Machine Learning (ML) 基础知识

1. 如何处理 overfitting?
2. 描述 bias 和 variance 的区别
3. Regularization 的作用是什么？常见的 regularization 方法有 L1 和 L2
4. 解释 PR-AUC 和 ROC-AUC 这两种 evaluation metrics
5. 如何处理 imbalanced data?
6. 什么是 hyperparameter optimization?
7. 机器学习中常见的 data split 方法有哪些？比如 train/validation/test
8. 解释 logistic regression 的原理
9. 描述 decision tree 的原理，以及在 inference 时的时间复杂度
10. Random forest 和 XGBoost 有什么区别?
11. 比较 bagging 和 boosting 的异同
12. 描述 K-means 算法的原理和缺点
13. 如何实现 KNN 算法?
14. 解释 PCA 的原理和应用场景
15. 定义 SVM 并描述其优化过程

Deep Learning (DL) 模型和技术

16. 描述 CNN 的结构和原理
17. 比较 RNN 和 LSTM 的区别
18. 详细解释 Transformer 的结构和原理，包括:
19. Attention mechanism 的原理
20. Self-attention 的计算过程
21. Multi-head attention 的作用
22. Positional encoding 的作用和实现方法
23. Encoder-decoder 结构
24. Dropout 的原理和作用
25. 神经网络中常用的 weight initialization 方法
26. 比较 Adam 和 RMSprop 等 optimization algorithms 的异同
27. 什么是 vanishing 和 exploding gradients 问题？如何缓解?
28. 推导 Backpropagation 的数学公式，并描述实现步骤
29. 解释 VAE 的原理和应用场景

Large Language Models (LLMs)

30. 比较 GPT、BERT、T5 等模型的区别和特点
31. 有哪些常见的 tokenization 方法和类型？
32. 描述 LLM 常用的 fine-tuning 方法，如 LoRA
33. 介绍一些 LLM 的 efficient training 和 inference 技术
34. LLM 目前存在哪些局限性？
35. 什么是 prompt engineering？
36. LLM 在工业界有哪些应用？
37. 推荐系统
38. 推荐系统中的 candidate generation 模型通常有哪些？它们的复杂度如何？
39. 常见的 ranking 模型有哪些？它们的复杂度如何？
40. 推荐系统中的特征工程和特征选择有哪些考虑？
41. 介绍常见的 embedding 方法，如 user/item embedding
42. 编程和系统设计
43. 手写实现 logistic regression、decision tree、KNN、K-means 等 ML 模型
44. 手写实现 CNN、RNN、Transformer 等 DL 模型 （手写 self-attention 考过几次）
45. 实现 precision、recall、ROC 等常见评估指标
46. 如何设计一个推荐系统的架构？
47. 如何设计 LLM 的部署和服务流程？
48. 如何优化模型训练和推理的效率？
49. 如何设计实验和 A/B 测试？

LLM 近几个月比较常见的问题

50. What are some ways to adapt LLMs to new tasks?
51. Explain the Transformer architecture, including the encoder, decoder, QKV, and attention mechanisms.
52. Why do we choose decoder-only models instead of encoder-decoder models for certain tasks?
53. What is LoRA (Low-Rank Adaptation) and what is it used for in the context of LLMs?
54. What are the differences and advantages/disadvantages between continue training and LoRA training for LLMs?
55. If we want to train an LLM to output a specific style, what techniques should we use?
56. Describe the differences between BERT and vanilla Transformer models.
57. What are the differences in attention mechanisms and position embeddings across various LLMs like OPT and LLaMA?
58. How would you reduce the latency of an LLM at inference time?
59. What changes are needed when moving an LLM from offline to online usage?
60. Explain the tokenizer mechanism used in LLMs and describe the different types of tokenizers.