Linux 专版 · V3

部署你自己的私人超级AI管家

Linux云服务器部署保姆级教程 · 从零到完成

不懂命令行？没关系。手把手带你租一台云GPU服务器，用图形化面板管理一切，
让大模型在你的专属服务器上 24 小时运行，随时待命。

Linux 服务器 | 云GPU | Docker | Ollama | OpenClaw | MCP工具

如需手机访问完整教程，请访问 **互动网页版（支持手机 · 含付费解锁）**

适合人群：零基础小白 · Linux 初学者 · 想要私有AI但不会折腾的人

前置要求：会使用浏览器、会复制粘贴命令

全部章节

01 为什么需要私人AI管家

02 整体架构一览

03 租一台云GPU服务器

04 连接你的Linux服务器

05 安装1Panel可视化管理面板

06 配置Docker GPU支持

07 安装与配置Ollama

08 下载AI大模型

09 安装OpenClaw智能体

10 MCP工具生态（扩充版）

11 手机接入：随时随地用AI

12 常见问题排雷

13 完整部署清单

第01章 为什么需要私人AI管家

你每天都在用 ChatGPT、文心一言、Kimi 这些在线 AI，但你有没有想过——**如果有一个完全属于你自己的 AI，24 小时待命、不审查、不限速、数据不上传**，那会是什么体验？

1.1 公有云AI vs 私人AI管家

对比维度	公有云AI（ChatGPT等）	私人AI管家（本教程方案）
数据隐私	对话上传到第三方服务器，可能用于训练	**数据完全在你自己的服务器上，不外传**
使用限制	有速率限制、次数限制、内容审查	**无限制，随便用，不审查**
月费	ChatGPT Plus ¥140/月起	**云GPU ¥1~5/天，按需开关**
模型选择	只能用平台提供的模型	**任意切换开源模型，自己说了算**
工具能力	平台决定，无法扩展	**通过MCP协议自由添加工具**
网络依赖	必须联网，在国内还需梯子	**服务器在国内，无需梯子**
API接入	需要付费API Key	**本地API，免费无限调用**
可定制性	几乎不能定制	**完全可定制人格、记忆、工具**

1.2 部署完成后，你能做什么？

🎓 日常工作

- 写邮件、改简历、润色文章
- 翻译文档（中英日韩…）
- 总结长文、提炼要点
- 头脑风暴，提供创意

💻 技术辅助

- 写代码、Debug、代码审查
- 管理文件、操作数据库
- 自动化日常任务
- 服务器运维助手

📚 学习成长

🏠 生活助手

- 像名师一样解释复杂概念
- 练习外语对话
- 制定学习计划
- 答疑解惑，有问必答

- 通过Telegram随时聊天
- 提醒日程、记录待办
- 搜索信息、查资料
- 24小时在线，永不掉线

💡 **小白视角**

你可以把它理解为一个**永远在线、不会生气、知识面极广**的私人助手。它不需要你精通技术——本教程会用最通俗的方式，带你一步步搞定。

1.3 费用概览

项目	费用	说明
云GPU服务器	约 ¥1~5/天	按小时计费，不用时关机不计费
Ollama	免费	开源大模型运行工具
AI大模型	免费	开源模型，直接下载使用
1Panel	免费	开源服务器管理面板
OpenClaw	免费	开源AI智能体框架
MCP工具	免费	开源工具生态

✅ **预期效果**

全套部署完成后，你将拥有一个**月成本不到200元**、功能媲美付费AI服务的私人AI管家。而且用得越多，越觉得"真香"。

第02章 整体架构一览

在开始操作之前，先了解你要搭建的"大房子"长什么样。不用看懂每一个细节，有个整体印象就行。

2.1 系统架构图

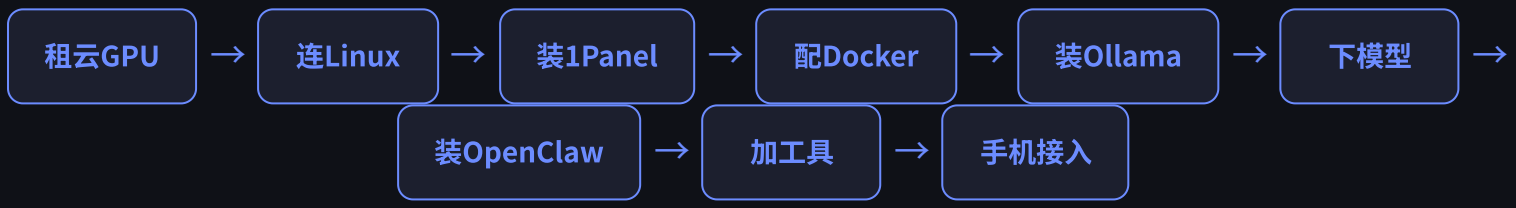
2.2 核心组件解释

组件	是什么（通俗解释）	作用
云GPU服务器	租一台带有NVIDIA显卡的远程Linux电脑	提供AI运算所需的强大算力
Linux	服务器的操作系统（就像电脑上的Windows）	运行所有软件的基础环境
1Panel	一个图形化管理面板，替代命令行操作	让你像用网站一样管理服务器

组件	是什么（通俗解释）	作用
📦 **Docker**	软件打包容器（像集装箱一样标准化运行）	让Ollama、OpenClaw等软件一键安装
🦙 **Ollama**	大模型运行工具	下载并运行AI大模型，让GPU执行推理
🧠 **AI大模型**	AI的"大脑"（如Qwen3、DeepSeek-R1）	真正理解你说话并生成回复的核心
🤖 **OpenClaw**	AI智能体框架	给AI"装上手脚"——让它能搜索、写文件、操作工具
🔌 **MCP工具**	可插拔的工具插件	扩展AI的能力（搜索、写代码、管数据库…）

2.3 本教程部署路线

我们将按照以下顺序一步步完成部署：



> ⚡ **要点**
>
> 整个过程大约需要 **1~2小时**（含模型下载时间）。如果网络慢或对命令行不熟悉，可能需要3小时。不用担心——每一步都有详细截图说明和"为什么"的解释。

第03章 租一台云GPU服务器

AI大模型需要GPU（显卡）来运行，而一张高端显卡动辄上万元。好消息是——**你可以按小时租**，用的时候开机，不用的时候关机，灵活又便宜。

3.1 三大云GPU平台对比

平台	GPU型号	价格参考（3090）	优点	缺点	推荐度
AutoDL	3090/4090/A100/多种	~¥1.5/h	价格便宜、机型丰富、社区活跃	高峰期经常没卡	⭐⭐⭐⭐⭐
优云智算	3090/4090/A100等	~¥1.8/h	界面友好、有Ollama预装镜像、稳定	机型比AutoDL少	⭐⭐⭐⭐⭐
矩池云	3090/A100等	~¥2.0/h	老牌平台、稳定	价格稍贵、界面较旧	⭐⭐⭐⭐

> **要点**
>
> 本教程以**优云智算**为主要示例（因为它提供Ollama预装镜像，最省事）。如果你更喜欢AutoDL，我也会提供简要说明。

3.2 AutoDL 快速上手（备选方案）

1. 注册并登录 `https://www.autodl.com`，充值 ¥50 即可开始（建议先充小额试用）。
2. 点击「社区实例」→ 选择一个带有GPU的镜像，推荐选择 Ubuntu + CUDA 的基础镜像。
3. 选择GPU型号（推荐RTX 3090，性价比最高），然后创建实例。
4. 实例创建成功后，通过网页终端或SSH连接即可。

> **进阶技巧**
>
> **AutoDL 抢卡攻略：**高峰期（工作日晚上8-11点）显卡经常被抢光。建议：① 提前在「无卡实例」中设置开机自动运行脚本；② 关注退卡时段（凌晨2-6点）；③ 使用自动抢卡脚本。

⚠ **踩坑提醒**

AutoDL 的基础镜像是容器类型，**不支持再次安装Docker**（因为容器里已经有Docker环境了）。如果你需要在AutoDL上用1Panel，建议选择**系统镜像**（虚机类型），这样可以自行安装Docker。

3.3 优云智算详细操作（重点！）

下面是本教程的核心操作步骤——在优云智算上创建一台带有Ollama预装的GPU服务器。

3.3.1 注册与充值

1 访问优云智算官网，注册账号并完成实名认证。

充值 ¥50~100 即可体验（按小时计费，随时可以关机省钱）。

[截图：优云智算官网首页]

优云智算官网，点击右上角"注册/登录"按钮

3.3.2 进入部署页面

2 登录后，在左侧导航栏找到 **「GPU算力服务」** 大目录，展开后点击 **「部署GPU实例」**。

[截图：优云智算左侧导航栏]

左侧导航栏 → GPU算力服务 → 部署GPU实例，点击进入

💡 **小白视角**

左侧导航栏就像网站的"大抽屉"，**「GPU算力服务」**是其中一个抽屉，展开后里面有几个小分类，**「部署GPU实例」**就是我们要去的地方——在这里创建你的AI服务器。

3.3.3 选择镜像

进入部署页面后，上方有两大分类标签：

🏷 **平台镜像 ← 选这个**

🌐 **社区镜像**

官方提供和维护的镜像，稳定可靠，包含常用软件。
适合大多数用户。

用户自己制作上传的镜像，种类多但质量参差不齐。
适合有特殊需求的进阶用户。

点击 **「平台镜像」**，你会看到三种镜像类型：

镜像类型	实例类型	能否装Docker	通俗解释	适合谁
基础镜像	容器类型	❌ 不支持再次安装	像住进了一个已经装修好的酒店房间，设施齐全但你不能自己改结构	✅ 新手首选
系统镜像	虚机类型	✅ Ubuntu可以	像租了一套毛坯房，Ubuntu系统可以自己装修	需要完全控制权的人
第三方镜像	虚机类型	✅ 可以	别人装修好的精装房，但装修风格固定，不能发布到社区	特定需求用户

> 💡 **小白视角**
>
> **容器 vs 虚机**的简单理解：容器是"精简版"系统，启动快、资源占用少，但有些高级操作不能做；虚机是"完整版"系统，像自己装了一台电脑，完全自由但启动慢一些。
>
> 我们选择**基础镜像中的Ollama**——因为它已经帮我们把Ollama装好了，省去手动安装的步骤。

3 在 **基础镜像** 列表中找到 **Ollama v0.13.1**，点击选择它。

[截图：平台镜像选择页面]

平台镜像 → 基础镜像 → 找到并点击「Ollama v0.13.1」

3.3.4 右侧配置区详解

选择镜像后，右侧会出现配置区域。我们逐项来配置：

🖥️ 实例配置（选择GPU型号）

GPU型号	显存	参考价格	推荐模型	适合人群
RTX 3090	24GB	~¥1.5~2/h	Qwen3:32B、DeepSeek-R1:32B	⭐ 性价比之王，首选

GPU型号	显存	参考价格	推荐模型	适合人群
RTX 4090	24GB	~¥3~4/h	任意32B以下模型	追求速度、预算宽裕
A100	40/80GB	~¥5~10/h	大模型、多模型并行	专业用户 / 企业

> **要点**
> **推荐选择RTX 3090**——24GB显存足够运行32B模型，价格便宜，性能出色。
> 如果你只是想先用8B小模型试试水，甚至可以选择更便宜的型号。

🔢 GPU数量

可选 1、2、4、8 张GPU。

> **小白视角**
> **选1个就够！**GPU数量是给大规模训练用的。我们只是运行一个AI管家，单张3090绰绰有余。多张GPU = 多倍价格，没必要。

⚙️ CPU配置

- **16C 64G** — 16核CPU，64GB内存 → ✅ **够用了**
- **16C 94G** — 16核CPU，94GB内存 → 更宽裕，跑多个模型不挤

> **小白视角**
> **64G内存已经足够**。简单来说：内存是工作台，GPU是工人。工人干活需要GPU，但也要有足够大的工作台放工具和材料。64G够放好几个大模型了。

💾 系统盘

每100GB收费 ¥0.05/小时，默认200GB，最高可设1000GB。

- **200GB（默认）** — 够下载3~4个大模型
- **500GB** — 模型爱好者推荐，可存10+个模型
- **1000GB** — 重度用户

> **小白视角**
> 系统盘就是服务器的"硬盘空间"。一个AI大模型通常4~20GB，200GB够你先下载几个试试。**先从默认200GB开始**，不够了再扩容。

💾 数据盘（可选）

可以勾选开启，支持部署实例后挂载。

> 💡 **小白视角**
>
> 数据盘 = 额外的存储空间。就像你的电脑有一个C盘（系统盘）和一个D盘（数据盘）。
>
> **暂时不用开**。等系统盘快满了，可以在实例创建后再挂载数据盘，很方便。数据盘的好处是：即使实例被删除，数据盘里的东西还在。

☁️ 云存储

可创建。这是对象存储服务，适合存放大量文件。

> 💡 **小白视角**
>
> **暂时不需要**。这是给需要存大量数据（比如训练数据集、海量图片）的用户用的。我们只是运行AI管家，用不到。

🔧 更多配置

展开「更多配置」后有两个选项：

CPU平台：可选「自动分配」「Intel(x86-64)」「AMD(x86-64)」。一般选**自动分配**即可。

防火墙预设（重要！）

防火墙预设	开放端口	说明
cuda130_torch291_py312	基础端口	通用CUDA环境预设
非Web服务器推荐(22, 3389)	22, 3389	仅SSH和远程桌面
Web服务器推荐(22, 3389, 80, 443)	22, 3389, 80, 443	SSH + 远程桌面 + 网站
Ollama v0.13.1	**含Ollama端口**	**✅ 最省事！自动开放Ollama所需端口**

> ✅ **选择建议**
>
> **选「Ollama v0.13.1」防火墙预设！** 它会自动帮我们开放Ollama需要的所有端口，不用手动去安全组里一条条添加端口规则，省时省力。

> ⚠ **踩坑提醒**
> 如果你不选Ollama防火墙预设，后面就要手动在安全组中开放端口（如11434），这是很多新手卡住的地方。选了预设就能跳过这一步。

3.3.5 点击部署，等待创建

4 确认所有配置无误后，点击 **「立即部署」**。等待1~3分钟，实例创建完成。

[截图：配置确认页面]

右侧配置区汇总：Ollama v0.13.1 + RTX 3090 + 1个GPU + 16C 64G + 200GB + Ollama防火墙 → 点击「立即部署」

3.3.6 记录连接信息 ⚠ 重要！

实例创建完成后，页面会显示以下信息，**请务必截图或记录下来**：

> ⚠ **注意**
> **以下信息后续连接服务器时必须用到，请截图保存！**

- **公网IP地址** — 类似 `123.45.67.89`
- **SSH端口** — 通常为 `22`（也可能不同）
- **登录用户名** — 通常为 `root`
- **登录密码** — 创建时设置或平台生成的密码

[截图：实例详情页 - 连接信息]

实例列表中找到刚创建的实例，查看详情页中的公网IP、端口、用户名、密码信息，截图保存

第04章 连接你的Linux服务器

服务器创建好了，但它现在是一台"看不见摸不着"的远程电脑。我们需要通过终端（命令行窗口）连接上去才能操作它。

4.1 三种连接方式

方式	工具	优点	适合场景
网页终端	优云智算/AutoDL自带	不用装软件，打开浏览器就能用	临时操作、测试
SSH客户端	终端/PowerShell/PuTTY	稳定、速度快、体验好	**推荐日常使用**
1Panel终端	1Panel自带Web终端	在浏览器中操作，和1Panel一起用	装好1Panel之后用

4.2 使用SSH连接（推荐）

打开系统自带的终端应用：

- **macOS / Linux**：打开「终端」（Terminal）
- **Windows**：打开 PowerShell 或 CMD（开始菜单搜索"PowerShell"）

输入以下命令：

```
ssh root@你的服务器IP -p 端口号
```

例如：

```
ssh root@123.45.67.89 -p 22
```

首次连接会提示确认指纹，输入 `yes` 回车，然后输入密码即可。

💡 小白视角

SSH是什么？ 简单来说，SSH就是一种"远程遥控"协议。通过SSH，你的电脑可以安全地连接到远方的服务器，就像坐在服务器面前一样操作。

`ssh` = 远程连接命令
`root` = 管理员用户名
`@` = "at"（在）
`123.45.67.89` = 服务器的"门牌号"（IP地址）
`-p 22` = 连接端口（"门牌号上的房间号"）

4.3 Linux终端基础命令

连接成功后，你会看到一个黑色的命令行界面。不用害怕——以下是你要用到的所有命令：

命令	功能	示例
`ls`	列出当前目录下的文件和文件夹	`ls`
`ls -la`	列出所有文件（包括隐藏文件），详细信息	`ls -la`
`cd`	进入某个目录（Change Directory）	`cd /opt`
`cd ..`	返回上一级目录	`cd ..`
`pwd`	显示当前所在路径（Print Working Directory）	`pwd`
`mkdir`	创建新目录（Make Directory）	`mkdir myapp`
`rm`	删除文件	`rm file.txt`
`rm -rf`	强制删除文件夹及其内容（⚠️谨慎使用）	`rm -rf myapp`
`cat`	查看文件内容	`cat file.txt`
`cp`	复制文件	`cp a.txt b.txt`
`mv`	移动/重命名文件	`mv a.txt b.txt`
`sudo`	以管理员权限执行命令	`sudo apt update`
`systemctl`	管理服务（启动/停止/重启）	`systemctl restart ollama`

命令	功能	示例
`docker`	Docker容器管理	`docker ps`
`curl`	网络请求工具	`curl http://localhost:11434`
`nvidia-smi`	查看GPU状态	`nvidia-smi`
`clear`	清屏	`clear`

4.4 终端操作小技巧

必会技巧

- **Tab键自动补全** — 输入命令或文件名的前几个字母，按Tab键自动补全。例如输入 `cd /o` 然后按Tab，自动变成 `cd /opt`
- **Ctrl + C 中断命令** — 命令运行太久想停？按 Ctrl+C 立即中断
- **↑ ↓ 方向键翻历史** — 按上方向键可以调出之前输入过的命令，不用重新打
- **Ctrl + L 清屏** — 和 `clear` 效果一样，清空屏幕显示
- **Ctrl + A / Ctrl + E** — 光标快速跳到行首/行尾
- **粘贴命令** — 在终端中粘贴通常用 Ctrl+Shift+V（不是Ctrl+V）

4.5 root用户说明

小白视角

root用户是什么？

Linux中，`root` 就是"超级管理员"，拥有最高权限。就像Windows的Administrator账户。

在云GPU服务器上，你默认就是以root身份登录的，所以不需要额外切换。

但要注意：root权限很大，误操作可能导致系统出问题。所以复制粘贴命令时要仔细检查，不要乱删文件。

如果你是以普通用户登录，需要切换到root：

```
sudo su -
```

输入当前用户的密码后，就变成root了。命令行前面的提示符会从 `$` 变成 `#`。

第05章 安装1Panel可视化管理面板

5.1 什么是1Panel?

💡 **小白视角**

1Panel 是一个开源的**Linux服务器可视化管理面板**。简单来说，它把所有命令行操作变成了一个漂亮的网页界面——就像给Linux服务器装了一个"Windows控制面板"。

有了1Panel，你可以：

- ✅ 在浏览器中管理文件
- ✅ 用图形界面安装Docker应用
- ✅ 管理防火墙端口
- ✅ 查看系统资源使用情况
- ✅ 内置终端，不用单独开SSH

1Panel 只能安装在 Linux 上，不支持Windows。这也是为什么本教程是Linux专版。

⚠️ **踩坑提醒**

如果你选的是优云智算的「基础镜像」（容器类型），1Panel可能无法安装——因为容器环境不支持安装systemd服务。

解决方案：选择**系统镜像（虚机类型）**，比如Ubuntu 22.04，然后在虚机中安装1Panel和Docker。

本教程假设你选择的是**系统镜像（虚机类型）**。如果你坚持用基础镜像，可以跳过1Panel，直接在终端中操作。

5.2 安装步骤

1. 先确保系统是最新的：

```
apt update && apt upgrade -y
```

2. 运行1Panel官方安装脚本：

```
curl -sSL https://resource.fit2cloud.com/1panel/package/quick_start.sh -o quick_start.sh && sudo bash quick_start.sh
```

[截图：1Panel安装脚本运行中]

终端显示1Panel安装向导，开始交互式配置

3 安装向导会问你几个问题，按照以下方式回答：

```
# 选择语言
请选择语言：zh              # 输入 zh 选中文

# 选择1Panel安装目录
请设置1Panel安装目录：/opt/1panel   # 默认就是这个，直接回车

# 设置1Panel端口
请设置1Panel端口：直接回车使用默认端口   # 默认通常是随机端口，记住它！

# 设置管理员用户名
请设置1Panel用户名：admin      # 或者你自己想用的用户名

# 设置管理员密码
请设置1Panel密码：输入一个强密码   # 务必记住！
```

⚠ 重要：截图保存登录信息！

安装完成后，终端会显示类似以下信息，**请立即截图保存**：

```
面板地址：http://123.45.67.89:8888/xxxx
用户名：admin
密码：xxxxxxxxxxx
```

这是你登录1Panel的唯一凭证，丢失了需要重置。

[截图：1Panel安装完成信息]

终端显示面板地址、用户名、密码 → 立即截图保存

5.3 浏览器登录1Panel

4 打开浏览器，输入面板地址（如 `http://123.45.67.89:8888/xxxx`），使用用户名和密码登录。

[截图：1Panel登录页面]

1Panel登录界面，输入用户名和密码

[截图：1Panel主界面]

登录成功后看到1Panel主面板，左侧导航栏包含仪表盘、网站、容器、文件管理、防火墙等功能

5.4 防火墙端口开放

为了让你能从外面访问服务器上的各种服务，需要开放一些端口。

> 💡 **小白视角**
>
> **端口是什么？**服务器就像一栋大楼，IP地址是门牌号，端口就是大楼里不同的房间号。
> - 22号房间 = SSH远程连接
> - 80号房间 = 网站访问
> - 443号房间 = 安全网站访问
> - 11434号房间 = Ollama AI服务
> - 3000号房间 = OpenClaw界面
>
> 防火墙就是"门卫"，默认会锁上所有房间。我们需要告诉门卫"这些房间要对外开放"。

5.4.1 在优云智算中开放端口（安全组）

如果你的防火墙预设选择了「Ollama v0.13.1」，则Ollama所需的端口已经自动开放，可以跳过11434端口的配置。但其他端口可能仍需手动添加。

1 进入优云智算控制台 → 找到你的实例 → 安全组/防火墙设置

2 添加需要开放的端口规则：

端口	用途	是否必须	是否已被预设开放
`22`	SSH远程连接	✅ 必须	通常已开放
`80`	HTTP网站	推荐	视预设而定
`443`	HTTPS安全网站	推荐	视预设而定
`11434`	Ollama AI服务	✅ 必须	✅ **Ollama预设已开放**
`3000`	OpenClaw Web界面	推荐	需手动开放
1Panel端口	1Panel管理面板	✅ 必须	需手动开放

[截图：优云智算安全组设置]

在安全组规则中添加端口3000和1Panel端口，协议选TCP，来源填0.0.0.0/0

5.4.2 在1Panel中开放端口

除了云平台的安全组，服务器自身的防火墙也需要放行端口。

1. 登录1Panel → 左侧菜单找到 **「防火墙」**
2. 点击「添加规则」，依次添加需要的端口

[截图：1Panel防火墙设置]

1Panel → 防火墙 → 添加规则 → 端口填3000，协议TCP → 保存

✅ **为什么选了Ollama防火墙预设就不用手动开11434**

因为你创建实例时选择了「Ollama v0.13.1」防火墙预设，优云智算已经在云平台层面（安全组）自动帮你放行了11434端口。

这就好比你入住酒店时告诉前台"我要用WiFi"，前台直接帮你开通了——不用自己再去拉网线。

但其他端口（如3000、1Panel端口）仍需手动开放。

第06章 配置Docker GPU支持

6.1 为什么需要这个步骤?

> 💡 **小白视角**
>
> Docker默认是"看不见"GPU的。想象你请了一个工人（Docker容器），他干活很能干，但你没给他"钥匙"（GPU驱动）去打开GPU这个工具箱。
>
> `nvidia-container-toolkit` 就是这把钥匙。装上它之后，Docker容器就能访问GPU了。

6.2 安装 nvidia-container-toolkit

```
apt-get update
apt-get install -y nvidia-container-toolkit
```

> ⚠️ **踩坑提醒**
>
> **注意拼写！**是 `nvidia-container-toolkit`（中间是连字符 `-`），不是 `nvidia_container_toolkit`（下划线）。拼写错误会导致安装失败。
>
> 正确：`nvidia-container-toolkit` ✅
> 错误：`nvidia_container_toolkit` ❌

6.3 配置Docker运行时

```
nvidia-ctk runtime configure --runtime=docker
```

这个命令会自动修改Docker的配置文件，告诉Docker"以后启动容器时，要支持NVIDIA GPU"。

6.4 重启Docker

```
systemctl restart docker
```

6.5 验证GPU是否可用

运行以下命令来测试Docker是否能正确使用GPU：

```
docker run --rm --gpus all nvidia/cuda nvidia-smi
```

[截图：GPU验证结果]

如果看到GPU型号、显存、驱动版本等信息表格，说明GPU配置成功 ✅

✅ **预期结果**

如果输出类似以下内容，说明一切正常：

```
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 550.xx    Driver Version: 550.xx    CUDA Version: 12.x           |
|-------------------------------+----------------------+----------------------+
| GPU  Name                 Persistence-M | Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp   Perf          Pwr:Usage/Cap |         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  NVIDIA GeForce RTX 3090    On      | 00000000:00:04.0 Off |                  N/A |
| 30%   35C    P8             20W / 350W  |      4MiB / 24576MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
```

⚠️ **踩坑提醒**

如果你选择的是优云智算的「基础镜像」（容器类型），Docker可能已经预装好了，甚至nvidia-container-toolkit也已经配置好了。可以先直接跑验证命令试试——如果能看到GPU信息，就跳过本章。

如果你选择的是「系统镜像」（虚机类型），则需要按照上述步骤完整安装。

第07章 安装与配置Ollama

7.1 两种情况

情况A：选了Ollama基础镜像

恭喜你！Ollama已经预装好了，你只需要做配置（第7.3节），跳过安装步骤。

情况B：没选Ollama镜像

需要手动安装Ollama，按第7.2节操作。

7.2 手动安装Ollama（情况B）

1 运行官方安装脚本：

```
curl -fsSL https://ollama.com/install.sh | sh
```

等待安装完成，会显示 "Ollama installed successfully"。

[截图：Ollama安装完成]

终端显示Ollama安装成功的信息

7.3 配置Ollama监听地址（两种情况都要做！）

> 💡 **小白视角**
>
> Ollama默认只监听 `localhost`（本机），意思是只有服务器自己能访问它。但OpenClaw（稍后安装）需要通过网络连接Ollama。
>
> 我们要把监听地址改为 `0.0.0.0`，意思是"接受所有网络地址的连接"——就像把办公室的门从"只有内部员工能进"改成"有门禁卡就能进"。

1 编辑Ollama的配置文件：

```
mkdir -p /etc/systemd/system/ollama.service.d
cat > /etc/systemd/system/ollama.service.d/override.conf << 'EOF'
[Service]
Environment="OLLAMA_HOST=0.0.0.0"
EOF
```

进阶说明

如果你的Ollama是通过Docker运行的（基础镜像情况），则配置方式不同。需要修改Docker容器的环境变量或docker-compose配置：

```
# Docker方式运行时，添加环境变量
docker run -d -e OLLAMA_HOST=0.0.0.0 -v ollama:/root/.ollama -p 11434:11434 ollama/ollama
```

2 重启Ollama服务：

```
systemctl daemon-reload
systemctl restart ollama
```

3 验证Ollama是否正常运行：

```
curl http://localhost:11434/api/tags
```

预期结果

如果返回类似 `{"models":[...]}` 的JSON数据（可能是空数组，因为还没下载模型），说明Ollama已经正常运行了。

[截图：Ollama验证结果]

终端显示 {"models":[]} — Ollama正常运行，等待下载模型

第08章 下载AI大模型

Ollama已经准备好了，现在需要给它下载一个AI大模型。模型就是AI的"大脑"——下载不同的模型，AI就有不同的能力。

8.1 模型推荐表

小白推荐（上手快，要求低）

模型	大小	需要显存	特点	适合场景
Qwen3:8B	4.9GB	6GB	中文效果最好的小模型	日常聊天、翻译、写文案
DeepSeek-R1:8B	4.7GB	6GB	推理能力强，善于分析	数学、逻辑、分析思考

进阶推荐（更聪明，需要更多显存）

模型	大小	需要显存	特点	适合场景
Qwen3:14B	9GB	12GB	更聪明，理解力更强	进阶使用，复杂对话
Qwen3:32B	19GB	24GB	接近GPT-4水平	专业使用，高质量输出
DeepSeek-R1:32B	19GB	24GB	超强推理，思维链详尽	复杂任务、深度分析

8.2 GPU与模型匹配表

GPU型号	显存	推荐模型	可运行最大模型
RTX 3060	12GB	Qwen3:8B / R1:8B / Qwen3:14B	Qwen3:14B
RTX 3090	24GB	**Qwen3:32B / R1:32B**	Qwen3:32B / R1:32B
RTX 4090	24GB	**任意32B模型（速度更快）**	Qwen3:32B / R1:32B
A100 40GB	40GB	70B模型	Qwen3:72B
A100 80GB	80GB	70B+模型，多模型并行	任意开源模型

> **要点**
>
> **核心原则：模型需要的显存不能超过GPU显存。**
>
> 如果模型比显存大，会用内存来补充（offload），但速度会慢很多。所以尽量选显存够用的模型。

8.3 下载模型

在1Panel的终端中，或者在SSH终端中，使用以下命令下载模型：

下载 Qwen3:8B（推荐新手第一个模型）

```
ollama pull qwen3:8b
```

下载 DeepSeek-R1:8B

```
ollama pull deepseek-r1:8b
```

下载 Qwen3:32B（需要24GB显存）

```
ollama pull qwen3:32b
```

> **小白视角**
>
> `ollama pull` 就是"下载"命令。后面的 `qwen3:8b` 是模型名。冒号后面是版本/大小标识。
>
> 下载速度取决于网络，通常4.9GB的模型需要5~15分钟。下载期间不要关闭终端。

[截图：模型下载进度]

终端显示模型下载进度条和百分比，等待100%完成

8.4 验证模型

下载完成后，查看已安装的模型：

```
ollama list
```

[截图：已下载的模型列表]

终端显示已下载模型的名称、大小、修改时间等信息

测试一下模型是否能正常对话：

```
ollama run qwen3:8b "你好，请用一句话介绍自己"
```

✅ **预期结果**

如果AI回复了一段中文自我介绍，说明模型运行成功！🎉

第09章 安装OpenClaw智能体

OpenClaw是整个系统的"大脑皮层"——它让AI不仅能聊天，还能使用工具、操作文件、搜索信息……就像给AI装上了手和脚。

9.1 在1Panel中安装OpenClaw

1. 登录1Panel，在左侧菜单找到 **「应用商店」** 或 **「智能体」** 模块。
2. 搜索 **"OpenClaw"**，找到应用并点击 **「安装」**。

[截图：1Panel应用商店搜索OpenClaw]

在1Panel应用商店中搜索"OpenClaw"，找到后点击安装

9.2 配置模型账号

安装过程中或安装完成后，需要配置模型连接信息——告诉OpenClaw"AI大脑"在哪里。

3. 在OpenClaw的设置页面，找到 **「模型配置」**：

配置项	填写内容	解释
模型提供商	`Ollama`	告诉OpenClaw，你要用本地的Ollama来运行AI模型（而不是OpenAI等第三方服务）
API Key	`sk-local-ollama`	随便填就行！因为我们的Ollama是本地运行的，不需要真正的API密钥。填什么都行，但不能为空。
Base URL	`http://127.0.0.1:11434/v1`	Ollama的API地址。`127.0.0.1` = 本机（localhost），`11434` = Ollama端口，`/v1` = OpenAI兼容的API格式

💡 **小白视角**

Base URL 的含义拆解：

`http://` = 网络协议（"用HTTP方式通信"）

`127.0.0.1` = "本机地址"（就像说"我自己的电话号码"）

`:` = 端口分隔符（"门牌号和房间号之间用冒号隔开"）

`11434` = Ollama监听的端口（"11434号房间"）

`/v1` = API路径（"走v1版本的接口"）

连起来就是："在本机的11434端口上，用v1版本的API格式和Ollama对话"

[截图：OpenClaw模型配置页面]

OpenClaw设置 → 模型配置 → 填写Ollama提供商、API Key（sk-local-ollama）、Base URL（http://127.0.0.1:11434/v1）

9.3 创建OpenClaw实例

4 配置完成后，创建OpenClaw实例。在模型选择中，手动输入你下载的模型名称：

```
openclaw/qwen3:8b
```

💡 **格式说明**

格式是 `openclaw/模型名:版本` 。这里的 `openclaw/` 前缀是OpenClaw的命名规范，后面跟着Ollama中的模型名。

5 等待部署完成（首次启动可能需要1~2分钟），然后点击 **「WebUI」** 或对应的端口号链接。

[截图：OpenClaw实例列表]

OpenClaw实例运行中，点击WebUI端口（如3000）进入对话界面

9.4 首次使用配置

进入OpenClaw的WebUI后，你会看到一个聊天界面。首次使用时：

- 设置AI的名字和人格（比如叫它"小智"、"阿管家"等）
- 选择对话模型（选择你下载的qwen3:8b等）
- 发送第一条消息测试：**"你好！请用一句话介绍你能做什么"**

[截图：OpenClaw首次对话]

OpenClaw WebUI聊天界面，AI成功回复第一条消息 🎉

> ✅ **恭喜！核心系统部署完成**
>
> 如果你看到AI成功回复了你的消息，说明**核心系统已经部署完成！** 🎉🎉🎉
>
> 接下来是让它变得更强——安装MCP工具。

第10章 MCP工具生态（扩充版）

10.1 什么是MCP?

💡 小白视角

MCP = Model Context Protocol（模型上下文协议）

用一个比喻：**AI大脑就像一个人，MCP工具就像他的手机里安装的APP。**

没有APP，手机只能打电话发短信；装了APP，手机就能导航、点外卖、拍照、听音乐……

同理，没有MCP工具，AI只能聊天；装了MCP工具，AI就能搜索、读写文件、写代码、查数据库……

10.2 七大工具类别

🔍 1. 搜索类

🌐 Web Search（网络搜索）

用途：让AI能联网搜索最新信息

使用场景：查新闻、查资料、获取实时数据、验证事实

📰 RSS Feed（信息订阅）

用途：订阅和读取RSS信息源

使用场景：跟踪博客更新、新闻动态、技术资讯

📁 2. 文件类

📂 Filesystem（文件系统）

用途：读写服务器上的文件

使用场景：让AI帮你写配置文件、读日志、管理文档

📝 Notes / Obsidian（笔记）

用途：管理和搜索Markdown笔记

使用场景：知识管理、笔记整理、个人Wiki

💻 3. 开发类

🔨 GitHub（代码托管）

用途：管理GitHub仓库、提交代码

使用场景：让AI帮你创建仓库、提交PR、查看Issue、Code Review

🖥️ Code Interpreter（代码执行）

用途：在沙盒中执行代码

使用场景：让AI跑Python脚本、数据处理、生成图表

🗄️ 4. 数据库类

🐘 PostgreSQL

用途：查询和管理PostgreSQL数据库

使用场景：让AI帮你看数据、写SQL、生成报表

🗂️ SQLite

用途：操作轻量级SQLite数据库

使用场景：管理本地数据、快速查询、数据分析

📊 5. 生产力类

📧 Email（邮件）

用途：读取和发送邮件

使用场景：让AI帮你整理收件箱、草拟回复、总结重要邮件

📅 Calendar（日历）

用途：管理日程安排

使用场景：让AI帮你查看/创建日程、设置提醒、规划时间

⚙️ 6. 自动化类

🔁 n8n / Zapier（工作流自动化）

用途：创建自动化工作流

使用场景：自动化重复操作、集成多个服务、定时任务

📡 Webhook

用途：发送和接收HTTP请求

使用场景：与其他服务集成、自动化触发、数据推送

🛠️ 7. 实用工具类

🌤️ Weather（天气查询）

用途：获取天气信息

使用场景：问AI"明天穿什么"、"周末天气怎么样"

🗺️ Maps（地图导航）

用途：地图搜索和路线规划

使用场景：问AI"附近有什么好吃的"、"怎么去XX地方"

🔢 Calculator（计算器）

用途：精确数学计算

使用场景：复杂计算、数据统计、单位换算

10.3 推荐安装清单

⚡ 推荐优先安装的工具（按优先级排序）

1. 🔍 **网络搜索** — 最实用的工具，没有之一。让AI不再"断网"
2. 📁 **文件系统** — 让AI能帮你管理服务器上的文件
3. 🌤️ **天气查询** — 日常最常用的查询之一
4. 📧 **邮件** — 如果你有邮件管理需求
5. 💻 **GitHub** — 如果你做开发工作
6. 🗄️ **数据库** — 如果你有数据库需要管理
7. 🗓️ **日历** — 如果你需要AI帮你管理时间

🧠 进阶建议

MCP工具不需要一次全部安装。建议**先装网络搜索和文件系统**这两个最实用的，用熟悉了再根据需求添加其他工具。工具安装得太多反而会让AI变慢（因为需要选择使用哪个工具）。

第11章 手机接入：随时随地用AI

AI管家已经跑在你的服务器上了，但你不能只在电脑前才能用它。这一章教你如何从手机（以及任何设备）访问你的AI管家。

11.1 局域网访问 vs 外网访问

访问方式	条件	速度	适合场景
局域网	手机和服务器在同一网络	极快	办公室、家里（不实用，服务器在云端）
外网（直接IP）	直接用公网IP + 端口访问	快	临时使用，不够安全
Cloudflare Tunnel	通过Cloudflare隧道	快	**推荐！安全、稳定、免费**
ngrok	通过ngrok隧道	中等	快速测试用
Telegram Bot	通过Telegram聊天	快	**最方便的手机入口**

11.2 方案一：Cloudflare Tunnel（推荐）

> 💡 **小白视角**
>
> **Cloudflare Tunnel是什么？** 简单来说，它在你的服务器和互联网之间建立了一条"加密隧道"，让你的AI管家可以通过一个好看的域名访问（比如 `ai.yourname.com` ），而不用暴露服务器的真实IP地址。
>
> 就像给你的服务器装了一个"门牌翻译器"——别人只需要记住你的域名，不需要知道你的真实地址。

1. 注册 Cloudflare 账号，并将你的域名添加到 Cloudflare（如果你没有域名，可以在 Cloudflare 或其他注册商购买一个，通常几十元一年）。

2. 在服务器上安装 Cloudflare Tunnel 客户端（cloudflared）：

```
curl -L https://github.com/cloudflare/cloudflared/releases/latest/download/cloudflared-linux-amd64 -o cloudflared
chmod +x cloudflared
mv cloudflared /usr/local/bin/
```

3 登录并创建隧道：

```
cloudflared tunnel login
```

按提示在浏览器中授权，选择你的域名。

4 创建隧道并配置路由：

```
cloudflared tunnel create my-ai
cloudflared tunnel route dns my-ai ai.yourname.com
```

5 配置隧道指向 OpenClaw 和 Ollama 端口。创建配置文件：

```
cat > ~/.cloudflared/config.yml << 'EOF'
tunnel: my-ai
credentials-file: /root/.cloudflared/xxxx.json

ingress:
  - hostname: ai.yourname.com
    service: http://127.0.0.1:3000
  - hostname: api.yourname.com
    service: http://127.0.0.1:11434
  - service: http_status:404
EOF
```

6 启动隧道：

```
cloudflared tunnel run my-ai
```

建议用 systemd 设置为后台服务，开机自启动。

[截图：Cloudflare Tunnel 配置完成]

在浏览器中访问 ai.yourname.com，能看到OpenClaw的界面，说明隧道配置成功

✅ **预期效果**

配置完成后，你可以通过 `https://ai.yourname.com` 从手机、平板、任何设备访问你的AI管家。而且自带HTTPS加密，安全可靠。

11.3 方案二：ngrok（快速测试）

💡 **小白视角**

ngrok 比 Cloudflare Tunnel 更简单——一行命令就能把你的本地服务暴露到公网。但免费版每次启动地址会变，适合临时测试，不适合长期使用。

1. 下载安装 ngrok：

```
snap install ngrok
```

2. 注册 ngrok 账号，获取 authtoken：

```
ngrok config add-authtoken 你的token
```

3. 启动隧道：

```
ngrok http 3000
```

4. 终端会显示一个公网地址（如 `https://xxxx.ngrok-free.app`），在手机浏览器打开即可。

11.4 方案三：Telegram Bot（最方便）

💡 **小白视角**

Telegram Bot就是你在Telegram里的一个"AI聊天机器人"。你给它发消息，它用AI大模型回复你。这比打开浏览器访问网页方便多了——就像和朋友聊天一样自然。

1. 在Telegram中搜索 **@BotFather**，发送 `/newbot`，按提示创建一个Bot，获取Bot Token。

2. 在OpenClaw的配置中添加Telegram Bot集成，填入Bot Token。

3 配置完成后，在Telegram中找到你创建的Bot，直接发消息即可与AI对话。

[截图：Telegram Bot对话界面]

在Telegram中给Bot发消息，AI回复，和正常聊天一样流畅

✅ **最佳方案总结**

- **日常手机使用** → Telegram Bot（最方便，随时随地）
- **电脑浏览器访问** → Cloudflare Tunnel（安全、稳定、有自定义域名）
- **临时测试** → ngrok（最快上手，但不适合长期）

第12章 常见问题排雷

部署过程中难免会遇到各种问题。这一章汇总了最常见的坑和对应的解决方案。

❓ 连接不上服务器，SSH超时？

可能原因：

1. **实例没开机** — 检查优云智算控制台，确认实例状态为"运行中"
2. **安全组没开放22端口** — 检查安全组规则，确保22端口已放行
3. **IP或端口填错** — 仔细核对实例详情页上的公网IP和SSH端口
4. **密码错误** — 在控制台重置密码后重试

❓ 1Panel网页打不开？

排查步骤：

1. 确认1Panel服务正在运行：

```
systemctl status 1panel
```

2. 确认安全组/防火墙已开放1Panel端口
3. 确认浏览器中输入的地址和端口正确（注意HTTP不是HTTPS）
4. 尝试用 `http://IP:端口` 而不是 `https://IP:端口`

❓ Ollama启动失败，报错 "address already in use"？

说明11434端口被占用了。可能是之前的Ollama进程还在运行：

```
lsof -i :11434        # 查看谁占用了11434端口
kill -9 进程号        # 杀掉占用进程
systemctl restart ollama  # 重启Ollama
```

❓ Docker无法使用GPU，报错 "could not select device driver"?

解决步骤：

1. 确认nvidia-container-toolkit已安装：

```
which nvidia-ctk
```

2. 重新配置Docker运行时：

```
nvidia-ctk runtime configure --runtime=docker
systemctl restart docker
```

3. 如果还是不行，检查NVIDIA驱动：

```
nvidia-smi
```

   如果nvidia-smi也报错，说明GPU驱动有问题，可能需要重新选择镜像。

❓ 模型下载很慢怎么办?

加速方案：

1. 设置Ollama镜像源（国内加速）：

```
export OLLAMA_ORIGINS="*"
export OLLAMA_HOST=0.0.0.0
```

部分云平台（如优云智算）可能已经内置了国内镜像加速，下载速度会快很多。

如果仍然很慢，可以尝试在本地下载模型文件后上传到服务器。

❓ OpenClaw连接不上Ollama?

排查步骤：

1. 确认Ollama正在运行：

```
curl http://localhost:11434/api/tags
```

2. 确认Ollama监听地址是 `0.0.0.0`（不是 `127.0.0.1`）
3. 确认OpenClaw中的Base URL填写正确：`http://127.0.0.1:11434/v1`
4. 注意最后的 `/v1` 不能漏掉——这是OpenAI兼容接口的路径

❓ 基础镜像（容器类型）不能安装1Panel怎么办？

这是正常的——基础镜像创建的是容器实例，不支持systemd服务，因此无法安装1Panel。

解决方案：

1. 方案一：换用**系统镜像（虚机类型）**创建实例，然后安装1Panel + Docker + Ollama
2. 方案二：继续用基础镜像，跳过1Panel，所有操作在SSH终端中完成
3. 方案三：在基础镜像中手动安装Docker Compose，通过docker-compose管理服务

❓ 显存不够，大模型跑不动？

解决方案：

1. **换小模型** — 32B换14B，14B换8B
2. **开量化版本** — Ollama默认使用Q4量化，已经比较省显存了
3. **升级GPU** — 在优云智算中更换更大显存的GPU实例

> **经验法则**
> 模型需要的显存 ≈ 模型参数量(十亿) × 1.5GB。比如8B模型 ≈ 12GB显存（Q4量化后约6GB）。留出2~4GB余量给系统和Ollama。

❓ 不用的时候怎么省钱？

省钱技巧：

1. **关机** — 在优云智算控制台点击「关机」，关机后不计费（但可能收少量存储费）
2. **选择按量计费** — 不要选包月/包年，按小时计费最灵活
3. **用小模型** — 小模型可以用便宜的小显卡
4. **定时开关机** — 可以写个脚本，工作时间自动开机，下班自动关机

❓ 怎么更新Ollama和模型?

更新Ollama:

```
curl -fsSL https://ollama.com/install.sh | sh
```

更新模型:

```
ollama pull qwen3:8b  # 重新pull就会更新到最新版
```

⚠️ 踩坑通用建议

- 遇到问题先看**错误信息**——错误信息就是最好的线索
- 复制错误信息去搜索引擎搜——大概率有人遇到过同样的问题
- 修改配置后记得**重启服务**（systemctl restart xxx）
- 不确定命令会不会搞坏系统？先在测试环境试，或者做好快照备份
- 记录你做的每一步修改——这样出问题时可以回溯

第13章 完整部署清单

恭喜你读完了全部教程！以下是完整的部署清单，你可以逐项对照检查：

📋 基础设施

- [ ] 注册优云智算账号并充值
- [ ] 选择平台镜像 → 基础镜像 → Ollama v0.13.1
- [ ] 配置GPU型号（推荐RTX 3090）、GPU数量（1个）、CPU（16C 64G）、系统盘（200GB+）
- [ ] 选择防火墙预设「Ollama v0.13.1」
- [ ] 点击「立即部署」，等待实例创建
- [ ] 截图保存公网IP、SSH端口、用户名、密码

🛠️ 系统配置

- [ ] 通过SSH连接到Linux服务器
- [ ] 更新系统：`apt update && apt upgrade -y`
- [ ] 安装1Panel（如使用系统镜像）
- [ ] 截图保存1Panel登录信息（地址、用户名、密码）
- [ ] 在浏览器中登录1Panel
- [ ] 在安全组中开放端口（3000、1Panel端口等）
- [ ] 在1Panel防火墙中开放端口

🎮 GPU与Docker

- [ ] 安装 nvidia-container-toolkit（如使用系统镜像）

- [ ] 配置Docker运行时：`nvidia-ctk runtime configure --runtime=docker`
- [ ] 重启Docker：`systemctl restart docker`
- [ ] 验证GPU：`docker run --rm --gpus all nvidia/cuda nvidia-smi`

🦙 Ollama与模型

- [ ] 确认Ollama已安装（基础镜像已预装 / 手动安装）
- [ ] 配置Ollama监听地址：`OLLAMA_HOST=0.0.0.0`
- [ ] 重启Ollama：`systemctl restart ollama`
- [ ] 验证Ollama：`curl http://localhost:11434/api/tags`
- [ ] 下载AI大模型：`ollama pull qwen3:8b`
- [ ] 验证模型：`ollama list` 和 `ollama run qwen3:8b`

🤖 OpenClaw

- [ ] 在1Panel中安装OpenClaw
- [ ] 配置模型提供商：Ollama
- [ ] 配置API Key：sk-local-ollama
- [ ] 配置Base URL：http://127.0.0.1:11434/v1
- [ ] 创建OpenClaw实例，输入模型名
- [ ] 点击WebUI，测试对话

🔌 MCP工具（可选）

- [ ] 安装网络搜索工具
- [ ] 安装文件系统工具

- [ ] 根据需要安装其他工具

📱 手机接入（可选）

- [ ] 配置Cloudflare Tunnel / ngrok
- [ ] 或者配置Telegram Bot
- [ ] 从手机测试访问

🎯 部署完成后的日常使用

✅ 你现在拥有了

- ✅ 一台运行在云端的AI服务器，24/7在线
- ✅ 一个强大的AI大模型，能聊天、写文、分析、编程
- ✅ 一个可视化管理面板（1Panel），方便管理
- ✅ 一个智能体框架（OpenClaw），支持工具扩展
- ✅ 手机端随时访问的能力
- ✅ 完全私密的数据环境
- ✅ 按需付费，灵活省钱

⚡ 日常使用建议

1. **不用时关机** — 养成习惯，不使用时在控制台关机省钱
2. **定期更新** — Ollama和模型偶尔会有新版本，关注更新
3. **备份数据** — 重要对话和配置定期备份
4. **尝试不同模型** — 不同模型各有优势，多试试找到最适合你的
5. **逐步加工具** — 随着使用深入，按需添加MCP工具

部署完成！享受你的私人AI管家吧！

如果在部署过程中遇到问题，欢迎随时回来查阅本教程。

祝你用AI提升效率，享受科技带来的便利！

Linux GPU Ollama OpenClaw 1Panel MCP