context across turns
C. Disk space
D. The training set

Answer: B. Stored conversation history that gives the model context across turns
Buffer and summary memory are common types.

100. What is Gradio used for in the capstone projects?

A. Training models
B. Building a simple web interface for a model or app
C. Tokenising
D. Quantising

Answer: B. Building a simple web interface for a model or app
It lets you demo an app in a few lines of Python.

101. What is a hallucination in a RAG system usually a sign of?

A. Too many GPUs
B. Poor retrieval or the model ignoring the provided context
C. Low temperature
D. Too many chunks

Answer: B. Poor retrieval or the model ignoring the provided context
Better chunking, retrieval and prompts reduce it.

102. What is the role of an 'output parser' in LangChain?

A. Tokenising input
B. Converting the model's raw text into a structured format such as JSON
C. Embedding documents
D. Training

Answer: B. Converting the model's raw text into a structured format such as JSON
It makes outputs usable by downstream code.