---

marp: true

学前准备——Ollma 本地部署大模型

什么是 Ollama

Ollama 是一个开源框架，旨在简化大型语言模型（LLM）在本地计算机上的部署和管理，使开发者能够轻松加载、运行和微调模型，如 Llama、Mistral、Gemma 等，支持文本生成、代码补全、问答系统等多种自然语言处理任务

安装 Ollama

- 前往 https://ollama.com 下载安装

下载并运行本地模型

- 前往 https://ollama.com/search
- 搜索想要部署的模型，如“Deepseek”
  - 从搜索结果点击进入模型详情页面
  - 选择模型尺寸（如 Deepseek 有 1.5b~671b）
    - DS 1.5b 只需要 cpu 即可运行体验。
    - 其他模型，大约需要 **（2/3 * 数字）GB 的显卡**。
      - 如 6G 显卡可以跑 8B，8G 显卡可以勉强跑 14B

- 拷贝右侧的指令，在 cmd 终端中运行
  - 比如 1.5b 的运行指令是：`ollama run deepseek-r1:1.5b`
  - 首次运行会触发后台模型下载，时间较长