1. Define data in the context of AI systems.

Data refers to raw facts, figures, or information (text, images, audio, video, sensor readings, etc.) that are collected and used by AI systems to learn patterns, make predictions, and improve performance.

2. Define data visualization in AI.

Data visualization in AI is the process of representing data and analysis results graphically (charts, plots, graphs, dashboards) to better understand patterns, relationships, and insights.

3. Name two common types of plots used in AI data analysis.

- Scatter plot

- Histogram

4. Name two roles of data in training AI models.

- Training: Data is used to teach the model patterns and relationships.

- Validation/Testing: Data is used to check the accuracy and generalization of the model.

5. What is Natural Language Processing (NLP)?

NLP is a field of AI that focuses on enabling machines to understand, interpret, and generate human language.

6. Give two applications of NLP.

- Machine translation (e.g., Google Translate)

- Chatbots and virtual assistants (e.g., Siri, Alexa)

7. Explain tokenization and stemming in NLP.

- Tokenization: Splitting text into smaller units like words, phrases, or sentences.

- Stemming: Reducing words to their base or root form (e.g., “playing” → “play”).

8. Describe the role of embeddings in representing text data.

Embeddings convert words or sentences into numerical vectors that capture their meaning and relationships, allowing AI models to process and compare text efficiently.

9. Explain why high-quality data is essential for AI performance.

High-quality data ensures accuracy, reduces bias, improves generalization, and helps models make reliable predictions. Poor-quality data can lead to errors and misleading results.

10. Describe the difference between structured and unstructured data.

- Structured data: Organized in rows and columns (e.g., databases, spreadsheets).

- Unstructured data: Not organized in a predefined format (e.g., text, images, videos).

11. What is data acquisition?

Data acquisition is the process of collecting data from various sources (sensors, databases, web, surveys, etc.) to be used in AI systems.

12. Define data pre-processing in AI.

Data pre-processing is the step of cleaning, transforming, and organizing raw data into a suitable format for model training (e.g., handling missing values, normalization).

13. What is an outlier?

An outlier is a data point that is significantly different from other observations, often due to variability or errors, and may distort analysis if not handled properly.

14. What is a Transformer model?

A Transformer model is a deep learning architecture based on the attention mechanism, designed to process sequential data (like text) in parallel. It captures long-range dependencies and contextual relationships more effectively than traditional models like RNNs or LSTMs.

15. Name one popular Transformer model used in NLP.

- BERT (Bidirectional Encoder Representations from Transformers)

16. Apply normalization or standardization to a numerical dataset.

- Normalization (Min-Max scaling): x' = (x - x_min) / (x_max - x_min)

Example: For dataset [10, 20, 30], normalized values → [0, 0.5, 1]

- Standardization (Z-score scaling): x' = (x - μ) / σ

Example: For dataset [10, 20, 30], mean = 20, std ≈ 8.16 → standardized values ≈ [-1.22, 0, 1.22]

17. Explain why labelled data is essential for supervised learning.

Labelled data provides input-output pairs that allow the model to learn the mapping between features and target values. Without labels, the model cannot understand what outcome to predict or measure accuracy against.

18. Apply preprocessing steps (tokenization, stopword removal) to a sample text dataset.

Sample text: “AI is transforming the world with intelligence.”

- Tokenization → ["AI", "is", "transforming", "the", "world", "with", "intelligence"]

- Stopword removal → ["AI", "transforming", "world", "intelligence"]

19. Explain methods to handle missing data in a dataset.

- Deletion: Remove rows or columns with missing values (if few).

- Imputation: Replace missing values with mean, median, mode, or predicted values.

- Use algorithms that handle missing values: e.g., decision trees.

20. What is the main difference between overfitting and underfitting?

- Overfitting: Model learns training data too well, including noise, and fails to generalize.

- Underfitting: Model is too simple and fails to learn important patterns in data.

21. Give one example of a situation that can cause overfitting.

Training a deep neural network on a very small dataset without regularization.

22. Give one example of a situation that can cause underfitting.

Using a linear regression model to predict a complex, non-linear relationship (e.g., predicting housing prices based on multiple non-linear features).

23. Explain the difference between batch and real-time data acquisition.

- Batch data acquisition: Data is collected, stored, and processed at intervals (e.g., daily sales reports).

- Real-time data acquisition: Data is collected and processed continuously as events occur (e.g., stock market updates).

24. Analyze the challenges commonly faced during data acquisition in real-world applications.

- Data inconsistency (different formats, missing fields).

- Noise and errors in raw data.

- Privacy and security concerns in data collection.

- High cost of acquiring large-scale quality data.

- Latency in real-time data collection.

25. Explain the common challenges encountered in data acquisition for real-world applications.

- Heterogeneous data sources.

- Incomplete or missing data.

- Data duplication or redundancy.

- Scalability with large volumes.

- Compliance with data regulations (GDPR, HIPAA).

26. Explain the various types of data annotation techniques, such as image, text, audio, and video annotation.

- Image annotation: Labeling objects in images (e.g., bounding boxes for cars in self-driving datasets).

- Text annotation: Tagging entities, sentiment, or parts of speech in text.

- Audio annotation: Labeling speech segments, phonemes, or emotions in audio.

- Video annotation: Frame-by-frame labeling of objects, activities, or events.

27. Describe the architecture of Transformer models and compare their efficiency and scalability with recurrent neural networks.

- Architecture: Input embeddings + positional encoding, stacked encoder-decoder blocks, self-attention mechanism, feed-forward layers, normalization, residual connections.

- Comparison: Transformers process sequences in parallel (faster), handle long-range dependencies better, and scale efficiently, while RNNs are sequential and struggle with vanishing gradients.

28. Evaluate the advantages of using Transformers over traditional sequence models for NLP tasks like translation or summarization.

- Capture long-term dependencies without forgetting.

- Parallel training for faster computation.

- Better performance on large datasets.

- State-of-the-art results in machine translation, summarization, and text generation.

- Flexible scalability (GPT, BERT, etc.).