LLM (Large Language Model) QUESTIONS

1. What is an LLM (Large Language Model)?

An LLM is an artificial intelligence model trained on a very large amount of text data to understand and generate human-like language.
It can perform tasks such as text generation, translation, summarization, and question answering.

---

2. Define Large Language Model.

A Large Language Model is a deep learning model with a large number of parameters that learns language patterns from massive datasets.
It uses neural networks to predict the next word or token in a sequence.

---

3. What is Natural Language Processing (NLP)?

Natural Language Processing is a field of Artificial Intelligence that focuses on enabling computers to understand, interpret, and generate human language.
It allows interaction between humans and machines using natural language.

---

4. What is a Transformer?

A transformer is a deep learning architecture used in LLMs that processes text using an attention mechanism instead of recurrence.
It enables parallel processing and better understanding of long-range context.

---

5. What is Tokenization?

Tokenization is the process of breaking text into smaller units called tokens, such as words or sub-words.
It helps LLMs convert text into a format that can be processed by the model.

---

6. What is a Token in LLM?

A token is the smallest unit of text processed by a Large Language Model.
It can be a word, part of a word, punctuation, or symbol.

---

7. What is Pre-training in LLMs?

Pre-training is the initial training phase where an LLM learns general language patterns from large unlabeled text data.
The model learns grammar, context, and word relationships during this stage.

---

8. What is Fine-tuning?

Fine-tuning is the process of training a pre-trained LLM on a smaller, task-specific dataset.
It helps the model perform better for a particular application or domain.

---

9. What is Prompt Engineering?

Prompt engineering is the technique of designing effective input prompts to get accurate and useful outputs from an LLM.
It improves model responses without changing the model’s parameters.

10. What is Generative AI?

Generative AI is a type of artificial intelligence that can create new content such as text, images, audio, or code.
It learns patterns from existing data and generates outputs similar to human-created content.

---

11. What is Inference in LLM?

Inference is the process where a trained Large Language Model generates output based on a given input prompt.
It is the prediction stage after training is completed.

---

12. What is Overfitting in LLMs?

Overfitting in LLMs occurs when the model learns training data too closely and performs poorly on new or unseen data.
It reduces the model’s ability to generalize.

---

13. What is Hallucination in LLMs?

Hallucination is a situation where an LLM generates incorrect or misleading information that appears confident.
It happens when the model predicts plausible but false content.

---

14. What is Temperature in LLMs?

Temperature is a parameter that controls the randomness of the model's output.
Lower temperature gives more predictable responses, while higher temperature produces more creative outputs.

---

15. What is Zero-Shot Learning?

Zero-shot learning is the ability of an LLM to perform a task without seeing any example during training.
The model relies only on instructions given in the prompt.

---

16. What is Few-Shot Learning?

Few-shot learning is when an LLM performs a task using only a small number of examples provided in the prompt.
It helps improve accuracy with minimal training data.

---

17. What is Transfer Learning?

Transfer learning is a technique where knowledge gained from one task is reused for another related task.
In LLMs, a pre-trained model is adapted to specific tasks through fine-tuning.

---

18. What are Model Parameters?

Model parameters are internal numerical values learned during training that determine how the LLM processes data.
They store the learned language patterns and relationships.

---

19. What is Bias in LLMs?

Bias in LLMs refers to unfair or skewed outputs caused by biased training data or model design.
It may lead to discrimination or inaccurate responses.

---

20. What are Ethical Issues in LLMs?

Ethical issues in LLMs include data privacy, bias, misinformation, and misuse of generated content.
Responsible use is necessary to avoid harmful social impacts.

5 MARKS QUESTIONS

1. Explain Large Language Models (LLMs)

Definition:
Large Language Models (LLMs) are advanced deep learning models trained on massive text data to understand and generate human-like language.

Working Principle:
LLMs predict the next token in a sentence based on previous tokens using probability and learned language patterns.

Features:

- Handles large context
- Generates coherent text
- Performs multiple NLP tasks

Examples:
ChatGPT, GPT-4, BERT, Gemini

---

2. Explain the Architecture of a Transformer Model

Encoder–Decoder Structure:
The encoder converts input text into meaningful representations, while the decoder generates output text based on these representations.

Attention Mechanism:
Attention helps the model focus on important words in a sentence, improving understanding of context.

Importance in LLMs:
Transformers enable parallel processing, faster training, and better handling of long-range dependencies.

---

3. Explain the Training Process of LLMs

Data Collection:
Large volumes of text data are collected from books, articles, and websites.

Pre-training:
The model learns general language patterns using unlabeled data.

Fine-tuning:
The model is trained on task-specific or domain-specific datasets.

Inference:
The trained model generates responses for new inputs.

---

4. Explain Attention and Self-Attention Mechanism

Definition:
Attention allows the model to focus on relevant words in a sentence.

Working:
Self-attention compares each word with other words in the same sentence to understand relationships.

Advantage over RNN:
It processes data in parallel and captures long-distance dependencies better than RNNs.

---

5. Explain Tokenization and Embeddings in LLMs

Types of Tokenization:
Word-based, subword-based, and character-based tokenization.

Word vs Contextual Embeddings:
Word embeddings give fixed meanings, while contextual embeddings change meaning based on sentence context.

Role in Language Understanding:
They convert text into numerical form for better language comprehension.

---

6. Explain Prompt Engineering with Examples

Definition:
Prompt engineering is the process of designing effective input prompts to guide LLM responses.

Types of Prompts:

- Zero-shot
- Few-shot
- Instruction-based

Importance:
It improves accuracy and usefulness of model outputs.

Example:
“Summarize this text in 5 lines.”

---

7. Explain Applications of Large Language Models

Education:
Automated tutoring, content generation.

Healthcare:
Medical report analysis, symptom checking.

Chatbots:
Customer support and virtual assistants.

Code Generation:
Writing, debugging, and explaining code.

---

8. Explain Challenges and Limitations of LLMs

Hallucination:
Models may generate incorrect information confidently.

Bias:
Outputs may reflect biases present in training data.

High Computational Cost:
Training requires powerful hardware and high energy.

Data Privacy:
Risk of using sensitive or personal data.

Explain Zero-Shot, One-Shot and Few-Shot Learning

Definition

Zero-Shot Learning:
Zero-shot learning is a technique where a model performs a task without seeing any

example during training.
The model relies only on instructions or prior knowledge to complete the task.

One-Shot Learning:
One-shot learning is a learning approach where the model learns to perform a task using only one example.
It helps the model generalize from very limited data.

Few-Shot Learning:
Few-shot learning is a method where a model learns a task using a small number of examples (usually 2–10).
It improves performance compared to zero-shot and one-shot learning.

Zero-Shot Example:
Asking an LLM: *“Translate this sentence into French”* without providing any example.

One-Shot Example:
Giving one sample translation and asking the model to translate another sentence.

Few-Shot Example:
Providing 3–5 sample question–answer pairs and asking the model to answer a new question