1. Could you elucidate the fundamental differences between discriminative and generative models?
2. What types of generative models have you worked with, and in what contexts?
3. How do you assess the quality of generated samples from a generative model?(évaluations techniques)
4. Can you describe a challenging project involving generative models that you've tackled?
5. How do you handle mode collapse in Generative Adversarial Networks (GANs)?
6. What ethical considerations are crucial when deploying generative models, and how do you address them?
7. Can you explain the concept of latent space in generative models?
8. Have you implemented conditional generative models? If so, what techniques did you use for conditioning?
9. Discuss the trade-offs between different generative models, such as GANs vs. VAEs.
10. Describe your experience working with text generation using generative models.
11. What is Retrieval-Augmented Generation (RAG)?
12. Can you explain the basic difference between RAG and traditional language models?
13. What are some common applications of RAG in AI?
14. How does RAG improve the accuracy of responses in AI models?
15. What is the significance of retrieval models in RAG?
16. What types of data sources are typically used in RAG systems?
17. How does RAG contribute to the field of conversational AI?
18. What is the role of the retrieval component in RAG?
19. How does RAG handle bias and misinformation?
20. What are the benefits of using RAG over other NLP techniques?
21. Can you discuss a scenario where RAG would be particularly useful?
22. How does RAG integrate with existing machine learning pipelines?
23. What challenges does RAG solve in natural language processing?
24. How does RAG ensure the retrieved information is up-to-date?
25. Can you explain how RAG models are trained?

<https://huggingface.co/facebook/rag-token-nq>

1. What is the impact of RAG on the efficiency of language models?
2. How does RAG differ from Parameter-Efficient Fine-Tuning (PEFT)?
3. In what ways can RAG enhance human-AI collaboration?
4. Can you explain the technical architecture of a RAG system?
5. How does RAG maintain context in a conversation?
6. What are the limitations of RAG?
7. How does RAG handle complex queries that require multi-hop reasoning?
8. Can you discuss the role of knowledge graphs in RAG?
9. What are the ethical considerations when implementing RAG systems?
10. What is Fine-tuning?
11. Describe the Fine-tuning process.
12. What are the different Fine-tuning methods?
13. When should you go for fine-tuning?
14. What is the difference between Fine-tuning and Transfer Learning?
15. Explaining RLHF in Detail.
16. Explaining PEFT in Detail.
17. What is LoRA and QLoRA?
18. Define “pre-training” vs. “fine-tuning” in LLMs.
19. How do models like Stability Diffusion leverage LLMs to understand complex text prompts and generate high-quality images?
20. How do you train LLM models with billions of parameters?
21. How does LoRA work?
22. How do you train an LLM model that prevents prompt hallucinations?
23. How do you prevent bias and harmful prompt generation?
24. How does proximal policy gradient work in a prompt generation?
25. How does knowledge distillation benefit LLMs?
26. What’s “few-shot” learning in LLMs?(RAG)
27. Evaluating LLM performance metrics?
28. How would you use RLHF to train an LLM model?(RLHF)
29. What techniques can be employed to improve the factual accuracy of text generated by LLMs?(RAGA)
30. How would you detect drift in LLM performance over time, especially in real-world production settings?(monitoring and evaluation metrics)
31. Describe strategies for curating a high-quality dataset tailored for training a generative AI model.
32. What methods exist to identify and address biases within training data that might impact the generated output?(eval metrics)
33. How would you fine-tune LLM for domain-specific purposes like financial and medical applications?
34. Explain the algorithm architecture for LLAMA and other LLMs alike.

Transformer architecture

LLM 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?
12. Describe the concept of learning rate scheduling and its role in optimizing the training process of generative models over time.
13. Discuss the concept of transfer learning in the context of natural language processing. How do pre-trained language models contribute to various NLP tasks?
14. Highlight the key differences between models like GPT (Generative Pre-trained Transformer) and BERT (Bidirectional Encoder Representations from Transformers)?
15. What problems of RNNs do transformer models solve?
16. Why is incorporating relative positional information crucial in transformer models? Discuss scenarios where relative position encoding is particularly beneficial.
17. 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?
18. 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?
19. How does self-attention work?
20. What pre-training mechanisms are used for LLMs, explain a few
21. Why is multi-head attention needed?
22. What is RLHF, how is it used?
23. What is catastrophic forgetting in the context of LLMs
24. 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?
25. Why is positional encoding crucial in transformer models, and what issue does it address in the context of self-attention operations?
26. 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?
27. 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?
28. 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.
29. 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?
30. 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?
31. 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?
32. 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?
33. 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?
34. 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?
35. What is Masked language-image modeling?
36. 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?
37. What are the unique challenges in training multimodal generative models compared to unimodal generative models?
38. How do multimodal generative models address the issue of data sparsity in training?
39. Explain the concept of Vision-Language Pre-training (VLP) and its significance in developing robust vision-language models.
40. How do models like CLIP and DALL-E demonstrate the integration of vision and language modalities?
41. How do attention mechanisms enhance the performance of vision-language models?
42. 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?
43. Compare and contrast word embeddings and sentence embeddings. How do their applications differ, and what considerations come into play when choosing between them?
44. 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?
45. 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.
46. When training word embeddings, how can models be designed to effectively capture representations for rare words with limited occurrences in the training data?
47. Discuss common regularization techniques used during the training of embeddings to prevent overfitting and enhance the generalization ability of models.
48. 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?
49. What is quantization in the context of embeddings, and how does it contribute to reducing the memory footprint of models while preserving representation quality?
50. 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
51. 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.
52. In scenarios where an LLM encounters out-of-vocabulary words during embedding generation, propose strategies for handling such cases.
53. 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?
54. Explain the concept of triplet loss in the context of embedding learning.
55. In loss functions like triplet loss or contrastive loss, what is the significance of the margin parameter?
56. 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?
57. 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?
58. When generating sequences with LLMs, how can you handle long context lengths efficiently? Discuss techniques for managing long inputs during real-time inference.
59. What evaluation metrics can be used to judge LLM generation quality
60. Hallucination in LLMs a known issue, how can you evaluate and mitigate it?
61. What are mixture of experts models?
62. Why might over-reliance on perplexity as a metric be problematic in evaluating LLMs? What aspects of language understanding might it overlook?

**Note: prepare the answer for these questions and submit to me through linkedin DM or you can mail me also.**

**Giving you 3 days for preparing the detail and best possible answer of these question**

**95% questions in interview you will get around to these interview questions and from the fundamental concept which i discussed**

Hope for the best