2024年最全AI大模型面试题合集

机器学习AI算法工程 2024年12月04日 10:11 广西

向AI转型的程序员都关注公众号 机器学习AI算法工程

1. 你了解ReAct吗，它有什么优点?
2. 解释一下langchain Agent的概念
3. langchain 有哪些替代方案?
4. langchain token计数有什么问题？如何解决?
5. LLM预训练阶段有哪几个关键步骤?
6. RLHF模型为什么会表现比SFT更好?
7. 参数高效的微调（PEFT）有哪些方法?
8. LORA微调相比于微调适配器或前缀微调有什么优势?
9. 你了解过什么是稀疏微调吗?
10. 训练后量化（PTQ）和量化感知训练（QAT）有什么区别?
11. LLMs中，量化权重和量化激活的区别是什么?
12. AWQ量化的步骤是什么?
13. 介绍一下GPipe推理框架
14. 矩阵乘法如何做数量并行?
15. 请简述TPPO算法流程，它跟TRPO的区别是什么?
16. 什么是检索增强生成（RAG）?
17. 目前主流的中文向量模型有哪些?
18. 为什么LLM的知识更新很困难?
19. RAG和微调的区别是什么?
20. 大模型一般评测方法及其准是什么?
21. 什么是Kv cache技术，它具体是如何实现的?
22. DeepSpeed推理对算子融合做了哪些优化?
23. 简述一下FlashAttention的原理
24. MHA、GQA、MQA三种注意力机制的区别是什么?
25. 请介绍一下微软的ZeRO优化器
26. Paged Attention的原理是什么，解决了LLM中的什么问题?
27. 什么是投机采样技术，请举例说明?
28. 简述GPT和BERT的区别
29. 讲一下GPT系列模型的是如何演进的?
30. 为什么现在的大模型大多是decoder-only的架构?
31. 讲一下生成式语言模型的工作机理
32. 哪些因素会导致LLM中的偏见?
33. LLM中的因果语言建模与掩码语言建模有什么区别?

34. 如何减轻LLM中的“幻觉”现象?
35. 解释ChatGPT的“零样本”和“少样本”学习的概念
36. 你了解大型语言模型中的哪些分词技术?
37. 如何评估大语言模型（LLMs）的性能?
38. 如何缓解LLMs复读机问题?
39. 请简述下Transformer基本原理
40. 为什么Transformer的架构需要多头注意力机制?
41. 为什么transformers需要位置编码?
42. transformer中，同一个词可以有不同的注意力权重吗?
43. Wordpiece与BPE之间的区别是什么?
44. 有哪些常见的优化LLMs输出的技术?
45. GPT-3拥有的1750亿参数，是怎么算出来的?
46. 温度系数和top-p、top-k参数有什么区别?
47. 为什么transformer块使用LayerNorm而不是BatchNorm?
48. 介绍一下post layer norm和pre layer norm的区别
49. 什么是思维链（CoT）提示?
50. 你觉得什么样的任务或领域适合用思维链提示?
51. 目前主流的开源模型体系有哪些?
52. prefix LM和causal LM区别是什么?
53. 涌现能力是啥原因?
54. 大模型LLM的架构介绍?
55. 什么是LLMs复读机问题?
56. 为什么会出现LLMs复读机问题?
57. 如何缓解LLMs复读机问题?
58. llama输入句子长度理论上可以无限长吗?
59. 什么情况下用Bert模型，什么情况下用LLama、ChatGLM类大模型，咋选?
60. 各个专长领域是否需要各自的大模型来服务?
61. 如何让大模型处理更长的文本?
62. 为什么大模型推理时显存涨的那么多还一直占着?
63. 大模型在gpu和cpu上推理速度如何?
64. 推理速度上，int8和fp16比起来怎么样?
65. 大模型有推理能力吗?
66. 大模型生成时的参数怎么设置?
67. 有哪些省内存的大语言模型训练/微调/推理方法?
68. 如何让大模型输出台规化
69. 应用模式变更
70. 大模型怎么评测?
71. 大模型的honest原则是如何实现的?
72. 模型如何判断回答的知识是训练过的已知的知识，怎么训练这种能力?
73. 奖励模型需要和基础模型一致吗?
74. RLHF在实践过程中存在哪些不足?
75. 如何解决人工产生的偏好数据集成本较高，很难量产问题?
76. 如何解决三个阶段的训练（SFT->RM->PPO）过程较长，更新迭代较慢问题?

77. 如何解决PPO的训练过程中同时存在4个模型（2训练，2推理），对计算资源的要求较高问题?
78. 如何给LLM注入领域知识?
79. 如果想要快速检验各种模型，该怎么办?
80. 预训练数据Token重复是否影响模型性能?
81. 什么是位置编码?
82. 什么是绝对位置编码?
83. 什么是相对位置编码?
84. 旋转位置编码RoPE思路是什么?
85. 旋转位置编码RoPE有什么优点?
86. 什么是长度外推问题?
87. 长度外推问题的解决方法有哪些?
88. ALiBi（Attention with Linear Biases）思路是什么?
89. ALiBi（Attention with Linear Biases）的偏置矩阵是什么？有什么作用?
90. ALiBi（Attention with Linear Biases）有什么优点?
91. Layer Norm的计算公式写一下?
92. RMS Norm的计算公式写一下?
93. RMS Norm相比于Layer Norm有什么特点?
94. Deep Norm思路?
95. 写一下Deep Norm代码实现?
96. Deep Norm有什么优点?
97. LN在LLMs中的不同位置有什么区别么？如果有，能介绍一下区别么?
98. LLMs各模型分别用了哪种Layer normalization?
99. 介绍一下FFN块计算公式?
100. 介绍一下GeLU计算公式?
101. 介绍一下Swish计算公式?
102. 介绍一下使用GLU线性门控单元的FFN块计算公式?
103. 介绍一下使用GeLU的GLU块计算公式?
104. 介绍一下使用Swish的GLU块计算公式?