课程机器配置要求

本课程中涉及到多个模型：

- 大语言模型
- embedding模型
- rerank模型

在本地化部署大语言模型时，对于个人电脑配置有一定的要求，其他模型对配置无特别要求，一般电脑都可以支持（其他模型也都可以在CPU上直接执行）。

大语言模型如果采用大模型厂商api（如讯飞星火、deepseek等），个人电脑的配置无特别要求，可以参见一般配置，大模型厂商API很多都有赠送和免费的额度

常规电脑配置

配置	规格
CPU 核数	8核
内存	8G
硬盘	>=256G

如果需要本地部署大语言模型，则需要对个人电脑配置有点硬的要求，特别是没有独立显卡如NVIDIA GPU。建议采用ollama进行部署，可支持GPU和CPU部署，没有GPU的也可以部署。注意的是，不同模型大小对GPU显存和内存大小有不同的要求。

以ollama中int4量化的大语言模型（CPU部署），最低配置参考如下：

模型大小	内存	CPU核数	硬盘
1.5B	1.1G	8核	256G
7B	4.7G	8核	256G
14B	9G	8核	256G
32B	20G	16核	256G
70B	43G	16核	256G
671B	403G	16核	2T

如果非int4量化的模型，可以简单按倍数增加，如fp16 按照4倍

也就是，如果是int4量化的模型，常用1.5B最少需要1.1G、7B最少需要4.7G、14B最少需要9G，32B最少需要20G、70B最少需要43G，671B的最少需要403G。 如果是GPU部署，显卡的显存需要高于最低要求，比如1张4090Ti可以部署32B的模型，2张部署70B， 如果CPU部署，机器的内存要高于最低要求，比如16G内存的个人电脑可以部署14B的模型，以此类推。对于deepseek r1和qwen2.5等系列模型都适用。另外，如果是个人电脑，CPU核数最低8核以上，模型越大核数要求也越高，70B模型CPU部署最好是16核或者更高。磁盘也是根据模型大小不同而不同（7B模型需要5G空间）。

下一节：本章简介(01:40)

下一节