# 大模型本地部署使用手册

## **目录**

1. 硬件要求
2. 环境准备
3. 模型下载
4. 部署步骤

## **硬件要求**

### **最低配置**

* ****CPU****: Intel i7 或同等性能的AMD处理器(10代以上)

****内存****: 32GB DDR4

****GPU****: NVIDIA RTX 3090 (24GB显存)

****存储****: 至少100GB可用SSD空间

### **推荐配置**

****CPU****: Intel Xeon 或 AMD EPYC 系列

* ****内存****: 64GB以上

****GPU****: NVIDIA A100 40GB/80GB 或多卡配置

* ****存储****: NVMe SSD 1TB以上
* **环境准备**

### **操作系统**

Ubuntu 20.04/22.04 LTS

CentOS 7/8 (需额外配置)

Windows WSL2 (仅限开发测试)

### **软件依赖**

1. ****Python****: 3.8-3.10
2. sudo apt updatesudo apt install python3.8 python3.8-venv
3. ****CUDA****: 11.7或更高

wget https://developer.download.nvidia.com/compute/cuda/repos/ubuntu2204/x86\_64/cuda-ubuntu2204.pinsudo mv cuda-ubuntu2204.pin /etc/apt/preferences.d/cuda-repository-pin-600sudo apt-key adv --fetch-keys https://developer.download.nvidia.com/compute/cuda/repos/ubuntu2204/x86\_64/3bf863cc.pubsudo add-apt-repository "deb https://developer.download.nvidia.com/compute/cuda/repos/ubuntu2204/x86\_64/ /"sudo apt-get updatesudo apt-get -y install cuda

1. pip3 install torch torchvision torchaudio --extra-index-url https://download.pytorch.org/whl/cu117
2. ****其他依赖****:
3. pip install transformers accelerate sentencepiece protobuf

**模型下载**

### **官方渠道**

1. Hugging Face Hub:

from transformers import AutoModelForCausalLM, AutoTokenizer

model\_name = "meta-llama/Llama-2-7b-chat-hf"

tokenizer = AutoTokenizer.from\_pretrained(model\_name)

model = AutoModelForCausalLM.from\_pretrained(model\_name)

手动下载:

git lfs installgit clone https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

1. **模型量化(减少显存占用)**

python

from transformers import BitsAndBytesConfig

quantization\_config = BitsAndBytesConfig(

 load\_in\_4bit=True,

 bnb\_4bit\_compute\_dtype=torch.float16,

 bnb\_4bit\_quant\_type="nf4",

 bnb\_4bit\_use\_double\_quant=True,)

model = AutoModelForCausalLM.from\_pretrained(

 model\_name,

 quantization\_config=quantization\_config,

 device\_map="auto")

## **部署步骤**

### **基础部署**

1. 创建Python虚拟环境:

 python3 -m venv llm-envsource llm-env/bin/activate

安装依赖:

pip install -r requirements.txt

from transformers import pipeline

pipe = pipeline("text-generation", model="your/local/model/path")

result = pipe("你好，大模型!")print(result)