***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***

1. 描述CNN的结构和原理
2. 比较RNN和LSTM的区别
3. 详细解释Transformer的结构和原理，包括：
4. Attention mechanism的原理
5. Self-attention的计算过程
6. Multi-head attention的作用
7. Positional encoding的作用和实现方法
8. Encoder-decoder结构
9. Dropout的原理和作用
10. 神经网络中常用的weight initialization方法
11. 比较Adam和RMSprop等optimization algorithms的异同
12. 什么是vanishing和exploding gradients问题？如何缓解？
13. 推导Backpropagation的数学公式，并描述实现步骤
14. 解释VAE的原理和应用场景

***Large Language Models (LLMs)***

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

***LLM近几个月比较常见的问题***

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