Question on Artificial Neural Network(ANN):

1. What is an Artificial Neural Network, and how does it work?

2. What are activation functions, tell me the type of the activation functions and why are they used in neural networks?

3. What is backpropagation, and how does it work in training neural networks?

4. What is the vanishing gradient and exploding gradient problem, and how can it affect neural network training?

5. How do you prevent overfitting in neural networks?

6. What is dropout, and how does it help in training neural networks?

7. How do you choose the number of layers and neurons for a neural network?

8. What is transfer learning, and when is it useful?

9. What is a loss function, and how do you choose the appropriate one for your model?

10. Explain the concept of gradient descent and its variations like stochastic gradient descent (SGD) and mini-batch gradient descent.

11. What is the role of a learning rate in neural network training, and how do you optimize it?

12. What are some common neural network based architectures, and when would you use them?

13. What is a convolutional neural network (CNN), and how does it differ from an artificial neural network?

14. How does a recurrent neural network (RNN) work, and what are its limitations?

Questions on Classical Natural Language Processing:

1. What is tokenization? Give me a difference between lemmatization and stemming?

2. Explain the concept of Bag of Words (BoW) and its limitations.

3. How does TF-IDF work, and how is it different from simple word frequency?

4. What is word embedding, and why is it useful in NLP?

5. What are some common applications of NLP in real-world systems?

6. What is Named Entity Recognition (NER), and where is it applied?

7. How does Latent Dirichlet Allocation (LDA) work for topic modeling?

8. What are transformers in NLP, and how have they impacted the field?

9. What is transfer learning, and how is it applied in NLP?

10. How do you handle out-of-vocabulary (OOV) words in NLP models?

11. Explain the concept of attention mechanisms and their role in sequence-to-sequence tasks.

12. What is a language model, and how is it evaluated?

Questions on Transformer and Its Extended Architecture:

1. Describe the concept of learning rate scheduling and its role in optimizing the training process of generative models over time.

2. Discuss the concept of transfer learning in the context of natural language processing. How do pre-trained language models contribute to various NLP tasks?

3. Highlight the key differences between models like GPT (Generative Pre-trained Transformer) and BERT (Bidirectional Encoder Representations from Transformers)?

4. What problems of RNNs do transformer models solve?

5. How is the transformer different from RNN and LSTM?

6. How does BERT work, and what makes it different from previous NLP models?

7. Why is incorporating relative positional information crucial in transformer models? Discuss scenarios where relative position encoding is particularly beneficial.

8. What challenges arise from the fixed and limited attention span in the vanilla Transformer model? How does this limitation affect the model's ability to capture long-term dependencies?

9. Why is naively increasing context length not a straightforward solution for handling longer context in transformer models? What computational and memory challenges does it pose?

10. How does self-attention work?

11. What pre-training mechanisms are used for LLMs, explain a few

12. Why is multi-head attention needed?

13. What is RLHF, how is it used?

14. What is catastrophic forgetting in the context of LLMs

15. In a transformer-based sequence-to-sequence model, what are the primary functions of the encoder and decoder? How does information flow between them during both training and inference?

16. Why is positional encoding crucial in transformer models, and what issue does it address in the context of self-attention operations?

17. When applying transfer learning to fine-tune a pre-trained transformer for a specific NLP task, what strategies can be employed to ensure effective knowledge transfer, especially when dealing with domain-specific data?

18. Discuss the role of cross-attention in transformer-based encoder-decoder models. How does it facilitate the generation of output sequences based on information from the input sequence?

19. Compare and contrast the impact of using sparse (e.g., cross-entropy) and dense (e.g., mean squared error) loss functions in training language models.

20. How can reinforcement learning be integrated into the training of large language models, and what challenges might arise in selecting suitable loss functions for RL-based approaches?

21. In multimodal language models, how is information from visual and textual modalities effectively integrated to perform tasks such as image captioning or visual question answering?

22. Explain the role of cross-modal attention mechanisms in models like VisualBERT or CLIP. How do these mechanisms enable the model to capture relationships between visual and textual elements?

23. For tasks like image-text matching, how is the training data typically annotated to create aligned pairs of visual and textual information, and what considerations should be taken into account?

24. When training a generative model for image synthesis, what are common loss functions used to evaluate the difference between generated and target images, and how do they contribute to the training process?

25. What is perceptual loss, and how is it utilized in image generation tasks to measure the perceptual similarity between generated and target images? How does it differ from traditional pixel-wise loss functions?

26. What is Masked language-image modeling?

27. How do attention weights obtained from the cross-attention mechanism influence the generation process in multimodal models? What role do these weights play in determining the importance of different modalities?

28. What are the unique challenges in training multimodal generative models compared to unimodal generative models?

29. How do multimodal generative models address the issue of data sparsity in training?

30. Explain the concept of Vision-Language Pre-training (VLP) and its significance in developing robust vision-language models.

31. How do models like CLIP and DALL-E demonstrate the integration of vision and language modalities?

32. How do attention mechanisms enhance the performance of vision-language models?

Questions on Fundamental of LLMs:

1. Describe your experience working with text generation using generative models.

2. Could you illustrate the fundamental differences between discriminative and generative models?

3. With what types of generative models you worked, and in what contexts?
   Hint: Different LLM models and in which project you have used it

4. What is multimodal AI, and why is it important in modern machine learning applications?

5. Discuss how multimodal AI combines different types of data to improve model performance, enhance user experience, and provide richer context for decision-making in applications like search engines and virtual assistants.

6. Can you explain the concept of cross-modal learning and provide examples of how it is applied?

7. Explore how cross-modal learning enables models to leverage information from one modality (e.g., text) to improve understanding in another (e.g., images), citing applications such as image captioning or visual question answering.

8. What are some common challenges faced in developing multimodal models, and how can they be addressed?

9. Identify issues such as data alignment, the complexity of model architectures, and the difficulty in optimizing for multiple modalities. Discuss potential solutions like attention mechanisms or joint embedding spaces.

10. How do architects like CLIP and DALL-E utilize multimodal data, and what innovations do they bring to the field?

11. Explain how CLIP combines text and image data for tasks like zero-shot classification, while DALL-E generates images from textual descriptions, emphasizing their impact on creative applications and content generation.

12. Describe the importance of data preprocessing and representation in multimodal learning. How do you ensure that different modalities can be effectively combined?

13. Discuss techniques for normalizing and embedding different data types, such as using CNNs for images and transformers for text, and how these representations facilitate integration in a unified model.

14. In the context of sentiment analysis, how can multimodal approaches improve accuracy compared to text-only models?

15. Analyze how incorporating visual or audio cues alongside textual data can enhance the understanding of sentiment, especially in complex contexts like social media or video content.

16. What metrics would you use to evaluate the performance of a multimodal model, and why are they different from traditional models?

17. Discuss evaluation metrics that specifically address the challenges of multimodal data integration, such as precision and recall for each modality and overall task performance.

18. How do you handle the issue of imbalanced data when working with different modalities in a multimodal dataset?

19. Explore strategies such as data augmentation, balancing techniques, or synthetic data generation to ensure that models receive sufficient training from all modalities.

20. Can you give examples of industries or applications where multimodal AI is making a significant impact?

21. Highlight fields like healthcare (combining medical images with patient records), entertainment (personalized recommendations), and autonomous systems (integrating sensory data for navigation).

22. What future trends do you foresee in the development of multimodal AI, and how might they shape the way we interact with technology?

23. Discuss anticipated advancements such as improved integration techniques, more sophisticated models capable of understanding context across modalities, and potential ethical considerations in their application.

Questions on Word and Sentence Embeddings:

1. What is the fundamental concept of embeddings in machine learning, and how do they represent information in a more compact form compared to raw input data?

2. Compare and contrast word embeddings and sentence embeddings. How do their applications differ, and what considerations come into play when choosing between them?

3. Explain the concept of contextual embeddings. How do models like BERT generate contextual embeddings, and in what scenarios are they advantageous compared to traditional word embeddings?

4. Discuss the challenges and strategies involved in generating cross-modal embeddings, where information from multiple modalities, such as text and image, is represented in a shared embedding space.

5. When training word embeddings, how can models be designed to effectively capture representations for rare words with limited occurrences in the training data?

6. Discuss common regularization techniques used during the training of embeddings to prevent overfitting and enhance the generalization ability of models.

7. How can pre-trained embeddings be leveraged for transfer learning in downstream tasks, and what advantages does transfer learning offer in terms of embedding generation?

8. What is quantization in the context of embeddings, and how does it contribute to reducing the memory footprint of models while preserving representation quality?

9. When dealing with high-cardinality categorical features in tabular data, how would you efficiently implement and train embeddings using a neural network to capture meaningful representations

10. When dealing with large-scale embeddings, propose and implement an efficient method for nearest neighbor search to quickly retrieve similar embeddings from a massive database.

11. In scenarios where an LLM encounters out-of-vocabulary words during embedding generation, propose strategies for handling such cases.

12. Propose metrics for quantitatively evaluating the quality of embeddings generated by an LLM. How can the effectiveness of embeddings be assessed in tasks like semantic similarity or information retrieval?

13. Explain the concept of triplet loss in the context of embedding learning.

14. In loss functions like triplet loss or contrastive loss, what is the significance of the margin parameter?

15. Discuss challenges related to overfitting in LLMs during training. What strategies and regularization techniques are effective in preventing overfitting, especially when dealing with massive language corpora?

16. Large Language Models often require careful tuning of learning rates. How do you adapt learning rates during training to ensure stable convergence and efficient learning for LLMs?

17. When generating sequences with LLMs, how can you handle long context lengths efficiently? Discuss techniques for managing long inputs during real-time inference.

18. What evaluation metrics can be used to judge LLM generation quality

19. Hallucination in LLMs is a known issue, how can you evaluate and mitigate it?

20. What is a mixture of expert models?

21. Why might over-reliance on perplexity as a metric be problematic in evaluating LLMs? What aspects of language understanding might it overlook?

22. How do models like Stability Diffusion leverage LLMs to understand complex text prompts and generate high-quality images?(internal mechanism of stable diffusion model)

Questions on RAG and Multimodel RAG:

1. What is Retrieval-Augmented Generation (RAG)?

2. Can you explain the text generation difference between RAG and direct language models?

3. What are some common applications of RAG in AI?

4. How does RAG improve the accuracy of responses in AI models?

5. What is the significance of retrieval models in RAG?

6. What types of data sources are typically used in RAG systems?

7. How does RAG contribute to the field of conversational AI?

8. What is the role of the retrieval component in RAG?

9. How does RAG handle bias and misinformation?

10. What are the benefits of using RAG over other NLP techniques?

11. Can you discuss a scenario where RAG would be particularly useful?

12. How does RAG integrate with existing machine learning pipelines?

13. What challenges does RAG solve in natural language processing?

14. How does the RAG pipeline ensure the retrieved information is up-to-date?

15. Can you explain how RAG models are trained?

16. What is the impact of RAG on the efficiency of language models?

17. How does RAG differ from Parameter-Efficient Fine-Tuning (PEFT)?

18. In what ways can RAG enhance human-AI collaboration?

19. Can you explain the technical architecture of a RAG system?

20. How does RAG maintain context in a conversation?

21. What are the limitations of RAG?

22. How does RAG handle complex queries that require multi-hop reasoning?

23. Can you discuss the role of knowledge graphs in RAG?

24. What are the ethical considerations when implementing RAG systems?

25. What is Retrieval-Augmented Generation (RAG), and how does it differ from traditional generation models?
    Hint: Discuss the core idea of RAG models, which combine generative and retrieval mechanisms to enhance output quality and accuracy, and contrast it with standard generative models that rely solely on pre-trained knowledge.

26. How can multimodal data be utilized within RAG frameworks to improve information retrieval and generation?
    Hint: Explore how incorporating various modalities (text, images, audio) into the RAG architecture allows for richer context during retrieval and generates more relevant and nuanced responses.

27. What are the challenges of implementing multimodal RAG, particularly regarding data integration and model training?

Hint: Identify difficulties such as aligning representations from different modalities, ensuring data quality across diverse sources, and managing the complexity of training models that handle multiple input types.

28. Can you describe a specific application of multimodal RAG in a real-world scenario? What are its benefits over unimodal approaches?
    Hint: Provide examples from industries like healthcare (combining patient records and medical images) or education (integrating textbooks and videos) to illustrate how multimodal RAG can lead to better decision-making and content understanding.

29. What evaluation metrics would be suitable for assessing the performance of multimodal RAG systems? How do they differ from those used in traditional RAG models?
    Hint: Discuss metrics such as BLEU for text generation, precision and recall for retrieval tasks, and user engagement metrics, focusing on how they need to account for the performance of both retrieval and generation components in a multimodal context.

30. How would you design a multimodal RAG system for a specific industry, such as healthcare or education? What key components would you include, and how would they interact?

31. What techniques can be used to ensure effective alignment and integration of different modalities in a RAG pipeline?

32. In a multimodal RAG setup, how would you evaluate the quality and relevance of generated content? What metrics or benchmarks would you consider?

33. What challenges do you anticipate when scaling a multimodal RAG system to handle large datasets, and how would you address them?

34. Can you provide an example of a potential ethical concern associated with multimodal RAG systems? How would you mitigate this issue?

    Hint: Candidates should be able to discuss ethical implications, such as biases in training data or privacy concerns regarding sensitive information, and propose strategies for bias detection, transparency, and responsible AI practices.

Questions on fine tuning:

1. What is Fine-tuning?
2. Describe the Fine-tuning process.
3. What are the different Fine-tuning methods?
4. When should you go for fine-tuning?
5. What is the difference between Fine-tuning and Transfer Learning?
6. Write about the instruction finetune and explain how does it work
7. Explaining RLHF in Detail.
8. Write the different RLHF techniques
9. Explaining PEFT in Detail.
10. What is LoRA and QLoRA?
11. Define “pre-training” vs. “fine-tuning” in LLMs.
12. How do you train LLM models with billions of parameters?(training pipeline of llm)
13. How does LoRA work?
14. How do you train an LLM model that prevents prompt hallucinations?
15. How do you prevent bias and harmful prompt generation?
16. How does proximal policy gradient work in a prompt generation?

17. How does knowledge distillation benefit LLMs?

18. What's "few-shot" learning in LLMs?(RAG)

19. Evaluating LLM performance metrics?

20. How would you use RLHF to train an LLM model?(RLHF)

21. What techniques can be employed to improve the factual accuracy of text generated by LLMs?(RAGA)

22. How would you detect drift in LLM performance over time, especially in real-world production settings?(monitoring and evaluation metrics)

23. Describe strategies for curating a high-quality dataset tailored for training a generative AI model.

24. What methods exist to identify and address biases within training data that might impact the generated output?(eval metrics)

25. How would you fine-tune LLM for domain-specific purposes like financial and medical applications?

26. Explain the algorithm architecture for LLAMA and other LLMs alike.

    Transformer architecture

Questions on Vector Database:

1. What are vector databases, and how do they differ from traditional relational databases?

   Hint: Discuss the fundamental differences in data storage, retrieval methods, and the use cases that vector databases are designed to address, particularly in handling unstructured data and similarity search.

2. Explain how vector embeddings are generated and their role in vector databases.

   Hint: Describe the process of transforming data into vector representations using techniques like Word2Vec, BERT, or other neural network architectures, and how these embeddings facilitate efficient similarity searches.

3. What are the key challenges in indexing and searching through high-dimensional vector spaces?

   Hint: Explore issues such as the curse of dimensionality, efficient data structures (like KD-trees, LSH, or HNSW), and the importance of approximating nearest neighbor searches to improve performance.

4. How do you evaluate the performance of a vector database in terms of search efficiency and accuracy?

   Hint: Discuss relevant metrics for performance evaluation, such as recall, precision, latency, and throughput, and how these metrics influence the choice of a vector database for specific applications.

5. Can you describe a scenario where you would prefer using a vector database over a traditional database?

   Hint: Provide examples such as applications in recommendation systems, semantic search, or image retrieval, where the ability to quickly find similar items based on their vector representations is crucial.

6. What are some popular vector databases available today, and what unique features do they offer?

   Hint: Mention databases like Pinecone, Weaviate, Milvus, and Faiss, discussing their architectures, scalability options, and specific features that cater to different use cases.

7. How do vector databases support machine learning workflows, particularly in deploying AI models?

Hint: Explain how vector databases can be integrated into the ML lifecycle for tasks such as model serving, feature storage, and facilitating real-time inference.

8. What techniques can be employed to ensure the scalability of a vector database as the dataset grows?

   Hint: Discuss methods such as sharding, distributed computing, and efficient indexing strategies that help maintain performance in larger datasets.

9. How can you handle vector data that may have different dimensionalities or representations?

   Hint: Explore normalization techniques, dimensionality reduction methods (like PCA or t-SNE), and strategies for maintaining consistency across various data sources.

10. What role does vector similarity play in applications like recommendation systems or natural language processing?

    Hint: Discuss how vector similarity measures (like cosine similarity or Euclidean distance) are crucial for ranking and retrieval tasks in these domains.

Questions on LLMOPs & system design:

1. You need to design a system that uses an LLM to generate responses to a massive influx of user queries in near real-time. Discuss strategies for scaling, load balancing, and optimizing for rapid response times.

2. How would you incorporate caching mechanisms into an LLM-based system to improve performance and reduce computational costs? What kinds of information would be best suited for caching?

3. How would you reduce model size and optimize for deployment on resource-constrained devices (e.g., smartphones)?

4. Discuss the trade-offs of using GPUs vs. TPUs vs. other specialized hardware when deploying large language models.

5. How would you build a ChatGPT-like system?

6. System design an LLM for code generation tasks. Discuss potential challenges.

7. Describe an approach to using generative AI models for creating original music compositions.

8. How would you build an LLM-based question-answering system for a specific domain or complex dataset?

9. What design considerations are important when building a multi-turn conversational AI system powered by an LLM?

10. How can you control and guide the creative output of generative models for specific styles or purposes?

11. How do you monitor LLM systems once productionized?

Questions on evaluation methods:

1. What are some common evaluation metrics used in NLP, and how do you decide which one to use?

2. How do you approach model evaluation differently for generative AI tasks like text generation versus classification tasks?

3. What is the importance of human evaluation in NLP, especially for generative AI?

4. How do you evaluate models for bias and fairness, especially in NLP tasks?

5. What is perplexity, and why is it used to evaluate language models?

6. How do you evaluate the coherence and relevance of text generated by an NLP model?

7. Discuss metrics like BLEU, METEOR, and human evaluation for coherence and relevance, particularly in conversational AI or creative text generation.

8. What methods can be used to assess the diversity of generated text?

9. What role does prompt engineering play in evaluation, especially for models like GPT?

10. What are ROUGE scores, and why are they commonly used for summarization?

11. Explain the ROUGE metric and its variants (ROUGE-N, ROUGE-L) as measures of overlap between model-generated summaries and reference summaries.

12. How would you assess the informativeness and conciseness of a summarization model?

13. How do you evaluate retrieval quality in RAG models, and why is it important

14. What strategies do you use to reduce hallucination in RAG models?

15. How do you determine if fine-tuning has improved a model's performance on a specific task?

16. Discuss comparing baseline metrics with fine-tuned metrics, tracking loss curves, and using task-specific metrics to measure improvement.

17. What challenges arise when fine-tuning large language models, and how do you mitigate them?

18. Talk about overfitting, the need for robust validation datasets, and regularization techniques that ensure generalizability in fine-tuned models.

19. How do you assess the quality of generated samples from a generative model?
    Hint: explain all the évaluations techniques

20. How would you set up an A/B test to evaluate two NLP models?

21. Describe the importance of testing with a live audience, creating control/experimental groups, and using click-through rates or engagement metrics in addition to core NLP metrics.

22. How do latency and efficiency factor into evaluating NLP models, especially in production settings?

23. What's the role of explainability in NLP evaluation, especially for high-stakes applications?

24. How do you measure user satisfaction with an NLP model deployed in a real-world application?

25. What is domain adaptation, and how do you evaluate it after fine-tuning a model on domain-specific data?

26. How would you evaluate the robustness of an NLP model to adversarial attacks?

Some miscellaneous questions:

1. What ethical considerations are crucial when deploying generative models, and how do you address them?

2. Can you describe a challenging project involving generative models that you've tackled
   Hint: discuss the challenge which you faced inside your project managerial round or director round
3. Can you explain the concept of latent space in generative models?

4. Have you implemented conditional generative models? If so, what techniques did you use for conditioning?

5. Discuss the trade-offs between different generative models, such as GANs vs. VAEs.

6. What are the primary differences between Hugging Face Transformers, Datasets, and Tokenizers libraries, and how do they integrate to streamline NLP workflows?

7. Describe how to use Hugging Face Pipelines for end-to-end inference. What types of NLP tasks can pipelines handle, and what are the main advantages of using them?

8. How does Hugging Face's Accelerate library improve model training, and what challenges does it address in scaling NLP models across different
9. hardware setups?

10. How does Hugging Face's transformers library facilitate transfer learning, and what are the typical steps for fine-tuning a pre-trained model on a custom dataset?

11. What role does multi-modality play in the latest LLMs, and how does it enhance their functionality?

12. What are the implications of the rapid advancement of LLMs on industries such as healthcare, education, and content creation?