ollama 常用命令和操作

以下是 **Ollama** 的常用命令和操作指南，涵盖模型管理、运行配置、API 使用等核心功能，适用于本地大模型开发和测试：

1. 模型管理

下载模型

```
ollama pull <模型名>[:版本] # 默认下载最新版
```

常用模型示例:

```
ollama pull llama2         # Llama 2 7B
ollama pull phi            # Microsoft Phi-2 (2.7B)
ollama pull mistral        # Mistral 7B
ollama pull qwen:7b        # 阿里千问中文模型
```

查看已安装模型

```
ollama list
```

输出示例：

```
NAME          ID            SIZE    MODIFIED
llama2:latest  e6a7b3b4d5e6  3.8 GB  2 days ago
phi:latest     f1a2b3c4d5e6  1.9 GB  5 hours ago
```

删除模型

```
ollama rm <模型名>
```

2. 运行模型

交互式对话

```
ollama run <模型名> "你的提示词"
```

示例：

```
ollama run llama2 "用 Python实现快速排序"
```

持续对话

直接运行不输入提示词，进入多轮对话模式：

```
ollama run phi
```

输入 `/bye` 退出对话。

带参数运行

```
ollama run --temperature 0.7 --num_ctx 2048 mistral
```

常用参数：

- `--temperature`：控制随机性（0-1，越高越有创意）
- `--num_ctx`：上下文长度（默认 2048）
- `--seed`：设置随机种子（用于复现结果）

---

3. 高级配置

启动本地 **API**服务

```
ollama serve
```

默认监听 `127.0.0.1:11434`，可通过 HTTP调用：

```
curl http://localhost:11434/api/generate -d '{
  "model": "llama2",
  "prompt": "为什么天空是蓝色的?"
}'
```

GPU加速

```
ollama run --gpu llama2  # 启用 NVIDIA GPU
```

需确保:

1. 已安装 NVIDIA驱动和 CUDA
2. Ollama版本 ≥0.1.20

量化模型运行

```
ollama pull llama2:7b-q4_0  # 下载 4-bit量化版
ollama run llama2:7b-q4_0   # 显存占用降低 50%
```

---

4. 自定义模型

从 **Modelfile**创建

1. 创建 `Modelfile`:
   ```
   FROM llama2
   PARAMETER temperature 0.8
   SYSTEM """
   你是一个专业的技术助手，回答需简洁准确。
   """
   ```
2. 构建自定义模型:
   ```
   ollama create my-llama -f Modelfile
   ```
3. 运行:
   ```
   ollama run my-llama
   ```

导入 /导出模型

```
ollama export llama2 llama2.tar  # 导出
ollama import llama2.tar         # 导入
```

---

5. 实用技巧

查看模型信息

```
ollama show <模型名> --modelfile  # 显示模型配置
ollama show --license llama2  # 查看许可证
```

多模型并行

```
# 终端 1
ollama serve

# 终端 2
ollama run phi "解释量子力学"

# 终端 3
ollama run mistral "写一首英文诗"
```

日志调试

```
ollama serve > ollama.log 2>&1  # 输出日志到文件
```

---

6. 常见问题解决

内存不足

- 方案 **1**：使用量化模型（如 `q4` 版本）
- 方案 **2**：限制 CPU线程：

  ```
  OLLAMA_NUM_THREADS=4 ollama run llama2
  ```

中文输出不佳

1. 使用支持中文的模型：

   ```
   ollama pull qwen:7b
   ```

2. 明确提示词：

   ```
   ollama run qwen "用中文回答：如何学习机器学习？"
   ```

端口冲突

修改默认端口：

```
OLLAMA_HOST=0.0.0.0:12345 ollama serve
```

---

7. 常用模型推荐

模型名称	参数量	特点	适用场景
Llama 2	7B/13B	通用性强，英文优	文本生成、问答
Phi-2	2.7B	逻辑推理强，代码生成好	技术问答、代码辅助
Mistral	7B	性能接近 Llama 13B	长文本处理
Qwen	7B	中文支持优秀	中文对话、翻译
Gemma	2B/7B	Google轻量级模型	移动端部署

通过以上命令，您可以高效管理本地大语言模型。如需扩展功能（如接入 LangChain或 Web界面），可结合 Ollama的 API进一步开发。

Revision #1
Created 9 April 2025 09:03:12 by Admin
Updated 9 April 2025 09:04:05 by Admin