《实战进阶：AI 应用程序和大模型计算能力提升落地实务》

--课程大纲

课程背景：

当前，在全球步入“智能原生”深水区的背景下，人工智能已从辅助工具演变为国企数字化转型的“核心驱动引擎”。随着生成式大模型（LLM）的广泛应用，国企信息中心正面临从传统硬件维护向智能化算力调度的角色跃迁。然而，当前普遍存在“算力盲目投入”与“效率黑盒”的矛盾，员工往往在不理解张量计算（Tensor）、模型权重与硬件拓扑（Topology）逻辑的情况下进行作业，导致高价值算力资源的极大浪费。AI 应用的部署不仅仅是点击“运行”，更是一场关于模型推理能力与算力效率的博弈。

与此同时，AI 驱动的全链路自动化攻击体系已经形成，这使得 AI 应用和算力平台本身成为了黑客眼中的“皇冠上的明珠（Crown Jewels）”。在国企“新质战斗力”的建设过程中，算力的使用规范已直接挂钩国家数据安全合规要求。依据最新发布的 GB/T 45577-2025 标准，企业在进行 AI 模型开发、测试、发布与运维的全生命周期中，必须建立起可感知的“网络空间地形图”，以应对 AI 驱动的智能化、隐蔽化勒索攻击所带来的威胁。

针对上述挑战，本课程立足于“理解底层逻辑、掌握优化技巧、守住合规红线”三大维度，

旨在协助信息中心员工构建起一套安全、合规、高效的算力应用体系。我们将通过深度解析 Transformer、CNN、RNN 等主流架构的算力需求特性，指导学员如何在本地环境（如 Docker）与云端平台（如阿里 PAI）之间进行最优算力配置。课程不仅关注运行速度的提升，更强调在复杂的网络空间环境中，如何通过安全监测预警与异常行为识别，将 AI 算力转化为支撑国企高质量发展的稳健动力。

课程收益

1. **提升 AI 应用逻辑重构与理性判断能力**：透彻理解深度学习张量运算及不同架构（Transformer/CNN）的算力消耗特征，实现从“按经验盲目跑数”向“按逻辑科学调优”的高质量转变。
2. **掌握业务效能跃迁与模型部署调优技能**：掌握模型量化、剪枝及知识蒸馏等主流优化技术，学会在本地 Docker 环境及 GPU 云端算力集群中进行高性能配置，确保数据安全、显著提升系统运行稳定性。
3. **构建排错韧性与高效持续运营模式**：学会利用 RASP 动态防御及日志溯源技术进行实时排错，构建针对 AI 服务的监控仪表盘，确保业务连续性。
4. **领会与遵守合规底线**：深刻领会“管业务必须管合规”原则，掌握 GB/T 45577 标准下的数据分类分级保护实务，形成防御勒索软件及防止敏感数据泄露的操作习惯。

课程时间：3 天，6 小时/天

课程人员：信息中心团队

课程类型：技术赋能与管理风控融合型

综合性实战进阶课程，理论讲授 + 环境实操 + 案例分析 +场景模拟 + 分组讨论 + 课件移交(包括不限于操作手册、工具环境及代码部分)

课程大纲

第一天：算力基础与本地大模型部署推荐——通俗理解，落地实测

第一讲 ：打好算力地基——AI 应用运行逻辑与架构适配

一、精准认知：AI 模型运行的算力底层逻辑

1. 神经网络计算本质：张量创建（Tensor）、数值计算与张量拼接
2. 硬件需求评估：模型量级（如 7B/13B）对内存、显存与带宽的硬约束
3. “规则驱动”向“模型驱动”的跨越：理解 AI 算力作为新质战斗力的内涵

二、解构 AI 算力逻辑——通俗理解底层算力的运行原理

1. AI 大模型是如何“运转”起来的

(1) 代码执行到结果输出：计算图、张量与算子概念通俗理解

(2) CPU vs GPU：为何 AI 大模型运算效率更加依赖 GPU？（并行计算原理图解）

(3) 内存与显存的博弈：模型加载、中间变量与显存占用情况计算

2. 常见 AI 应用场景的算力需求画像

(1) **文本模式：**大模型对生成式文本的显存消耗量图解：通过大家熟知的模型参数量与显存占用的换算公式来说明

(2) **非文本模式：**图像识别与生成的计算密集特性与要求：批处理对算力的影响，突出并行计算高强度效果

(3) **数据分析类任务**的瓶颈识别：是 IO 瓶颈还是计算瓶颈？

三、架构透析：不同算法模型的算力指纹

1. Transformer 架构：多头注意力机制的并行计算优势解读
2. CNN 卷积神经网络：图像分析中的局部感知与计算密度
3. RNN 系列模型：序列数据的算力瓶颈与梯度消失问题
4. 为什么是 Transformer？--从 BERT 到 GPT 的技术演进趋势进行说明

第二讲 实操环节--AI 大模型本地环境基础配置与安全保障

一、本地硬件环境体检与效能最大化

1. 显卡驱动与环境配置：CUDA、cuDNN 的正确安装与版本兼容
2. 系统资源监控实战：如何用任务管理器和专业工具“看透”资源占用
3. 笔记本与工作站的优化策略：散热、电源管理与性能模式设置

二、掌握本地模型运行优化技巧

1. 模型量化技术入门：浮点数 16 位、整型 8 位量化对速度与精度的影响实测
2. 推理框架选择与配置：Ollama、LM Studio 等工具的后台参数解读

3. 上下文窗口管理：如何通过优化提示词长度降低显存消耗

三、性能测试与评估：算力基准测试

1. 基于 Python 的 Numpy/PyTorch 张量运算性能对比测试

2. 性能评估：计算设备（CPU vs GPU）在不同批处理规模下的吞吐率表现

四、本地大模型部署与测试

1. 课程实战：在本地私有环境中部署一个开源大模型。课前提供调研问卷，根据学员反馈情况，指定部署具体厂商的开源大模型，具体操作流程与效果目标如下：

(1) 模型选型与下载：讲解如何根据硬件条件选择参数规模适合大小的模型

(2) 配置文件修改：调整线程数、GPU 层数加载等关键参数

(3) 效果对比：优化前后推理速度与资源占用率对比记录

第三讲 算力成本构成与部署模式评比(本地模型部署的必要性与综合衡量指标)：

一、算力成本分析

1. 推理成本：Token 计费逻辑与优化

2. 训练/微调成本：GPU 租用与显存预估

3. 算力成本与定价策略（商业核心）

(1) 私有化部署工具：Ollama (本地开发)、vLLM (高并发推理加速)、TensorRT-LLM (英伟达)

(2) Token 消耗：输入/输出 Token 成本优化落地（以 DeepSeek“价格屠夫”策略为例）

(3) SaaS 模式定价：按席位 vs 按调用量 vs 混合定价

(4) GPU 选型指南（示例）：A100 vs A10 vs 4090 --显存需求与并发量估算公式

二、部署模式分类：

(1) 公有云 API vs 云端私有化部署 vs 本地私有化部署（Local LLM），从应用效率，数据安全性要求、成本投入几部分指标综合衡量所选模式

第二天：降本增效，云端资源管理与进阶调优——云部署模式

第四讲 ：云端算力平台搭建与 AI 服务调用

一、理解云端算力模式

1. 云厂商 AI 基础设施概览：从虚拟机到 Serverless 推理服务
2. 计费模式分类：按量计费、包年包月与竞价实例的选择策略
3. 成本控制实战：设置预算警报与资源自动释放机制，需要根据不同应用场景选择不同资源占用模式，统筹兼顾成本要求

二、API 调用优化与并发管理

1. API 调用的网络延迟与计算延迟：识别时间资源占用分布情况，选择最优渠道和方案
2. 并发控制策略：QPS 限制、重试机制与指数退避算法应用
3. 批量请求技术：如何通过合并请求提升吞吐效率

实操环节：编写脚本调用云端大模型 API

(1) 原始性能测试：记录单次调用耗时与并发表现

(2) 应用优化策略：实时异步调用与批量处理改造

(3) 压测对比：优化后的吞吐量提升效果分析

(4) 基于阿里云 PAI 平台的 DSW 环境搭建与资源清理实践

三、API 驱动应用：基于 Flask/Gradio 框架的服务化封装与交付方案介绍

第五讲 ：模型调优，提升运行速度与效率的实用方案

一、模型压缩技术：模型量化、剪枝代码实现与关键功能详解

二、知识蒸馏：如何针对大模型在性能与算力成本之间取得平衡？

三、训练策略优化：计划采样与权重绑定技术应用，如何减少模型参数量、提升训练效率，改善模型性能

第六讲 诊断排错：AI 服务的稳定性保障

一、系统日志溯源：快速定位进程异常退出、显存溢出原因分析

二、痕迹检测功能应用：识别模型运行中的异常调用链

三、实操演练：基于 RASP 技术的零日漏洞“免疫式”拦截

第七讲 提升 AI 运行速度与稳定性的核心方法

一、数据传输与预处理的加速

1. 数据管道优化：减少“木桶效应”，让数据足够支撑模型
2. 缓存机制应用：本地缓存与 Redis 在 AI 推理中的应用场景
3. 数据压缩与传输：减少网络 IO 对算力等待的影响

二、模型微调与推理加速进阶

1. 提示词工程对算力的节省：精准指令减少无效计算轮次

2. 常见报错与稳定性保障：内存溢出的预防与处理

3. 容器化部署入门：利用 Docker 实现环境隔离与快速迁移

第三天：提升排错技能水平，养成安全合规习惯--AI 算力合规使用与习惯养成

第八讲 合规导航：国企数据安全风险评估实务（GB/T 45577-2025）

一、数据资产盘点：识别 AI 模型投喂数据中的“皇冠明珠”

二、分类分级保护：个人信息、重要数据在算力平台上的隔离存储

三、全生命周期管控：从收集、训练到生成、删除的合规核查节点

第九讲 AI 应用常见故障排查实战

一、故障诊断方法论

1. 排错基本流程：复现问题、隔离变量、日志分析

2. 典型报错代码解读：CUDA 内存溢出、连接超时等

3. 日志分析基础：如何从海量日志中提取关键报错信息，支持追溯排查

二、典型场景排错演练

1. 场景一：模型加载失败或推理速度骤降

(1) 排查大模型驱动版本、显存碎片整理与进程冲突情况

2. 场景二：API 调用频繁报错或超时

(1) 排查网络代理、并发阈值与负载均衡问题

3. 场景三：服务器 CPU/GPU 利用率异常飙升

(1) 排查死循环代码、僵尸进程与挖矿病毒风险

4. 防御博弈演练：应对 AI 驱动的智能化威胁

(1) 勒索软件态势：Weaxor、LockBit5.0 攻击路径及针对算力节点的破坏模式分析

(2) 账号与鉴权安全：如何防止远程桌面协议弱口令与 VPN 漏洞导致算力被“肉鸡化”？

(3) 应急响应 SOP：制定发现内网系统感染后的第一时间“隔离、断网、凭证清理”清单

第十讲 ：安全合规与高效工作习惯养成

一、算力使用的安全与合规红线

1. 数据隐私保护：敏感数据脱敏处理与本地化运行优先原则

2. 合规使用开源模型：开源协议风险与境外模型供应链安全解析

3. 资源使用规范：禁止与业务无关的脚本调度，遵循最小权限账号管理原则

二、养成高效的 AI 算力使用习惯

1. 任务调度习惯：利用非高峰时段运行重算力任务

2. 资源释放习惯：任务结束后的显存清理与环境重置，根据业务应用环境需要确定资源释放的时间点与资源释放状态

3. 持续学习习惯：关注新技术动态，更新优化知识库

第十一讲 结项评估：数字化意识与工作习惯养成

一、决策支持：通过仪表盘实时监测算力利用率与安全风险指标

二、习惯塑造：全员签署《算力合规使用承诺书》，建立依规履职底线意识

三、知识测评：AI 算力、算法与安全融合能力综合考核

四、综合实战考核

1. 模拟故障排除流程：给定一个运行异常的 AI 环境，要求在规定时间内定位并解决

2. 优化方案设计：针对一个具体的 AI 业务场景，输出资源配置与优化方案书

五、课程总结与行动承诺

1. 重点知识回顾：构建个人 AI 计算能力知识体系

2. 制定行动计划：基于岗位特点，制定未来 3 个月的算力优化改进目标。

课程总结和展望

1. 重点内容回顾

2. 互动问题讨论

3. 课后任务跟进