# 《Deepseek R1 本地部署完全手册》


作者wechat：samirtan
版本：V2.0
更新日期：2025年2月8日

## 一、简介

Deepseek R1 是支持复杂推理、多模态处理、技术文档生成的高性能通用大语言模型。本手册为技术团队提供完整的本地部署指南，涵盖硬件配置、国产芯片适配、量化方案、云端替代方案及完整671B MoE模型的Ollama部署方法。
**核心提示**：

- **个人用户**：不建议部署32B及以上模型，硬件成本极高且运维复杂。
- **企业用户**：需专业团队支持，部署前需评估ROI（投资回报率）。

---

## 二、本地部署核心配置要求

### 1. 模型参数与硬件对应表

| 模型参数 | Windows 配置要求 | Mac 配置要求 | 适用场景 |
| --- | --- | --- | --- |
| **1.5B** | - RAM: 4GB<br>- GPU: 集成显卡/现代CPU<br>- 存储: 5GB | - 内存: 8GB（M1/M2/M3）<br>- 存储: 5GB | 简单文本生成、基础代码补全 |
| **7B** | - RAM: 8-10GB<br>- GPU: GTX 1680（4-bit量化）<br>- 存储: 8GB | - 内存: 16GB（M2 Pro/M3）<br>- 存储: 8GB | 中等复杂度问答、代码调试 |
| **14B** | - RAM: 24GB<br>- GPU: RTX 3090（24GB VRAM）<br>- 存储: 20GB | - 内存: 32GB（M3 Max）<br>- 存储: 20GB | 复杂推理、技术文档生成 |
| **32B+** | **企业级部署**（需多卡并联） | 暂不支持 | 科研计算、大规模数据处理 |

### 2. 算力需求分析

| 模型 | 参数规模 | 计算精度 | 最低显存需求 | 最低算力需求 |
| --- | --- | --- | --- | --- |
| DeepSeek-R1 (671B) | 671B | FP8 | ≥890GB | 2*XE9680（16*H20 GPU） |
| DeepSeek-R1-Distill-70B | 70B | BF16 | ≥180GB | 4*L20 或 2*H20 GPU |

---

## 三、国产芯片与硬件适配方案

### 1. 国内生态合作伙伴动态

| 企业 | 适配内容 | 性能对标（vs NVIDIA） |
| --- | --- | --- |
| **华为昇腾** | 昇腾910B原生支持R1全系列，提供端到端推理优化方案 | 等效A100（FP16） |
| **沐曦GPU** | MXN系列支持70B模型BF16推理，显存利用率提升30% | 等效RTX 3090 |
| **海光DCU** | 适配V3/R1模型，性能对标NVIDIA A100 | 等效A100（BF16） |

### 2. 国产硬件推荐配置

| 模型参数 | 推荐方案 | 适用场景 |
| --- | --- | --- |
| **1.5B** | 太初T100加速卡 | 个人开发者原型验证 |
| **14B** | 昆仑芯K200集群 | 企业级复杂任务推理 |
| **32B** | 壁彻算力平台+昇腾910B集群 | 科研计算与多模态处理 |

---

## 四、云端部署替代方案

### 1. 国内云服务商推荐

| 平台 | 核心优势 | 适用场景 |
| --- | --- | --- |
| | | |

| | | |
|---|---|---|
| **硅基流动** | 官方推荐API，低延迟，支持多模态模型 | 企业级高并发推理 |
| **腾讯云** | 一键部署+限时免费体验，支持VPC私有化 | 中小规模模型快速上线 |
| **PPIO派欧云** | 价格仅为OpenAI 1/20，注册赠5000万tokens | 低成本尝鲜与测试 |

**2. 国际接入渠道（需魔法或外企上网环境🚫）**

- **英伟达NIM**：企业级GPU集群部署（链接）
- **Groq**：超低延迟推理（链接）

---

## 五、完整671B MoE模型部署（Ollama+Unsloth）

**1. 量化方案与模型选择**

| 量化版本 | 文件体积 | 最低内存+显存需求 | 适用场景 |
|---|---|---|---|
| **DeepSeek-R1-UD-IQ1_M** | 158 GB | ≥200 GB | 消费级硬件（如Mac Studio） |
| **DeepSeek-R1-Q4_K_M** | 404 GB | ≥500 GB | 高性能服务器/云GPU |

**下载地址**：

- HuggingFace模型库
- Unsloth AI官方说明

**2. 硬件配置建议**

| 硬件类型 | 推荐配置 | 性能表现（短文本生成） |
|---|---|---|
| **消费级设备** | Mac Studio（192GB统一内存） | 10+ token/秒 |
| **高性能服务器** | 4×RTX 4090（96GB显存+384GB内存） | 7-8 token/秒（混合推理） |

**3. 部署步骤（Linux示例）**

1. **安装依赖工具**：

```
# 安装llama.cpp（用于合并分片文件）
/bin/bash -c "$(curl -fsSL
https://raw.githubusercontent.com/Homebrew/install/HEAD/install.sh)"
brew install llama.cpp
```

2. **下载并合并模型分片**：

```
llama-gguf-split --merge DeepSeek-R1-UD-IQ1_M-00001-of-00004.gguf
DeepSeek-R1-UD-IQ1_S.gguf
```

3. **安装Ollama**：

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

4. **创建Modelfile**：

```
FROM /path/to/DeepSeek-R1-UD-IQ1_M.gguf
PARAMETER num_gpu 28  # 每块RTX 4090加载7层（共4卡）
PARAMETER num_ctx 2048
PARAMETER temperature 0.6
TEMPLATE "<｜end_of_thinking｜>{{ .Prompt }}<｜end_of_thinking｜>"
```

5. **运行模型**：

```
ollama create DeepSeek-R1-UD-IQ1_M -f DeepSeekQ1_Modelfile
ollama run DeepSeek-R1-UD-IQ1_M --verbose
```

#### 4. 性能调优与测试

- **GPU利用率低**：升级高带宽内存（如DDR5 5600+）。
- **扩展交换空间**：

```
sudo fallocate -l 100G /swapfile
sudo chmod 600 /swapfile
sudo mkswap /swapfile
sudo swapon /swapfile
```

---

## 六、注意事项与风险提示

1. **成本警示**：
   - **70B模型**：需3张以上80G显存显卡（如RTX A6000），单卡用户不可行。
   - **671B模型**：需8xH100集群，仅限超算中心部署。
2. **替代方案**：
   - 个人用户推荐使用云端API（如硅基流动），免运维且合规。
3. **国产硬件兼容性**：需使用定制版框架（如昇腾CANN、沐曦MXMLLM）。

---

## 七、附录：技术支持与资源

- **华为昇腾**：昇腾云服务
- **沐曦GPU**：免费API体验
- **李锡涵博客**：完整部署教程

---

## 结语

Deepseek R1 的本地化部署需极高的硬件投入与技术门槛，**个人用户务必谨慎**，企业用户应充分评估需求与成本。通过国产化适配与云端服务，可显著降低风险并提升效率。技术无止境，理性规划方能降本增效！

---

**手册更新与反馈**：如有补充或修正，请联系文档作者，接入细节请阅读详细文档硅基流动社区。

**全球企业个人渠道附表**

1. 秘塔搜索：https://metaso.cn
2. 360纳米AI搜索：https://www.n.cn/
3. 硅基流动：https://cloud.siliconflow.cn/i/OBklluwO
4. 字节跳动火山引擎：https://console.volcengine.com/ark/region:ark+cn-beijing/experience
5. 百度云千帆：https://console.bce.baidu.com/qianfan/modelcenter/model/buildIn/list
6. 英伟达NIM：https://build.nvidia.com/deepseek-ai/deepseek-r1
7. Groq：https://groq.com/
8. Fireworks：https://fireworks.ai/models/fireworks/deepseek-r1
9. Chutes：https://chutes.ai/app/chute/
10. Github：https://github.com/marketplace/models/azureml-deepseek/DeepSeek-R1/playground🚫
11. POE：https://poe.com/DeepSeek-R1🚫
12. Cursor：https://cursor.sh/🚫
13. Monica：https://monica.im/invitation?c=ACZ7WJJ9🚫
14. Lambda：https://lambdalabs.com/🚫
15. Cerebras：https://cerebras.ai🚫
16. Perplexity：https://www.perplexity.ai🚫
17. 阿里云百炼：https://api.together.ai/playground/chat/deepseek-ai/DeepSeek-R1

🚫 为需要魔法或外企上网环境

## 芯片企业支持附图

| 支持DeepSeek模型的国内AI芯片企业动态（智东西制表） | | |
|---|---|---|
| 日期 | 企业 | 官宣标题 |
| 2月1日 | 华为 | 首发！硅基流动x华为云联合推出基于昇腾云的DeepSeek R1&V3推理服务！ |
| 2月1日 | 沐曦 | Gitee AI联合沐曦首发全套DeepSeek R1千问蒸馏模型，全免费体验！ |
| 2月4日 | 天数智芯 | 一天适配！天数智芯联合Gitee AI正式上线DeepSeek R1模型服务 |
| 2月4日 | 摩尔线程 | 致敬DeepSeek：以国产GPU为基，燎原中国AI生态之火 |
| 2月4日 | 海光信息 | DeepSeek V3和R1模型完成海光DCU适配并正式上线 |
| 2月4日 | 华为 | 昇腾原生：潞晨科技推出基于昇腾算力的DeepSeek R1系列推理API及云镜像服务 |
| 2月5日 | 沐曦 | DeepSeek-V3满血版在国产沐曦GPU首发体验上线 |
| 2月5日 | 华为 | 昇腾蛇年开工送大礼，DeepSeek系列新模型正式上线昇腾社区 |
| 2月5日 | 海光信息 | 海光DCU成功适配DeepSeek-Janus-Pro多模态大模型 |
| 2月5日 | 壁仞科技 | DeepSeek R1在壁仞国产AI算力平台发布，全系列模型一站式赋能开发者创新 |
| 2月5日 | 太初元碁 | 基于太初T100加速卡2小时适配DeepSeek-R1系列模型，一键体验，免费API服务 |
| 2月5日 | 云天励飞 | DeepEdge10已完成DeepSeek R1系列模型适配 |
| 2月6日 | 燧原科技 | 燧原科技实现全国各地智算中心DeepSeek的全量推理服务部署 |
| 2月6日 | 昆仑芯 | 国产AI卡Deepseek训练推理全版本适配、性能卓越，一键部署等您来（附文档下载方式） |

## 云厂商智算企业支持附图

| 官宣支持DeepSeek模型的国内云服务及智算企业动态（智东西制表） | | |
|---|---|---|
| 日期 | 企业 | 官宣标题 |
| 1月28日 | 无问芯穹 | 无问芯穹Infini-AI异构云现已上架DeepSeek-R1-Distill，国产模型与异构云的绝妙组合 |
| 1月28日 | PPIO派欧云 | 重磅！DeepSeek-R1上线PPIO派欧算力云 |
| 1月28日 | 硅基流动 | SiliconCloud上线DeepSeek多模态模型：Janus-Pro-7B来了 |
| 2月1日 | 华为云 | 首发！硅基流动x华为云联合推出基于昇腾云的DeepSeek R1&V3推理服务！ |
| 2月1日 | 硅基流动 | 首发！硅基流动x华为云联合推出基于昇腾云的DeepSeek R1&V3推理服务！ |
| 2月1日 | 天翼云 | 神秘“东方力量”集结！DeepSeek-R1模型在天翼云上架！ |

| | | |
|---|---|---|
| 2月2日 | 腾讯云 | 一键部署，3分钟调用！DeepSeek-R1登陆腾讯云 |
| 2月2日 | 云轴科技 | 首发！ZStack智塔支持DeepSeek V3/R1/ Janus Pro，多种国产CPU/GPU可私有化部署 |
| 2月2日 | PPIO派欧云 | PPIO派欧算力云接入DeepSeek全模型，价格仅OpenAI o1 1/20，注册即送5000万tokens！ |
| 2月3日 | 阿里云 | 3步，0代码！一键部署DeepSeek-V3、DeepSeek-R1 |
| 2月3日 | 百度智能云 | 百度智能云千帆全面支持DeepSeek-R1/V3调用，价格超低 |
| 2月3日 | 超算互联网 | 超算互联网上线DeepSeek系列模型，提供超智融合算力支持 |
| 2月4日 | 腾讯云 | 一键部署+限免体验！腾讯云上架DeepSeek系列模型 |
| 2月4日 | 硅基流动 | 全家桶来了！硅基流动上线加速版DeepSeek-R1蒸馏模型 |
| 2月4日 | 火山引擎 | 全尺寸DeepSeek模型登陆火山引擎！ |
| 2月4日 | 青云科技 | 限时免费，一键部署！基石智算正式上线DeepSeek-R1系列模型 |
| 2月4日 | 算力互联 | 国产GPU与DeepSeek加速适配，算力互联携手天数智芯推出DeepSeek-R1模型服务 |
| 2月4日 | 京东云 | 一键部署！京东云全面上线DeepSeek-R1/V3 |
| 2月4日 | 超算互联网 | 再上新｜来超算互联网DeepSeek一下！ |
| 2月5日 | 联通云 | “哪吒闹海”！联通云上架DeepSeek-R1系列模型！ |
| 2月5日 | PPIO派欧云 | PPIO假期战报：99.9%可用性！连夜支持满血版DeepSeek，助力客户轻松应对流量高峰 |
| 2月5日 | 并济科技 | 并济科技携手燧原科技完成DeepSeek全量推理服务国产化部署，加速智算中心生态建设 |
| 2月5日 | 优刻得 | 优刻得基于国产芯片适配DeepSeek全系列模型 |
| 2月5日 | 移动云 | 全版本、全尺寸、全功能！移动云全面上线DeepSeek |
| 2月6日 | 青云科技 | 持续上线DeepSeek！基石智算Janus-Pro-7B文生图模型来了 |
| 2月6日 | 神州数码 | 3分钟部署高性能AI模型DeepSeek，神州数码助力企业智能化转型 |
| 2月6日 | 天翼云 | 国产AI生态新突破！“息壤”+DeepSeek王炸组合来了！ |
| 2月6日 | 并行科技 | 服务器繁忙？并行科技助您DeepSeek自由！ |
| 2月6日 | 优刻得 | 优刻得私有云上线DeepSeek系列模型 |
| 2月7日 | 浪潮云 | 浪潮云率先发布671B DeepSeek大模型一体机解决方案 |
| 2月7日 | 北京超算 | 北京超算×DeepSeek：双擎爆燃，驱动千亿级AI创新风暴 |

注：厂商支持图表版权归智东西