面试经验

Friday, September 19, 2025 11:25

LLM算法岗位面试

笔试
Rag
Ai agent
模型微调 -LLM Fine Tuning
提效和自动化： 业务和开发方向，
多模态
推理优化. -- 剪枝，蒸馏，量化
MCP -- Model Context Protocol
Monitoring and Observability

外企：

外企- 公司1

1. Python 中 list 与元组的区别。
2. Python list.append() 和 list.extends() 方法的区别。
3. 一个数组里， 有正数与负数，你写一算法 找出 它们加起来是0 的。

4. Lora 调优中 用到的参数说一下。
5. **开源的Vector db 有哪些?**
6. 你的加速在那里， 相比于开源的 vector db.
7. **Jira 数据中文本里的表格， 你们是怎么处理的?**
8. **如果是太多的 jira tickets 作为context , 创建prompt. 不会很多吗?**
9. 说一下 lora 理论的基础? 为什么用低秩? PEFT, parameter-eficent fine tune, adaptor ,lora
10. 你了解强化学习吗? ppo
11. Reward model 怎么训练的? 连续的数字。

外企 - 公司2

1. 给我说一下你最有成就的一个项目? 或者你最想说的一个项目?
2. Cross encoder rerank VS llm rerank, 你是怎么做实验发现 llm rerank 更好?

3. 在做smart jira 的时候 prompt 你是怎么调优的?
4. 在做Smart jira 项目的时候，遇到了哪些困难？ 怎么解决的?
5. 怎么测试知识数据库的召回率的？你们是事先标注过数据吗?
6. 为什么要rerank，在做知识检索的时候，你们具体遇到什么问题?
7. BizPro bot 里， 在上线之后，你们怎么监控生成答案的质量的?
8. 在生产环境，你们怎么debug的？ 比如生成的答案有问题。怎么debug？ 怎么fix?

外企 - 公司3

1. 挑一个你成功的项目说一下
2. Smart jira 项目 是怎么样的?
3. 出于什么目的要微调模型?
4. 你是怎么管理你团队的。
5. 传统项目你挑一个说一下。
6. 你们是敏捷开发团队吗?
7. 文本格式化这个项目说一下。
8. 你会go 语言吗?
9. Rag 的那个智能助手准确率多少?
10. 智能对话助手，如果是用户问你和rag 知识没有关系的时候，怎么办?
11. 你们做的都是toB的项目？ 有toc的项目吗?
12. 你的研究生学历是在职的吗?

1. Rag 项目怎么优化
2. 你的微调后的模型表现是怎么样?
3. 微调的目的是什么?
4. 微调后的模型用在什么地方?

外企 - 公司4

1. Rag 中评估方法有哪些?
2. Rrf 融合的方法不对。还有没有其它的融合方法?
3. Jira 的数据的预处理是怎么样的?
4. 你们为什么不用tf-idf 而用bm25? 你知道它们背后的逻辑吗?
5. 有没有build过 mvp client 和 mvp server
6. 为什么没有用语义分块?

外企 - 公司5

1. Beam search 原理是什么?
2. Agent 里面的步骤是什么优化的？比如原来得到一个结果原来是5个步骤，这个优化这个步骤呢？下次的时候
3. 我的客服智能对话机器人。 怎么适配每个不同的场景呢?

外企 - 公司6

1. 把一个文本和句子转成一个embedding， 请问embedding 怎么出来的?
2. 一个简单的rag 流程是什么?
3. 你们开发的工具有哪些?
4. 我的资料有word , pdf, excel , image. 。。。 怎么把它转成知识向量数据库。

外企 - 公司7

1. Smart jira
2. Automatic text format and update . 具体的流程和测试。
3. 流程 -你们公司的是怎么推进一个ai 的项目

外企 - 公司8

1. 给我介绍一下你认为最满意的项目。
2. 我的问答系统并发量（QPS）上千，请你给我设计一个架构能处理这样的场景。

名企:

名企- 公司1

1. 说一下rag 流程?
2. 你们用的什么向量数据库?
3. 评估 lora finetune 的模型有哪些方式? 技术（BLEU, ROUGE, METER, Bertsocre），人工评估，业务评估
4. 你的微调的数据是怎么来的?
5. Temperature 有什么作用
6. Lora 中 一些调优参数?
7. 训练时候模型的参数类型是什么?
8. 一个prompt 怎么写最好。

名企- 公司2

1. RAG 流程怎么优化的
2. Cross encoder 的原理是什么?
3. 你在RAG bizpro bot 里负责是什么?
4. 你印象最深的项目里挑战是什么?

4. 代码审核的项目里挑战是什么?
5. 如果模型出现错误了，怎么办?
6. 你的bizpro bot 或者smart jira 访问量多少?
7. 你的bizpro bot 准确率是多少?
8. 你认为rag 在未来的发展是会是怎么样?

名企 - 公司3

1. 梯度消失解决方法? 激活函数，参数初始化，正则化与归一化方法， 残差连接
2. 梯度爆炸的解决方法? 梯度裁剪， 合适的权重初始化， 归一化方法， 优化器选择， 残差连接，**L2 正则化**
3. 过拟合的解决方法? 数据增强， 降低模型复杂度， Dropout， Early Stopping，
4. Self-attention , q k, v 代表什么意思?
5. Ai agent 里幻觉是怎么解决呢?
6. Mb25 原理是什么?
7. Tf-idf 的原理是什么?

名企 - 公司 4

1. 长期记忆，短期记忆，用户画像， 你们有用吗?
2. Rag 中 retrieval 你们是怎么实现的?
3. 你做过ai agent 没有?
4. 有两个文档，有冲突，你是怎么解决的?
5. 你是怎么清理的你的数据的?
6. 你是怎么测试你的rag 系统的?
7. 你们用pdf 文件吗?

银行

银行- 公司1

1. Springboot 升级其实有很多问题。你们是怎么升级的吗?
2. 微调模型的时候，你们遇到什么问题?
3. 微调模型之后，是怎么测试的?
4. PV 与pvc 区别是什么?
5. 你现在面试的公司有哪些? 如果你要挑选一个，你会选哪一个?
6. 如果你现在还能继续待在你现在工作的公司，你还会继续待下去吗?
7. 你的ocr 代码review 系统，源代码放进去的话，没有token限制的问题?
8. 什么是微服务?
9. 你们开发的ai 流程是怎么样的?

车企

车企 - 公司1

1. Docker compose 什么意思?
2. 云服务上面，你们都做了什么?
3. Lora 微调的时候，学习率是怎么调的?
4. 你们是怎么使用lora 微调的? 比如数据集怎么准备的? 超参怎么调的?
5. Rag 的检索这块是怎么做的? 关键字检索和向量检索的比重是怎么调的?
6. Docker 与dockerfile 有什么关系?
7. 如果我要换镜像，怎么在dockerfile 换?
8. Rag 问答的时候，多模态是怎么实现的?
9. Deepseek 里的 grpo是什么原理?
10. 最近你读的论文有哪些
11. MCP 与function call 的区别是什么?

车企 - 公司2

1， Llama 的位置编码是怎么样的?

大厂

大厂- 公司1

1. 说一下RAG 的流程 ?
2. 说一下GPT 模型与transformer 模型关系?
3. 说一下transformer 里 self-attention 的公式。 Softmax(q*k^/d) *V
4. 你们用的是什么向量数据库?
5. 你们用的是自己写的向量数据库? 为什么不用开源的?
6. 你们用的是什么向量模型（Embedding 模型）， MiniLM. Mpnet， Genmini embedding api, azure openai Embeding api
7. 你们用的是什么大语言模型? Genimi 1.5 pro. -> Gemini 2.0 flash or Gemini 2.0 pro
8. 说一下你的这些创新项目是怎么上线的?
9. 笔试题： 矩阵相乘，二分查找，在一个列表里查找某一个值。

二面

10. 自我介绍
11. 介绍一下最近的项目，从收益，价值，效益，这些角度出发，
12. 你的text format 项目，怎么知道10这么多单词改变之后，你就要引入人工评估。

13. 你的这么多收益，有没有给公司带来裁员。 如果没有裁员，怎么衡量你的这个收益。
14. Lora 的原理是什么?
15. 低秩的两个矩阵可以表示满秩的矩阵吗?

    笔试， 我现在只有一个银币，正面和反面的概率各是1/2. 我现在要用它模拟骰子。 骰子每个面的概率是1/6. 用python 写一个程序，实现它。

大厂-公司2

1. 表单合并这种情况怎么处理。
2. 文档里既有图片又有表格怎么分片
3. Embedding model 你们用什么?
4. Cross encoder 重排是什么原理? 为什么用它?
5. 混合搜索的原理是什么?
6. Temperature 的作用是什么?
7. 你的非全研究生是怎么回事?

大厂 - 公司3

1. 用户在微博上每天会发表很多。我想实时的知道最热门的榜单有哪些? 用自然语言方式怎么实现?
2. 一个表里有 班级，学生，分数。 我想知道每个班里前三名的学生和分数。
3. Transformer 整个架构是怎么样的?
4. 你的smart jira 整个流程是怎么样的? Prompt 是怎么样的?
5. 问答助手是解决什么问题? 整个流程是怎么样的?
6. Ai agent 项目你做了什么?
7. 什么样的项目是AI agent? 它的定义是什么?
8. 你觉得transformer 和 rnn , lstm 相比，它的优势是什么?
9. 梯度消失怎么解决?
10. Sft 和强化学习， 强化学习微调在什么情况用? 它有哪些特点。
11. 你门微调的模型，是怎么微调的?
12. 问答助手的效率调优这块你们是怎么做的?
13. 怎么避免幻觉问题呢?

创业公司

创业 - 公司1

1. Reward model 里数据集怎么准备的?

2. bizProb bot 给我说一下。
3. 微调的时候参数类型是什么?
4. 你们用什么模型微调的?
5. 这些项目你负责什么?
6. 分片的方法是什么?如果是用句子分片的话，上下文不是丢掉了吗?

创业- 公司2

1. 你用sft 微调过模型吗?
2. 微服务怎么拆分的?有没有用ddd，gateway 用的是什么?
3. 智能助手的rag 流程是怎么样的?
4. 你们的知识数据库是什么?
5. 你们有没有用bge m3 。

创业 - 公司3

1. 你最近几年做的项目里，哪些是最后发现有问题的。做不下去了。
2. 你最近几年的长期规划是什么?

创业 - 公司4

1. Prefile and decode
2. Kv cache 的作用是什么?
3. Deepspeed 与 vllm 有什么区别?