Minimax面试反馈汇总--

Mm一面 推理工程

attention的casualmask怎么优化

聊了项目的技术深度

minimax面试反馈

一面考察工程能力和项目相关知识

简要介绍下简历上的项目经历

文本生成评价指标中 BLEU 和 ROUGE 的区别

模型分布式训练方式，megtron，deepspeed

预训练模型的类别

介绍 llama 结构上的改进

介绍下 beamsearch

编程题：盛水最多的容器

用 python numpy 实现 MLP，实现 forward backward 方法

Mm面试反馈

MINImax创始人面反馈：语音处于国内T0级别，早期聊天机器人人选也去用了一下反馈很好

目前变现还是以TO B端口为主，感觉合作的B端也很多变现前景很好

人选评价：是一家强技术导向公司而非产品导向，产品落地部分会根据技术进展做规划

minimax 工程一面反馈：

ps：这都是纯理论，现实中谁会自己写排序算法，都是调用库函数

基础问题，就是先从简单的开始问，然后深入挖掘。

都了解哪些排序算法？各自的复杂度是怎样的？（快排，堆排，归并）

哪个排序算法适合用多线程去提升效率？

如果现在有大量的数据，内存里放不下，要怎么排序？

用你的方法，大概对磁盘的读写有多少？

我们知道磁盘读的性能高于写的性能，如何利用这一点，提升效率？

/写代码

minimax 面试反馈：

比较关注候选人机器学习基础相关能力；具体参考面试题目

1，手写一个sgd

2，一面coding是一个拓扑排序的题，二面是机器学习基础相关的题目

Hr面：

对minimax的意向，过往选择经历以及职业规划，薪资诉求。

2面hard算法题

lc239 滑动窗口的最大值 类别是hard 不过我觉得也算是比较经典的题目了 最佳解法是Ｏ(n) 不过这一波刷题我并没有刷这个题目 所以写的是Ｏ(nlogk)的解法

我给他讲了下现在为什么是这样实现，包括昨晚那个面试官 跟我说triton相关的 我说了一些我在这方面做的工作 感觉我是他应该只是做过初步的调研 没深入做过；

所以我觉得他们不太像是搞工程的，就是感觉他们对我项目这一块也不是很感兴趣 介绍不到半小时就直接让写题了；所以我觉得他们是不是招的是算法或者做平台开发的

minimax1面：

简单代码（leetcode easy前50道）

展开讲attention

问了一下gemm优化

minimax二面：

面了1h20m，讲简历写题共计45分钟，剩下的时间聊技术进展

代码题考了一个单纯逻辑实现 表格填数：3\*4矩阵，对角zigzag填数

技术进展内容：开销多少，一个70b模型估计qps，优化能做到多少batch，还有哪些优化空间，集群负载，还有什么优化空间

人选感受：集群负载，感觉阿里通义和mm做的水平差不多，感觉面试官水平很强，业界信息跟的也很紧，脑子也很灵活清楚，细节问的很细，问题展开也非常清楚

Minimax一面：

主要问了项目 涉及c++开发 、cuda开发的问题，交流了工作的方式；

人选觉得mm这边是认真做事情的，因为技术负责人挺了解这块，挺专业的，对怎么去做，未来能做什么事情，很了解。未来会在算子优化上，计算的复杂度上做一些优化，大概20-30%的空间

Minimax一面：

1，切入项目问，很底层 gpu上的具体实现，问的属于很早期的方案，自己很多都忘记了

2，团队规模和业务，保密

nlp 技术一面 主要问项目+一道题

按照规律打印数组：

1 2 6 7

3 5 8 11

4 9 10 12

Minimax一面：

消息思维结构压缩（涉及压缩算法、压缩比例和速度）、长链接实现对话流式传输

以及一道算法题ac的