运维期末作业——作业要求

项目背景

某政府部门/某高校/某公司想推进智能化转型，希望在本地部署**全量 Deepseek R1** 赋能办公场景。项目选型过程中，甲方强调**高度自主可控**，不被美国卡脖子，采用国产信创产品。同时将大模型接入本地知识库 RAG ，能通过 MCP 协议自动帮助用户完成任务。

前世，你是该项目乙方公司的技术负责人，负责与甲方对接并提供一份可行的项目方案。你采购美国产品遇到美国的**芯片禁运**和**关税制裁**，你的公司破产了，你在搬离公司的马路上被一辆大货车撞倒。你穿越到了提交技术方案的一个月前，你发誓：**这一世，你将夺回属于你的一切！**

项目要求

写一篇项目技术选型报告，叙述一台能运行全量deepseek r1的计算集群的硬件和软件架构，估计该解决方案的各项性能和成本，并给出你选择该方案的理由。

1. 计算速度尽量快（以 tokens/s 或 TOPS 衡量）
2. 成本尽量低（以硬件成本（元）和能耗成本（元/年）衡量）
3. 功耗尽量低（以瓦衡量）
4. 禁止使用美国对中国禁运或限制出售的硬件，例如 A100/H100/H20 等
5. 使用国产硬件（例如华为升腾/昆仑芯）加分
6. 在使用国产芯片时，应考虑生态是否成熟

附加要求：

1. 将大模型接入本地知识库 RAG
2. 大模型可以调用 MCP 服务自动帮用户解决问题
3. 大模型可以自主进行多轮思考和调用 MCP 服务，直到解决问题
4. 保护内网机密数据的同时可以使用联网搜索功能

要点

作为技术负责人，而不是程序员，你的任务是**选型**，而不是实现。应重点叙述相对其他方案的**比较优势**以及你选择该方案的理由。

撰写报告

参考 运维期末作业-报告模板 书写报告，并在 Github 上提交。

可以使用 AI，可以问他人，可以参考网上的资料。

参考部分需要注明链接和原作者。如果他人为你提供了帮助，请在不透露其姓名的前提下将 TA 写入鸣谢部分。