

AI 工作负载与采购需求

先确认实际任务，再讨论 GPU 数量与平台。不确定的项目留空，由方案工程师继续确认。

01 / 项目信息

确认项目	填写 / 核对要点	记录
项目与部门	名称、对接人、记录日期	
任务类型	推理 / RAG / 微调 / 预训练 / 渲染 / CAE	
模型与框架	模型名称、版本、参数规模；PyTorch / TensorFlow 等	
输入与输出	文本长度、图像尺寸、视频长度；输出目标	
精度与训练方式	量化精度、全参数 / LoRA 等方式	
规模与并发	批量、上下文、并发用户；目标完成时间	
数据规模	现有数据量、增量与预处理方式	
预算与交期	预算范围、采购阶段、期望交付时间	

GPU 计算与 AI 训练平台不能只按卡数判断；显存、互联、框架支持和机房条件需要一起确认。

部署条件与工程确认

用于记录待确认事项，不构成报价、兼容性认证或固定交期承诺。

02 / 硬件与机房

确认项目	填写 / 核对要点	记录
GPU 与显存	拟选型号、数量、显存；现有验证结果	
互联与扩展	单机 / 多机；卡间与节点间通信需求	
CPU 与内存	数据预处理、单任务内存峰值和扩展需求	
NVMe 与容量层	热数据、检查点、结果文件和备份位置	
网络与管理	计算网络、管理网络、远程访问方式	
机柜与供电	机柜深度、U 位、PDU、可用功率与冗余	
散热与环境	现场散热条件、持续负载与噪声限制	

03 / 待确认记录

问题 / 风险	确认责任方	结论 / 日期

关联资料：[AI 平台方案](#) · [选型白皮书](#)