

AutoDL使用

网址: <http://autodl.com>

The screenshot shows the main landing page of the AutoDL website. At the top, there's a banner with the text "已恢复4090工作站整机售卖服务; 西北B区550+4090近日陆续上线". Below the banner, there's a large central image of a cloud with binary code floating around it. To the right of the image is a circular progress bar showing "49%" and some other metrics. The page features several navigation links at the top: AutoDL, 算力市场, AI服务器, 算法社区, 私有云, 帮助文档, 更多. On the right side, there's a "控制台" (Control Panel) and a dropdown menu for "炼丹师4540". The main content area includes a section titled "面向AIGC的解决方案" with a "弹性部署震撼上线" subtitle, a "了解详情" button, and a "注册礼包" section with "注册立送30天会员". There are also sections for "GPU选型" (How to choose the right GPU), "开具发票" (Issue an invoice), and "新手入门" (Newbie guide).

按需使用

The screenshot shows the "算力市场" (Power Market) section of the AutoDL website. It displays two main GPU rental options: "RTX 3060" and "RTX A4000". For the RTX 3060, it shows "西北B区 / 360机" (Northwest B Zone / 360 units) and "可租用至: 2027-01-01". The RTX 3060 details include: GPU数量(卡): 1 / 8, CPU: 7 核/GPU, Xeon(R) E5-2680 v4, 内存: 15 GB/GPU, 显存: 12 GB, 系统盘: 30 GB, 数据盘: 免费 60 GB, 可扩容 150 GB, 支持最高CUDA版本: 12.2. The price is listed as "¥0.60/时 · ¥0.63/时" with a "9.5折" discount. For the RTX A4000, it shows "西北B区 / 358机" (Northwest B Zone / 358 units) and "可租用至: 2027-01-01". The RTX A4000 details include: GPU数量(卡): 1 / 8, CPU: 16 核/GPU, Xeon(R) E5-2690 v4, 内存: 30 GB/GPU, 显存: 24 GB, 系统盘: 30 GB, 数据盘: 免费 60 GB, 可扩容 150 GB, 支持最高CUDA版本: 12.2. The price is listed as "¥0.60/时 · ¥0.63/时" with a "9.5折" discount. The page also features a sidebar with a "领优惠券" (Get a coupon) button and a progress bar showing "49%".

留意显存够不够用

The screenshot shows the AutoDL market page. At the top, there are tabs for GitHub - THUD, 智谱AI, 智谱AI开放平台, 智谱AI开放平台, 智谱AI开放平台, AutoDL算力云, AutoDL帮助文档, and initial commit. The main navigation bar includes AutoDL, 算力市场, AI服务器, 算法社区, 私有云, 帮助文档, and 更多. On the right, it shows 控制台 and 炼丹师4540. Below the navigation, there are dropdown menus for 选择地区 (Northwest A Zone), GPU型号 (All RTX 4090), and GPU数量 (1). Two GPU options are listed:

GPU	Region	可用机	可用至
RTX 4090	西北A区	031机	2024-01-10
RTX 4090	西北A区	008机	2024-01-10

Each listing provides detailed specifications: GPU数量(卡): 2 / 8, CPU: 15 核/GPU, Xeon(R) Platinum 8375C, 显存: 24 GB, 系统盘: 30 GB, 数据盘: 免费 50 GB, 支持最高CUDA版本: 12.0, 内存: 80 GB/GPU, 浮点算力: 单精度 82.58 TFLOPS / 半精度 165.2 Tensor TFLOPS. Pricing is shown as ¥2.68/时 (¥2.68 per hour).

选择基础镜像

The screenshot shows the AutoDL instance creation page. At the top, there are tabs for GitHub - THUD, 智谱AI, 智谱AI开放平台, 智谱AI开放平台, 智谱AI开放平台, AutoDL算力云, AutoDL帮助文档, and initial commit. The main navigation bar includes AutoDL, 算力市场, AI服务器, 算法社区, 私有云, 帮助文档, and 更多. On the right, it shows 控制台 and 炼丹师4540. The left sidebar shows GPU数量: 2, 数据盘: 30 GB, 实例规格: 高配版, and 镜像: PyTorch. A dropdown menu for PyTorch shows versions: 1.1.0, 1.5.1, 1.6.0, 1.7.0, 1.8.1, 1.9.0, 1.10.0, 1.11.0, and 2.0.0. The Python版本 is set to 3.8/ubuntu20.04 and the Cuda版本 is 11.8. A message at the bottom says "请选择框架名称/框架版本/Python版本/Cuda版本". Below the dropdown, there is a note: "创建完成后仍然可以更换其他镜像". At the bottom, there is a coupon dropdown labeled "优惠券: 请选择", and a summary section: "日常费用: ¥0.00/日" (¥0.00 per day), "配置费用: ¥2.68/时 费用明细" (Configuration cost: ¥2.68 per hour, Cost details), "账户余额: ¥2.80", and a "立即创建" (Create Now) button.

使用提出的登录指令和密码就能登陆服务器了

The screenshot shows the AutoDL instance list page. There are four instances listed:

- 西北A区 / 014机: RTX 4090 * 1卡, 状态: 运行中, 健康状态: 正常, 付费方式: 按量计费, 释放时间: 15天后释放, 登录指令: ssh*****. 实例ID: 8ae311943c-aad13247.
- 西北B区 / 142机: RTX 3080 Ti * 1卡, 状态: 已关机, 健康状态: 正常, 付费方式: 按量计费, 释放时间: 23小时后释放, 登录指令: ssh*****. 实例ID: ad3a418634-cf36e71e.
- 西北B区 / 026机: RTX 4090 * 1卡, 状态: 已关机, 健康状态: 正常, 付费方式: 按量计费, 释放时间: 14天14小时后释放, 登录指令: ssh*****. 实例ID: e92446ba49-5138036b.
- 西北B区 / 291机: RTX 3080 Ti * 1卡, 状态: 已关机, 健康状态: 正常, 付费方式: 按量计费, 释放时间: 14天09小时后释放, 登录指令: ssh*****. 实例ID: 794d4f961c-59cb97cd.

A large blue watermark in the center of the page says "然后现在搞好了以后我接下来干嘛".

The screenshot shows the same instance list page as above, but with an OpenSSH SSH client window overlaid. The terminal shows the following command being run:

```
root@autodl-container-8ae311943c-aad13247:~/glm3# ll
```

The terminal window has a black background and white text. The rest of the page is visible behind it.

因为是国内的机器，访问git和hugging face会非常慢
可以使用学术资源加速

The screenshot shows a browser window with multiple tabs open, all related to AutoDL. The main content is the '学术资源加速' (Academic Resource Acceleration) section of the AutoDL help documentation. It includes a statement about academic use, a list of accelerated sites (github.com, githubusercontent.com, githubassets.com, huggingface.co), usage instructions for the terminal and Jupyter Notebook, and a note about setting up across different regions. A large watermark in the center reads '需要到Hugging face上面去获取东西' (You need to get it from Hugging Face). On the right side, there are links for '目录' (Table of Contents), '使用方法' (Usage Methods), and '速度对比' (Speed Comparison). A circular progress bar at the bottom right indicates 49% completion.

接下来搭建一个ChatGLM3 (备注：目前主要在用的已经是ChatGLM4了)
先git clone

The screenshot shows a terminal window titled 'OpenSSH SSH client' running on an AutoDL container. The user is cloning the 'THUDM/ChatGLM3' repository from GitHub. The command used is 'git clone https://github.com/THUDM/ChatGLM3.git'. The terminal output shows the cloning process, including object enumeration, counting, compressing, and receiving. A large watermark in the center of the terminal window reads '我认为应该是几秒钟就搞完了' (I think it should be done in a few seconds). The GitHub repository page for 'THUDM/ChatGLM3' is visible in the background, showing it's a public repository with 528 forks and 5.5k stars. A circular progress bar at the bottom right indicates 49% completion.

修改一下 requirements.txt (调试的时候总是有一个包冲突，去掉发生冲突的那个包的版本，让 pip install 时自动选择版本)

我们打开给大家看一下

```
pip install -r requirements.txt
```

应该几秒钟

```
pip install -r requirements.txt
```

用chatglm3给的demo code测试一下

