项目需求

|  |  |
| --- | --- |
| 项目名称 | AI算力平台建设项目 |
| 需求内容：为了能够让医院业务尽快使用AI大模型技术赋能业务，加速医院的数智化转型，需要构建医院统一的AI算力平台，实现底层算力资源的统一和AI应用开发能力的统一，同时提高GPU算力资源的利用率和AI应用开发能力的共享。本次AI算力平台建设项目需求如下：1. 配置两颗 C86 CPU，每颗 CPU 配置：核数≥64，支持超线程，频率≥3.0Ghz，双散热片，内存≥2T，配置 4 块 1.92T SSD，千兆电口≥2个，万兆光口≥6 个。
2. 配置 8 块国产 GPU。每块 GPU 配置：显存≥64GB，全高全长双宽，PCle 接口≥16，FP16 精度下 TFLOPS≥196，INT8 精度下 TOPS≥393。
3. 提供配套麒麟、统信等服务器操作系统。
4. 提供≥8套GPU显卡授权。
5. 兼容主流开源大模型，提供模型管理、模型压缩、模型安全、模型推理服务部署等不同的阶段能力，降低推理成本，提供更高推理性能、更可靠的推理服务以及保障专属模型安全价值，降低管理使用大模型门槛。
6. 平台需实现自主可控，所投产品为国产品牌，要求AI算力平台产品的软件为完全自主研发。
7. 支持数据可视化及信息整合呈现能力，支持展示算力池相关信息，包括但不限于算力资源总量、已使用量、剩余量等，可查看显卡已占用、未占用或者健康状况；
8. 支持简单易用的模型上传管理功能，便于后续模型管理。支持标准大模型从平台外部导入和从训练任务导入2种方式上传模型。支持自定义模型通过本地以及指定存储路径便捷地导入自定义模型。
9. 为保障AI能力的可持续增长，平台兼容主流开源大模型，支持多种主流开源大模型管理，降低模型准备周期。支持的模型类型包括但不限于：LLama系列、Qwen系列、DeepSeek-R1蒸馏系列模型、DeepSeek-R1-671B模型等。
10. 支持微调和预训练能力。针对数据集的运用，需提供科学合理的数据集推荐配比，确保模型训练能达到最优效果。为降低操作难度，需提供向导式的训练配置，支持一键启动私有大模型的训练流程，极大提高训练效率，推动业务快速发展。
11. 支持可视化管理训练Checkpoint，根据时间线查看checkpoint，支持设置按频率周期性定时保存训练任务的checkpoint，支持选择某一checkpoint进行模型发布和断点续训。
12. 支持创建大模型服务时选择单实例算力分配方式，包括整卡分配和自由切分。自由切分支持通过制定主机和GPU，并按算力百分比和显存大小分配算力，显存支持GiB和MiB2种粒度，最小粒度为1MiB。整卡切分支持选择算力规格。
13. 平台提供模型压缩技术，如支持GPTQ量化压缩、量化对齐等，保持模型性能，减少模型的大小和计算复杂度。
14. 支持动态加密技术，对模型参数、结构实现加密，防止未经授权的访问和攻击，确保模型在各种应用场景中的安全和可靠运行。
 |
| 使用科室制定需求 |  |
| 信息工程师审核 |  |
| 信息管理处 |  |