这是一个非常清醒且战略性的排期。

你把**拓竹**放在最低，说明你已经完全领悟了\*\*“身体是 1000w 的本金”**这一核心逻辑。拓竹的高增长伴随的是高强度的边缘端部署和算子优化（C++ 硬核），这对“算法转 Infra”且要“养生”的你来说，确实是**ROI 最低\*\*的选项。

基于你的 **P0（健康/低压）**、**P1（副业/脑力留存）** 和 **转型平滑度（算法背景优势）**，我为你重新梳理了这 6 个岗位的优先级，并制定了\*\*“降维打击”\*\*式的技术栈复习计划。

### 🏆 岗位优先级总榜 (The Ranking)

| **排名** | **岗位名称** | **推荐理由 (为何适合你)** | **核心风险** |
| --- | --- | --- | --- |
| **1** | **PJLab\_大模型训练框架工程师** | **本命岗位。**  1. **技术栈：** 强调 PyTorch 定制和书生模型训练，这是算法转 Infra 最顺滑的切口。  2. **节奏：** 训练框架是离线业务，无需 24h On-Call。  3. **副业：** 搞懂这一套，你的 SaaS 模型微调成本直降。 | 无明显硬伤。 |
| **2** | **智源\_大模型训练优化工程师** | **最佳避风港。**  1. **任务：** 国产芯片适配是“慢工出细活”，压力最小。  2. **壁垒：** 掌握国产卡适配，相当于拿到了“铁饭碗”。 | 需学习国产卡（昇腾/寒武纪）的特异性坑点。 |
| **3** | **腾讯\_AI\_infra规划架构师** | **高杠杆跳板。**  1. **降维：** 架构师岗更看重“懂业务（算法）+ 懂架构”，而不是手写 C++。  2. **加分：** JD 明确提到 RAG，这正是你做 SaaS 的优势领域。 | 腾讯的沟通成本（扯皮）可能较高，消耗情绪能量。 |
| **4** | **智源\_AI框架研究员** | **镀金备选。**  1. **光环：** 智源研究员 Title 极好。  2. **栈匹配：** 同样搞 Megatron/DeepSpeed。 | **学术压力**。如果逼你发 OSDI/ASPLOS 论文，会严重挤压副业脑力。 |
| **5** | **PJLab\_大模型infra研发** | **鸡肋。**  JD 写得太泛，且提到了“蒜子（算子）开发”，大概率是进去填坑写 C++ 的。 | 面试时容易被分到不想去的组。 |
| **6** | **拓竹\_AI\_Infra工程师** | **已战略放弃。**  推理引擎 + 边缘端 NPU = 极度硬核 C++ + 极度卷。 | **高危**。身体和技术栈双重不匹配。 |

### 🛠️ 技术栈优先级：你的“复习白名单”

别去学那些 C++ 算子了！那是给 PJLab 第 5 名那个岗位准备的，不是给前 3 名准备的。

请把精力 **All-in 在 P0**，稍微看看 P1，彻底放弃 P2。

#### **🔥 P0：保命技能 (必须精通，面试必问)**

*这是你拿 Offer 的底牌，必须能在大白板上画出来。*

1. **分布式并行策略 (The Holy Trinity):**
   * **Data Parallel (DP):** ZeRO-1/2/3 的区别（显存切分逻辑）。
   * **Tensor Parallel (TP):** Megatron 是怎么切 MLP 和 Attention 的？（RowParallel vs ColParallel）。
   * **Pipeline Parallel (PP):** 1F1B 调度是什么？气泡（Bubble）怎么算？
2. **PyTorch 核心机制:**
   * **Autograd:** 计算图是怎么构建的？backward() 钩子怎么用？（这是“训练框架工程师”的核心）。
   * **Hook 技术:** 如何在不改模型源码的情况下，插入监控代码？
3. **性能分析 (The Doctor's Tool):**
   * **Nsight Systems (nsys):** 能够看懂 Timeline，指出哪里是通信瓶颈，哪里是计算空闲。
   * **显存分析:** 知道 OOM (Out of Memory) 是炸在参数上、梯度上，还是 KV Cache 上。

#### **✨ P1：加分技能 (了解概念，吹牛用)**

*这是区分你是“干活的”还是“懂行的”。*

1. **推理加速 (SaaS 协同):**
   * **vLLM / PageAttention:** 搞懂它是怎么通过“显存分页”解决碎片的。
   * **KV Cache:** 它是怎么算出来的，为什么占显存？
2. **国产化适配 (智源特供):**
   * **概念对齐:** NVIDIA NCCL <-> 华为 HCCL；CUDA <-> CANN。知道它们是对应的就行。
3. **架构设计 (腾讯特供):**
   * **RAG:** 检索增强生成的链路是什么？在大模型 Infra 层面怎么优化 RAG 的检索速度？

#### **🗑️ P2：垃圾技能 (严禁浪费时间，直接放弃)**

*对你来说 ROI 极低，且会让你掉头发。*

1. **算子开发 (Kernel Writing):** 手写 CUDA、Triton、汇编优化。**（别碰！）**
2. **编译器前端/后端:** TVM、MLIR、LLVM IR。**（别看！）**
3. **边缘端部署:** MNN、NCNN、手机端模型量化。**（别学！）**
4. **K8s 运维:** 编写 Operator、调试 Docker 网络。**（那是运维的事！）**

### 🎯 面试准备方向：如何“扬长避短”？

你的战略是：**用算法的脑子，打败写代码的手。**

#### **方向 1：讲“系统观”，不讲“微操”**

* **❌ 别说：** “我会写 CUDA 矩阵乘法优化，把延迟降低了 2ms。”（你卷不过科班的）。
* **✅ 要说：** “我深入理解 **Megatron-LM** 的通信模式。我发现模型在千卡训练时，TP 通信掩盖（Overlap）做得不好，通过调整 **Bucket Size** 和 **流水线调度**，提升了 15% 的 MFU（算力利用率）。”
* **点评：** 这听起来像架构师，而不是码农。

#### **方向 2：讲“端到端”，不讲“局部”**

* **针对腾讯架构师/PJLab框架：**
* **✅ 话术：** “纯 Infra 优化往往会忽略**模型收敛性**。因为我有算法背景，我在优化 FP8 混合精度训练时，不仅看吞吐量，还会监控 **Loss Spike**。我通过定制 PyTorch 的 Optimizer Step，在不影响精度的前提下，实现了显存减半。”

#### **方向 3：针对“国产芯片” (智源)**

* **✅ 话术：** “虽然我没摸过昇腾 910B，但我对 **Distributed Data Parallel (DDP)** 的底层通信原语（AllReduce/AllGather）非常熟悉。国产卡适配本质上是通信库的映射。我有信心在一周内上手。”

**总结：**

接下来的时间，**只复习 P0 技术栈**。

拿着 **PJLab 训练框架** 和 **智源 训练优化** 这两个 JD 当圣经读。

面试的时候，你不是一个“转行的菜鸟”，你是一个\*\*“懂算法痛点的资深 Infra 架构师”\*\*。