

中原算力算力方向工程师公开招聘岗位一览表

岗位	岗位职责	任职要求	招聘人数
算法移植与优化工程师	1. 负责完成模型训练/微调/推理框架新版本（如 Megatron、vLLM、SGLang）、主流模型/客户自研模型新版本（如 DeepSeek、GLM、QWen 系列）在国产 GPU 的移植适配、验证和优化，解决相关算子库的问题。 2. 负责完成客户 cuda 类应用、HPC 类应用、开源三方库在国产 GPU 上的移植适配和优化，并对接客户应用的工程化，解决国产 GPU 产品与客户人工智能业务场景对接过程中产生的训练、推理、精度等问题。 3. 负责大语言模型线上推理框架的性能优化，解决高并发、低延迟、高可靠性等核心问题，提升服务吞吐量与稳定性。优化多卡 GPU 集群资源调度与通信效率，支持千卡级训练/推理场景。 4. 适配国产硬件架构或 NVIDIA GPU 硬件架构，利用 CUDA、cuDNN、国产框架等工具链进行算子级优化，提升模型计算效率与显存利用率。 5. 调研并引入前沿技术（如异构计算、AI 编译器优化），推动模型量化、蒸馏等轻量化方案落地。	1. 年龄在 40 岁以下（1986 年 8 月 1 日之后），两年以上相关工作经验，硕士及以上学历，人工智能、计算机、数学或相关专业。 2. 熟练掌握 C/C++或 Python，具备扎实的数据结构与算法基础，有实际性能调优者优先。 3. 熟悉国产 GPU 架构或 NVIDIA GPU 架构及编程模型，掌握 CUDA 核函数优化、显存管理、多流并发等技术，有国产 GPU 实际性能调优经验者优先。 4. 至少熟悉一种主流推理框架，如 PyTorch、onnxruntime、vLLM/SGLang 等。 5. 有分布式系统开发经验，熟悉 MPI、NCCL 等通信库，或参与过大模型训练/推理项目者优先。 6. 有算子优化经验及熟悉 CUDA、Triton 等算子开发语言工具者优先。	7

token 技术工程师	1. 负责 Token 聚合平台核心技术研发与维护,主导平台 Token 相关模块的设计、开发、测试与上线全流程,保障平台高可用、高并发、高安全运行。 2. 对接多类 Token 相关资源(第三方模型 API、链上节点、云算力资源等),负责 API 接口的联调与维护,实现多源 Token 资源的统一聚合与标准化输出,支撑业务场景落地。 3. 负责平台 Token 安全体系搭建,防范智能合约漏洞、Token 泄露、恶意攻击等风险,开展安全审计与应急响应保障用户 Token 资产与数据安全。 4、推动业务需求的技术落地,参与技术方案评审与技术文档编写,沉淀技术资产,赋能团队成长。	1. 年龄在 40 岁以下(1986 年 8 月 1 日之后),两年以上相关工作经验,硕士及以上学历,人工智能、计算机、数学或相关专业。 2. 熟悉数据库(MySQL、PostgreSQL 等)、缓存(Redis)、消息队列(Kafka 等)的使用与优化;掌握 Docker、K8s 等容器化部署技术,能独立搭建开发、测试与生产环境,具备问题排查能力,能处理高并发、高可用场景下的技术问题。 3. 学习能力突出,能快速跟进前沿技术并应用于实际工作。具备较强的逻辑思维、问题拆解与独立解决问题的能力,责任心强,能承受一定的工作压力。 4. 具备良好的沟通协作能力,能清晰表达技术观点,有技术分享与团队赋能意识者优先。	3
-------------	---	---	---