AI ML Interview Questions

Here are **150 AI/ML interview-style questions** categorized by topic. These cover theory, coding, applications, and frameworks like TensorFlow, PyTorch, etc.

🧠 Core Machine Learning (30)

1. What is the difference between supervised, unsupervised, and reinforcement learning?
2. Define bias and variance. What is the bias-variance tradeoff?
3. What is overfitting and how can it be prevented?
4. What is cross-validation?
5. Explain the difference between L1 and L2 regularization.
6. What are precision, recall, F1-score, and accuracy?
7. What is the confusion matrix?
8. What is ROC-AUC and how do you interpret it?
9. What is feature engineering?
10. What is dimensionality reduction? Give examples.
11. What is Principal Component Analysis (PCA)?
12. What is the curse of dimensionality?
13. What is the difference between bagging and boosting?
14. Explain Gradient Boosting vs AdaBoost.
15. What are decision trees? How is a split made?
16. What is entropy in decision trees?
17. What is random forest and how does it work?
18. How does KNN work? What is its time complexity?
19. What is SVM? What is the role of the kernel?
20. What are hyperparameters vs model parameters?
21. What is a learning rate?

22. What is grid search vs random search?
23. What is early stopping?
24. How do you handle class imbalance?
25. What is SMOTE?
26. What is model drift?
27. What is a confusion matrix for multiclass classification?
28. What is cross-entropy loss?
29. What is log-loss?
30. What is the elbow method in clustering?

📊 Statistical Concepts (15)

31. What is the Central Limit Theorem?
32. Explain Bayes' Theorem.
33. What is a p-value?
34. What is hypothesis testing?
35. What are Type I and Type II errors?
36. What is a confidence interval?
37. What is the difference between correlation and causation?
38. What is multicollinearity?
39. What is autocorrelation?
40. What is heteroscedasticity?
41. What is maximum likelihood estimation (MLE)?
42. What is prior, likelihood, and posterior in Bayesian inference?
43. What is the law of large numbers?
44. What are skewness and kurtosis?
45. What is a z-score?

📚 Deep Learning (30)

46. What is the difference between a perceptron and a neuron?
47. What is the architecture of a neural network?
48. What is backpropagation?
49. What is gradient descent? Explain variants like SGD, Adam.
50. What is the vanishing gradient problem?
51. What are activation functions? Why is ReLU popular?
52. What is a loss function? How is it used?
53. What is dropout and why is it used?
54. What is batch normalization?
55. What is an epoch, batch size, and iteration?
56. What is a convolutional neural network (CNN)?
57. What are pooling layers in CNNs?
58. What is transfer learning?
59. What is fine-tuning in deep learning?
60. What is a recurrent neural network (RNN)?
61. What is the difference between RNN, LSTM, and GRU?
62. What is a transformer model?
63. What are attention mechanisms?
64. What is positional encoding?
65. What is multi-head attention?
66. What is self-supervised learning?
67. What is contrastive learning?
68. What is reinforcement learning? Define reward and policy.
69. What is Q-learning?
70. What are policy gradient methods?
71. What is an autoencoder?

72. What is a GAN? How does it work?
73. What is mode collapse in GANs?
74. What is the role of the discriminator in GANs?
75. How are embeddings generated and used?

🛠 Tools and Frameworks (TensorFlow, PyTorch, etc.) (25)

76. What are the main differences between TensorFlow and PyTorch?
77. How do you define a model in PyTorch?
78. How does `autograd` work in PyTorch?
79. What is the role of `DataLoader` and `Dataset` in PyTorch?
80. What is a Tensor in PyTorch/TensorFlow?
81. How do you use GPU with PyTorch?
82. What is Keras? How is it related to TensorFlow?
83. What are callbacks in Keras?
84. How do you save and load models in TensorFlow/Keras?
85. How to visualize a model architecture in TensorFlow?
86. What is the difference between static and dynamic computation graphs?
87. What is eager execution?
88. How do you freeze layers in transfer learning?
89. How do you implement custom loss in TensorFlow?
90. How do you monitor training performance?
91. How do you implement learning rate schedulers?
92. What is model checkpointing?
93. What is TensorBoard used for?
94. What are `nn.Module` and `nn.Sequential` in PyTorch?
95. How to avoid memory leak in training loop?

96. How to convert a model to ONNX?
97. What is `torchscript`?
98. What is `tf.function` used for?
99. How do you perform gradient clipping?
100. How do you perform inference in PyTorch?

📦 NLP and Computer Vision (20)

101. What is tokenization?
102. What is word embedding? Name popular types.
103. What is the difference between Word2Vec, GloVe, and BERT?
104. What is stemming vs lemmatization?
105. What is N-gram? How is it used?
106. What is BLEU score?
107. What is perplexity?
108. What are language models? Explain GPT.
109. What is BERT and how does it work?
110. What is masked language modeling?
111. What is image augmentation? Why is it useful?
112. What is transfer learning in computer vision?
113. What is object detection? How is YOLO different from RCNN?
114. What is semantic segmentation?
115. What is the difference between classification and localization?
116. What is facial recognition pipeline?
117. What is OCR?
118. What is caption generation?
119. How do ViT (Vision Transformers) work?

120. What is CLIP?

⚙️ Big Data & Deployment (20)

121. What is the role of Spark in ML?
122. What is MLlib in Spark?
123. How does distributed training work?
124. What is Horovod?
125. What is model serving?
126. What are options for model deployment?
127. What is a REST API for model inference?
128. What is Docker and how is it useful in ML?
129. What is the role of Kubernetes in deployment?
130. What is edge computing in ML?
131. How to use Apache Kafka in an ML pipeline?
132. What is MLFlow?
133. What is the role of Airflow?
134. What is feature store?
135. What is A/B testing in ML?
136. What is online vs offline inference?
137. What is streaming data vs batch data?
138. How to monitor a deployed model?
139. What is model retraining?
140. What is concept drift?

🧪 Applied/Scenario Questions (10)

141. How would you handle missing data in a dataset?
142. How do you select important features?
143. You have 10K samples and 1M features. What would you do?
144. How do you explain a black-box model?
145. What would you do if your model performs well on train but poorly on test?
146. How do you build a recommendation system?
147. What pipeline would you design for real-time fraud detection?
148. How do you approach hyperparameter tuning in a budget-constrained environment?
149. How do you combine multiple models (ensemble)?
150. How would you handle a model degrading in production?

Let me know if you'd like answers, or a subset with detailed explanations or code.

Here are **20 interview-style questions** focused on deploying and optimizing machine learning solutions on **massive datasets using big data technologies** like Spark, Hadoop, Kafka, and distributed systems:

🔄 Data Handling & Preprocessing at Scale

1. **How would you preprocess terabytes of data for a machine learning model?**
2. **What are the advantages of using Apache Spark over Pandas for large-scale data?**
3. **Explain how Spark DataFrames differ from RDDs and when to use each.**
4. **How can you join multiple massive datasets efficiently in Spark?**
5. **How do you handle skewed joins in distributed datasets?**

⚙️ Model Training on Big Data

6. **What is distributed training? How is it different from traditional training?**
7. **How does MLlib in Apache Spark handle machine learning training at scale?**
8. **What is the role of parameter servers in distributed deep learning?**
9. **How can you use mini-batch gradient descent in a distributed setting?**
10. **What are the challenges of model convergence when using big data platforms?**

🚀 Model Deployment in Big Data Environments

11. **How would you deploy a trained model to serve predictions on streaming data?**
12. **What is the role of Apache Kafka in real-time ML pipelines?**
13. **How do you deploy ML models using Docker and Kubernetes in production?**
14. **What are the trade-offs between batch inference and streaming inference?**
15. **What tools would you use to monitor model performance post-deployment on big data?**

🔧 Optimization, Scaling & Performance

16. **How do you optimize Spark jobs for ML pipelines?**
17. **What are broadcast variables in Spark, and how do they help with optimization?**
18. **How can you reduce data shuffling in distributed ML workflows?**
19. **What is model parallelism vs data parallelism? Give examples.**
20. **How do you handle fault tolerance and job recovery in a distributed ML environment?**

Let me know if you want sample answers, or if you'd like to focus on one stack like Spark or TensorFlow on distributed clusters.

Here are **20 interview-style questions** focused on the theme:
"Develop and prototype AI algorithms and software tools" — covering algorithm design, prototyping, evaluation, and tooling:

🧠 AI Algorithm Design

1. **How do you decide which AI algorithm is appropriate for a given problem?**
2. **What is the difference between heuristic-based AI and learning-based AI?**
3. **How would you design a custom loss function for a specialized problem?**
4. **What considerations go into selecting an optimization algorithm (e.g., Adam vs SGD)?**
5. **How do you balance accuracy with computational cost during algorithm development?**

🔁 Prototyping and Experimentation

6. **How do you structure an experimental pipeline for rapid prototyping of AI models?**
7. **What tools or frameworks do you use to quickly prototype deep learning ideas?**
8. **What is transfer learning and when would you use it to prototype a solution?**
9. **How do you validate the robustness of an AI prototype before production?**
10. **What role do automated tools (like AutoML) play in your prototyping workflow?**

🛠 Software Tools and Engineering

11. **How do you build reusable components for AI model training and evaluation?**
12. **What software engineering practices do you follow in AI projects (e.g., versioning, testing)?**
13. **How would you implement logging and debugging in an AI pipeline?**

14. **How do you use configuration management tools (e.g., Hydra, YAML, argparse) in prototyping?**
15. **How do you ensure reproducibility in your AI experiments?**

📊 Evaluation, Feedback, and Iteration

16. **How do you choose evaluation metrics aligned with business or user goals?**
17. **What strategies do you use to interpret black-box AI models?**
18. **How do you conduct error analysis to guide algorithm improvements?**
19. **What tools do you use to track and compare experiments? (e.g., MLflow, Weights & Biases)**
20. **How do you handle feedback loops in AI systems (e.g., human-in-the-loop)?**

Let me know if you'd like model answers, a breakdown by topic (e.g., NLP, CV, RL), or a Python code example for a prototyping workflow.

Here are **20 interview-style questions** focused on the theme:
"Implement enhancements to existing algorithmic / deep learning solutions" — covering optimization, debugging, architectural tweaks, and performance improvement.

🔍 Diagnosis & Bottleneck Identification

1. **How do you identify performance bottlenecks in an existing deep learning model?**
2. **What methods do you use to detect overfitting or underfitting in a model?**
3. **How do you interpret a learning curve to decide where improvements are needed?**
4. **How would you debug a model that performs well on training but poorly on test data?**
5. **What are common pitfalls when modifying existing models that initially worked well?**

⚙️ Architectural & Algorithmic Improvements

6. How would you improve the accuracy of a CNN without significantly increasing training time?
7. How can you make an existing LSTM model faster and more efficient?
8. What changes would you make to improve a model's generalization on unseen data?
9. When should you switch from a baseline algorithm to a more complex model like transformers?
10. How would you optimize a recommendation algorithm for scalability and precision?

🚀 Optimization & Efficiency

11. How do you reduce the memory footprint of a deep learning model?
12. What is quantization and how can it improve an existing deep learning solution?
13. How can you use mixed precision training to speed up deep models?
14. How do you decide which layers or hyperparameters to tune first for performance gain?
15. What’s the role of batch normalization or dropout in enhancing a model?

🔁 Training & Data Enhancements

16. How would you use data augmentation to improve a computer vision model?
17. What strategies do you use for improving models with imbalanced datasets?
18. How can transfer learning be applied to enhance a low-performing model?
19. How would you leverage ensembling to improve a model’s predictions?
20. What role does feature engineering still play in enhancing deep learning pipelines?

Want sample answers, code snippets for a few of these, or task-specific variants (e.g., NLP vs CV)?

Here are **20 interview-style questions** focused on the theme:
"Conduct quantitative data analysis using a variety of datasets" — covering exploratory data analysis, statistical techniques, hypothesis testing, and tools.

📊 Exploratory Data Analysis (EDA)

1. **What steps do you follow when starting quantitative analysis on a new dataset?**
2. **How do you handle missing or corrupted data in quantitative analysis?**
3. **What are the key differences between univariate, bivariate, and multivariate analysis?**
4. **How would you detect and treat outliers in a dataset?**
5. **What visualization techniques do you use to understand distributions and relationships?**

📐 Statistical Methods & Inference

6. **How do you choose between mean, median, and mode as a measure of central tendency?**
7. **What is p-value and how do you interpret it in hypothesis testing?**
8. **When would you use a t-test vs ANOVA?**
9. **What is the Central Limit Theorem and why is it important?**
10. **Explain Type I and Type II errors with an example.**

🧮 Modeling & Regression

11. **How do you decide which regression model to use (e.g., linear, logistic, Poisson)?**
12. **What is multicollinearity and how can it affect your regression analysis?**
13. **How do you interpret coefficients in a linear regression model?**
14. **What metrics do you use to evaluate a regression model's performance?**
15. **How would you perform feature selection in a dataset with 100+ variables?**

⚙️ Tools, Techniques & Real-World Scenarios

16. **Which libraries/tools do you use for quantitative analysis in Python or R?**
17. **Describe a time when quantitative analysis helped drive a business decision.**
18. **How do you handle datasets coming from different sources/formats (e.g., SQL, CSV, APIs)?**
19. **How would you compare two forecasting models on time series data?**
20. **What are your strategies for ensuring your data analysis is reproducible and interpretable?**

Let me know if you'd like Python code for any of these, or want to tailor the questions to a domain like finance, healthcare, or retail.

Here are **50 interview-style questions** on **NLP**, **linguistic data extraction**, and **modelling techniques** — with a **strong focus on LLMs** (encoder-based, decoder-based, encoder-decoder hybrids) and **Generative AI** applications:

📘 General NLP Concepts (10)

1. What are the main steps in an NLP pipeline?
2. How is tokenization handled in modern LLMs like BERT and GPT?

3. What is the difference between stemming and lemmatization?
4. How does POS tagging work, and why is it important?
5. What are word embeddings, and how are they different from one-hot encodings?
6. Explain the concept of Named Entity Recognition (NER).
7. How would you evaluate the performance of a sentiment analysis model?
8. What is dependency parsing and how is it used in NLP tasks?
9. How can topic modeling be applied to large document corpora?
10. Compare TF-IDF with embedding-based text representations.

🏗️ Language Model Architectures (Encoders, Decoders, Hybrids) (10)

11. What is the difference between encoder-only (BERT) and decoder-only (GPT) models?
12. Why is BERT not suitable for text generation tasks?
13. How do encoder-decoder models like T5 or BART work?
14. What is masked language modeling (MLM) and how does it differ from causal language modeling (CLM)?
15. How do attention mechanisms work in transformers?
16. What is self-attention and how does it differ from cross-attention?
17. Why do decoder-only models use causal masking?
18. How does positional encoding work in transformer models?
19. Explain the role of layer normalization and residual connections in LLMs.
20. How is the encoder-decoder attention different from self-attention?

🤖 Large Language Models (LLMs) & Fine-Tuning (10)

21. What is prompt tuning and how does it differ from full fine-tuning?
22. What is LoRA (Low-Rank Adaptation), and why is it used in LLMs?

23. How does instruction tuning improve LLM generalization?
24. What are retrieval-augmented generation (RAG) models?
25. How do you fine-tune a pre-trained LLM for domain-specific tasks?
26. What is the difference between supervised fine-tuning and reinforcement learning from human feedback (RLHF)?
27. How can you use LLMs for multi-turn conversations?
28. Explain the differences between GPT-2, GPT-3, and GPT-4.
29. What challenges arise when deploying LLMs for real-time applications?
30. What is the role of temperature and top-k/top-p sampling in generation?

🧠 Linguistic Data Extraction Techniques (10)

31. How would you extract relations from unstructured text using NLP?
32. What is coreference resolution and why is it useful?
33. How can dependency trees aid in syntactic pattern extraction?
34. What are slot-filling techniques in task-oriented dialogue systems?
35. Describe how you would perform keyword extraction from legal or medical documents.
36. What is zero-shot and few-shot information extraction using LLMs?
37. How do you handle ambiguity in entity disambiguation tasks?
38. How would you use a language model for automatic question generation?
39. What are contrastive learning methods in textual representation learning?
40. How do you extract structured knowledge from scientific papers using NLP?

🧪 Evaluation, Optimization, and Deployment (10)

41. What metrics are commonly used to evaluate generative NLP tasks (e.g., BLEU, ROUGE, METEOR)?
42. How do you prevent hallucination in LLMs during generation?

43. What are the trade-offs between fine-tuning a model vs using it via API with prompt engineering?
44. How can you compress a large NLP model for deployment on edge devices?
45. How would you deploy a multilingual NLP model for global use cases?
46. How do you perform model monitoring and error analysis in NLP systems?
47. How do you ensure fairness and reduce bias in LLM outputs?
48. How do you optimize inference speed for LLMs at scale?
49. What role does caching play in the deployment of generative AI systems?
50. How do you manage token limits and long-context modeling in LLM applications?

Would you like a breakdown into beginner/intermediate/advanced levels or answers/cheat sheet for any of these?

Here are **30 interview-style questions** focused on **analytical techniques** — specifically **outlier detection, clustering, classification, regression, and hypothesis testing** — often used in data science and machine learning.

🔍 Outlier Detection (6)

1. **What are the common techniques for detecting outliers in a dataset?**
2. **How would you detect outliers in high-dimensional data?**
3. **What is the role of Z-score and IQR in identifying outliers?**
4. **When is removing outliers appropriate, and when is it not?**
5. **How does DBSCAN help in detecting noise or outliers?**
6. **How can isolation forests be used for anomaly detection?**

🔗 Clustering (6)

7. What is the difference between K-means and hierarchical clustering?
8. How do you choose the optimal number of clusters in K-means?
9. Explain the concept of silhouette score and its use in clustering.
10. What are the advantages of DBSCAN over K-means?
11. How does t-SNE or PCA help in visualizing clustering results?
12. How would you handle categorical variables in clustering algorithms?

✅ Classification (6)

13. What’s the difference between precision, recall, and F1 score?
14. How does logistic regression differ from linear regression?
15. What is the role of a confusion matrix in classification tasks?
16. How do you handle imbalanced datasets in classification problems?
17. What’s the intuition behind decision boundaries in SVM?
18. How would you use cross-validation in a classification task?

📉 Regression (6)

19. What assumptions are made in linear regression?
20. How do you evaluate the performance of a regression model?
21. What is multicollinearity and how do you detect it?
22. When would you use Ridge or Lasso regression?
23. Explain the concept of residuals in regression analysis.
24. What are some non-linear regression techniques you've worked with?

📊 Hypothesis Testing (6)

25. **What are the null and alternative hypotheses in statistical testing?**
26. **How do you decide whether to use a t-test or z-test?**
27. **What is p-value and how is it interpreted?**
28. **Explain the difference between Type I and Type II errors.**
29. **What does statistical significance mean in A/B testing?**
30. **How would you test if two population means are significantly different?**

Let me know if you want **answers, Python code**, or **domain-specific variants** (e.g., for marketing, finance, healthcare).

Here are **20 interview-style questions** focused on the **core concepts of Machine Learning** and the **mathematics behind them**, including key ideas from linear algebra, probability, calculus, optimization, and learning theory.

🧠 Core ML Concepts (10)

1. **What is the difference between supervised, unsupervised, and reinforcement learning?**
2. **How does the bias-variance tradeoff affect model performance?**
3. **What are overfitting and underfitting, and how do you address them?**
4. **What is the difference between parametric and non-parametric models?**
5. **How do gradient descent and stochastic gradient descent differ?**
6. **Explain the concept of cross-validation. Why is it important?**
7. **What are hyperparameters and how do you tune them?**
8. **What is regularization, and how do L1 and L2 differ?**
9. **What is the ROC curve and AUC, and how are they used to evaluate models?**
10. **Explain how decision trees split nodes based on impurity measures.**

📐 Mathematics Behind ML (10)

11. **What is the role of eigenvalues and eigenvectors in PCA?**
12. **Explain how the dot product is used in linear models like logistic regression.**
13. **How is Bayes' Theorem used in machine learning?**
14. **Why is the sigmoid function used in binary classification problems?**
15. **What is the cost function in linear regression, and how is it minimized?**
16. **Explain the concept of convexity in optimization problems.**
17. **What is the Jacobian matrix, and where does it apply in ML models?**
18. **How do you interpret the Hessian matrix in second-order optimization methods?**
19. **Describe the role of probability distributions in generative models.**
20. **How is entropy used in information gain for decision trees?**

Let me know if you'd like the **answers**, **code implementations**, or **visual aids** for these concepts!

Here are **10 interview-style questions** on **Machine Learning tools and technologies** — covering libraries, platforms, frameworks, and their applications:

🧰 ML Tools & Technologies (10 Questions)

1. **What are the differences between TensorFlow and PyTorch in terms of usability and deployment?**
2. **When would you use Scikit-learn over deep learning frameworks like Keras or PyTorch?**
3. **What are some advantages of using MLflow in the machine learning lifecycle?**
4. **How does Apache Spark's MLlib support large-scale machine learning?**

5. **What is the role of ONNX in ML model deployment?**
6. **How do tools like Weights & Biases or Neptune help in experiment tracking?**
7. **What is the use of Hugging Face Transformers in modern NLP applications?**
8. **How can Docker and Kubernetes help in ML model deployment and scaling?**
9. **What is the difference between SageMaker, Vertex AI, and Azure ML?**
10. **How do you choose between CPU, GPU, and TPU for training ML models?**

Let me know if you want follow-up questions, detailed answers, or comparisons between specific tools!

Here are **20 interview-style questions** covering the key foundations of **statistical techniques**, **linear algebra**, and **numerical optimization** as used in **machine learning**:

📊 Statistical Techniques (7 Questions)

1. **What is the difference between probability and likelihood?**
2. **Explain how Maximum Likelihood Estimation (MLE) is used to fit statistical models.**
3. **What is the Central Limit Theorem, and why is it important in ML?**
4. **How do you interpret confidence intervals and p-values in model evaluation?**
5. **What is the difference between parametric and non-parametric models?**
6. **Describe a real-world scenario where Bayesian methods outperform frequentist approaches.**
7. **What is the curse of dimensionality in the context of statistical learning?**

📐 Linear Algebra (7 Questions)

8. **How are matrix operations used in training neural networks?**

9. **What is the role of eigenvectors and eigenvalues in PCA?**
10. **How is the dot product used in computing similarity between vectors?**
11. **Why is the rank of a matrix important in linear models?**
12. **How do SVD and QR decompositions relate to dimensionality reduction?**
13. **What is the geometric interpretation of a linear transformation?**
14. **How does the Gram-Schmidt process help in orthogonalizing vectors?**

🔽 Numerical Optimization (6 Questions)

15. **What is gradient descent, and how does it find local minima?**
16. **How do learning rate and momentum affect convergence in gradient-based methods?**
17. **What are the advantages of using second-order methods like Newton's method?**
18. **Explain how the Adam optimizer improves upon SGD.**
19. **What are saddle points, and why are they problematic in optimization?**
20. **How do constraints change the formulation of optimization problems in ML (e.g., Lagrange multipliers)?**

Let me know if you want **answers, derivations, or examples with code** for any of these!