国家超算互联网平台介绍

“国家超算互联网”由科技部指导发起，致力于链接我国算力产业上下游及供需双方资源，实现超算、智算等全国算力资源的统筹与调度，打造集算力、应用、数据、生态、社区等于一体的开放共享平台，让国产算力更加普惠易用，助力科技创新和数字经济高质量发展。

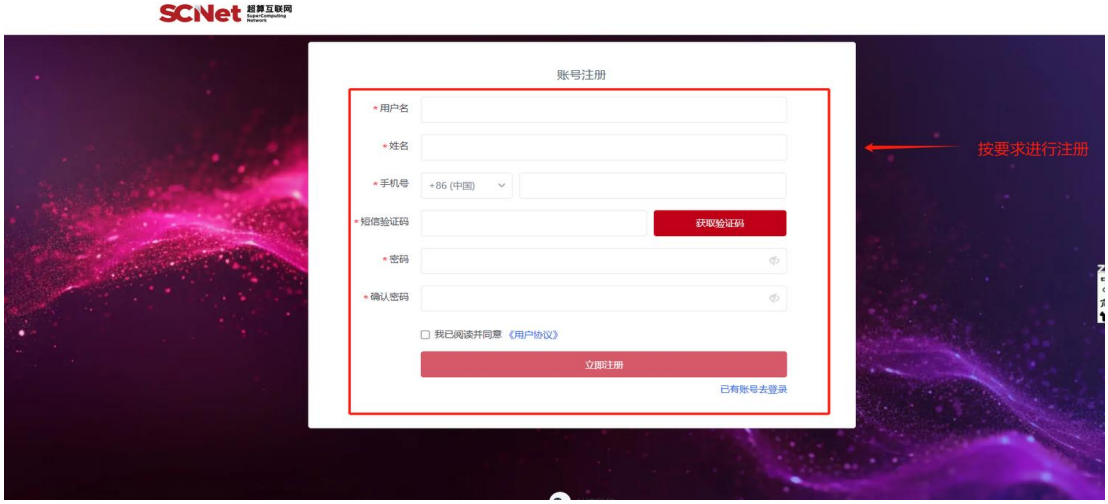
2024 年 4 月 11 日，首届超算互联网峰会暨国家超算互联网平台上线仪式在天津顺利举办，来自部委、省级科技厅、中国科学院、中国工程院、计算产业链相关企业等专家、代表数百人共聚一堂，见证了这一历史性时刻。

如何使用 Star 平台训练模型

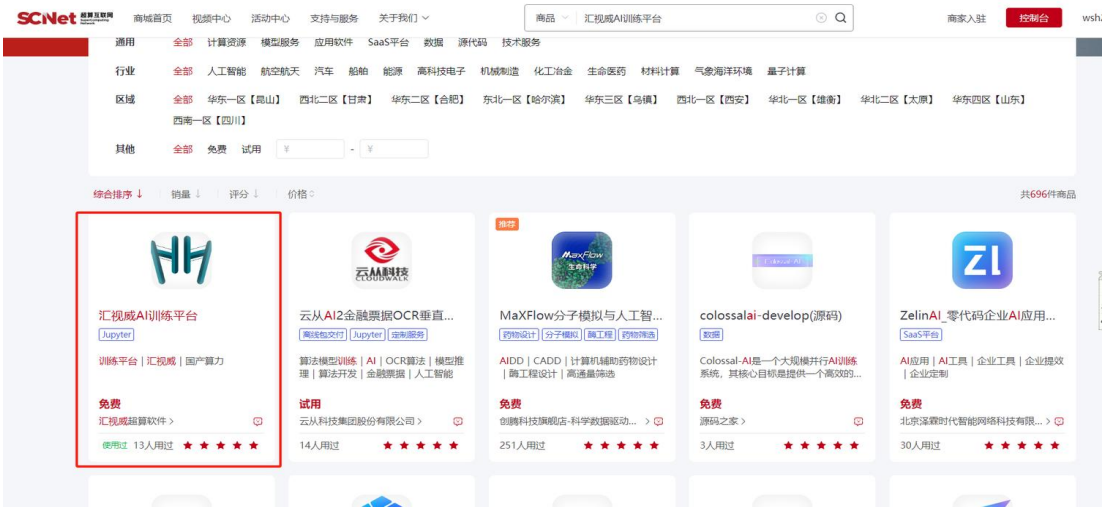
Star 平台是面向机器学习深度学习开发者，支持提供【模型微调】和【训练管理】等丰富建模工具、多框架高性能模型开发平台，数据托管、代码开发、模型训练等功能。

1. 注册并登录超算互联网平台

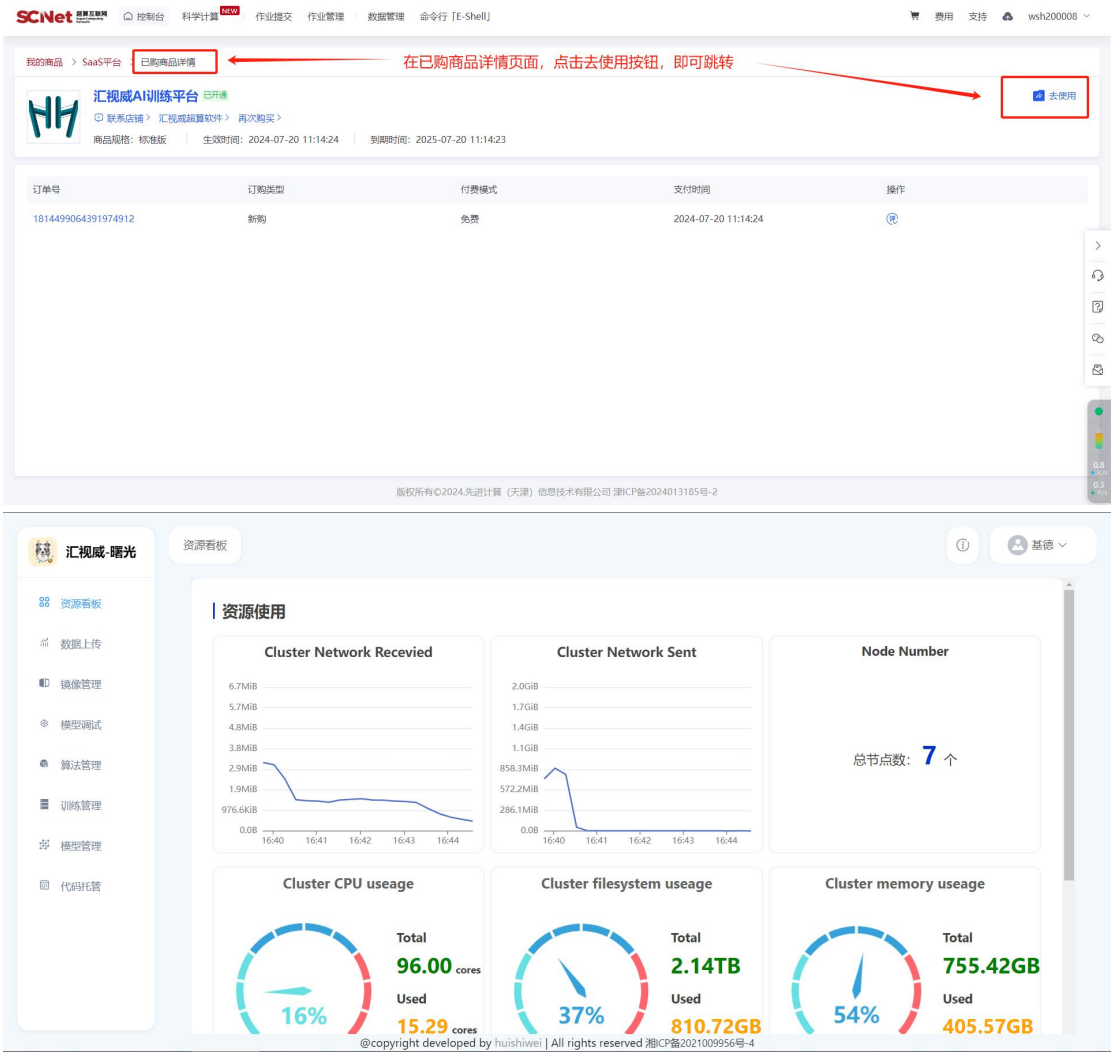
1. 打开您的浏览器，访问超算互联网平台：https://www.scnet.cn/。
2. 点击【登录/注册】按钮，按照提示填写相关信息以注册账户，如果已经有账户可以直接输入用户名密码登录。



3. 搜索“汇视威 AI 训练平台”，并点击进行使用

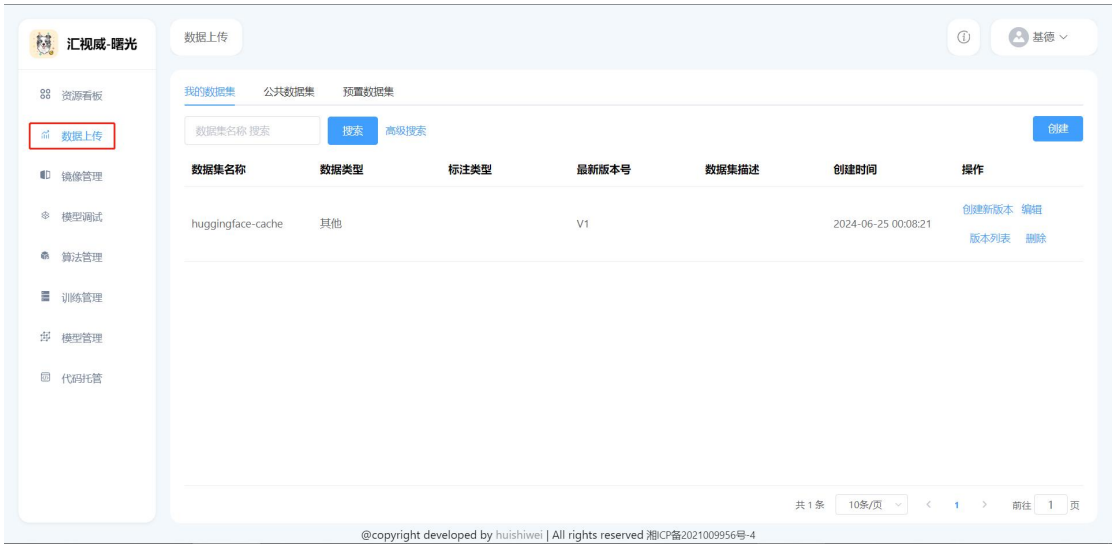


4. 在【已购商品详情页面】点击【去使用】按钮进入 Star 训练平台主页。



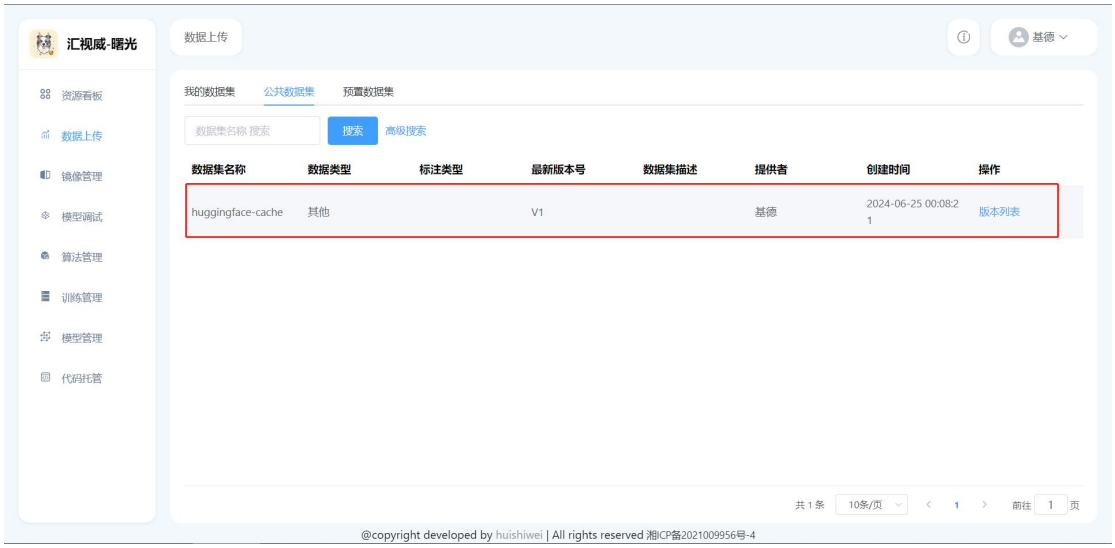
2. 准备数据集

1. 在平台界面左侧点击【数据上传】按钮可以跳转到数据模块。



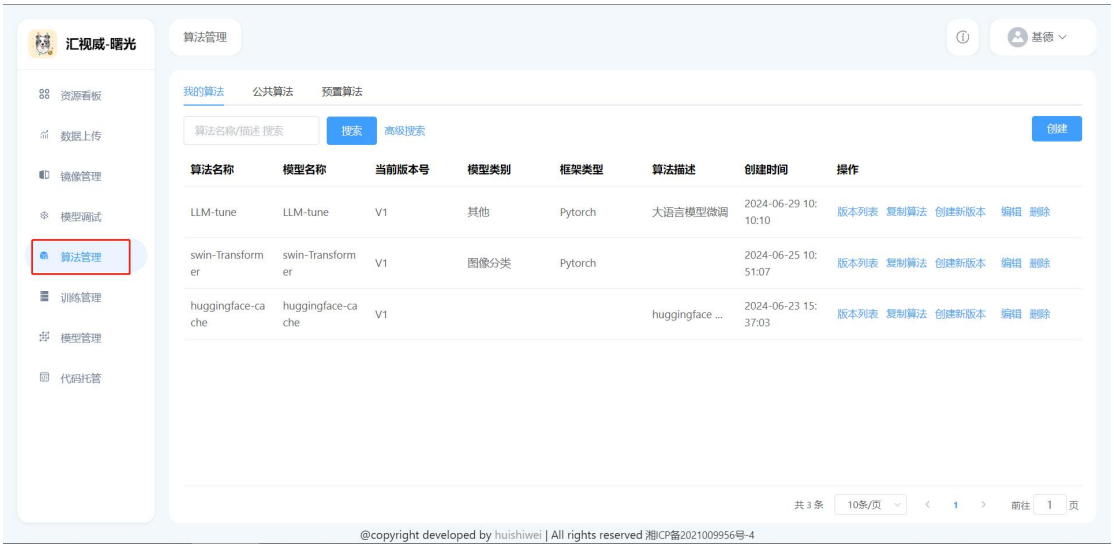
2. 在数据集页面上，点击【创建】按钮，可进入上传数据集页面，此时选择本地的 zip 格式文件进行上传，上传后平台会自动解压文件。

3. 数据集上传后会自动以 zip 文件名来命名并且作为数据集的 V1 版本，平台上可以根据改数据集名称来使用。
4. 对于微调大语言模型任务，平台已内置部分模型和数据集，数据集名称如下

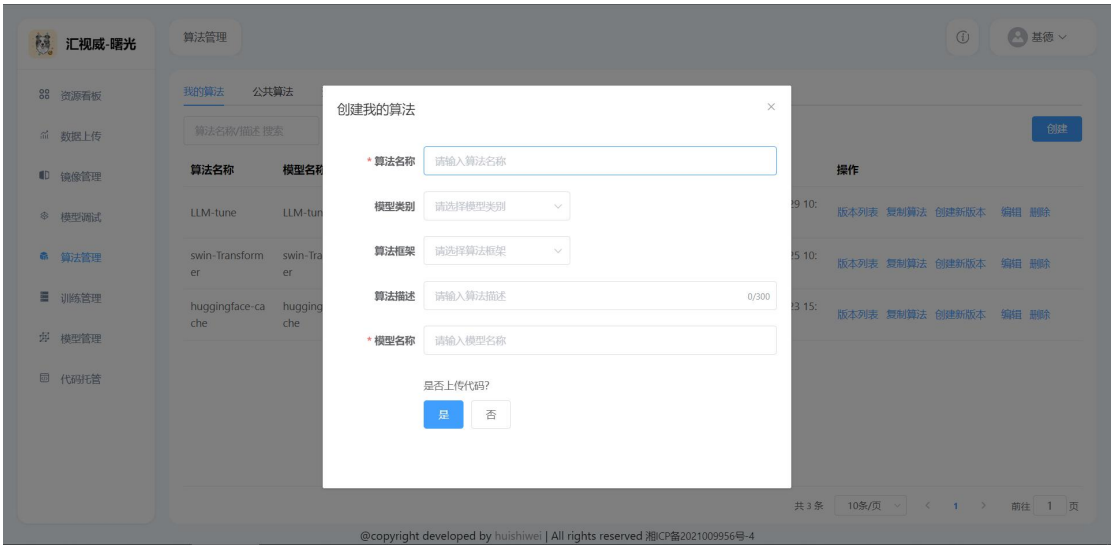


3. 选择合适的算法

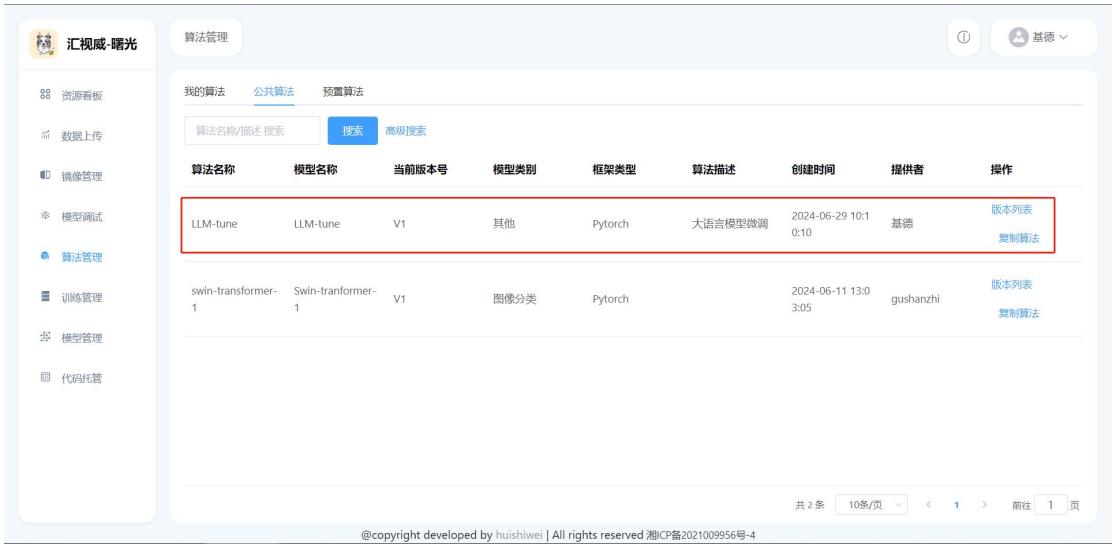
1. 根据您的任务类型和数据特点，从平台上选择合适的机器学习算法。平台左侧点击【算法管理】按钮可进入该模块。



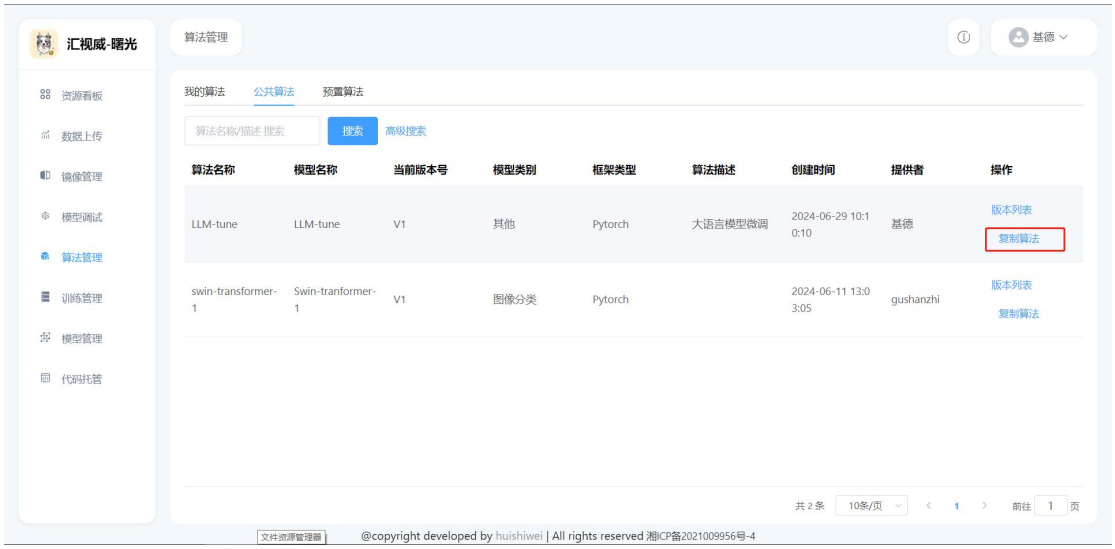
2. 点击创建可以进入算法创建页面，填写算法名称以及模型名称，比如“LLM-tune”；这里可以选择是否上传代码

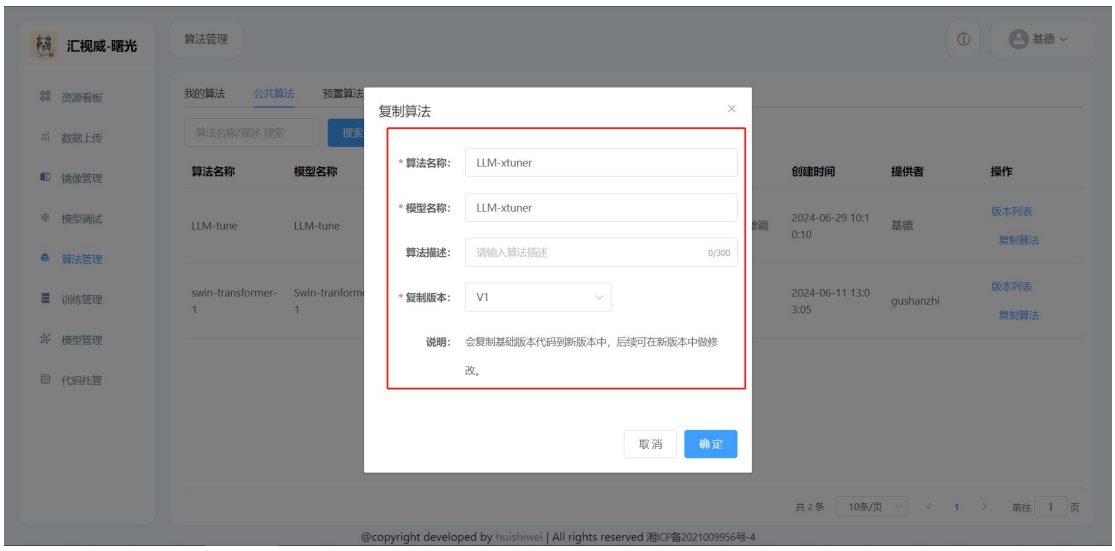


3. 对于微调大语言模型任务，大家可以选择不上传代码，之后在【模型调试】模块中编写或者上传，或者使用平台内置的【公共算法】【LLM-tune】，这里包括了本次实践课程中需要使用的代码。



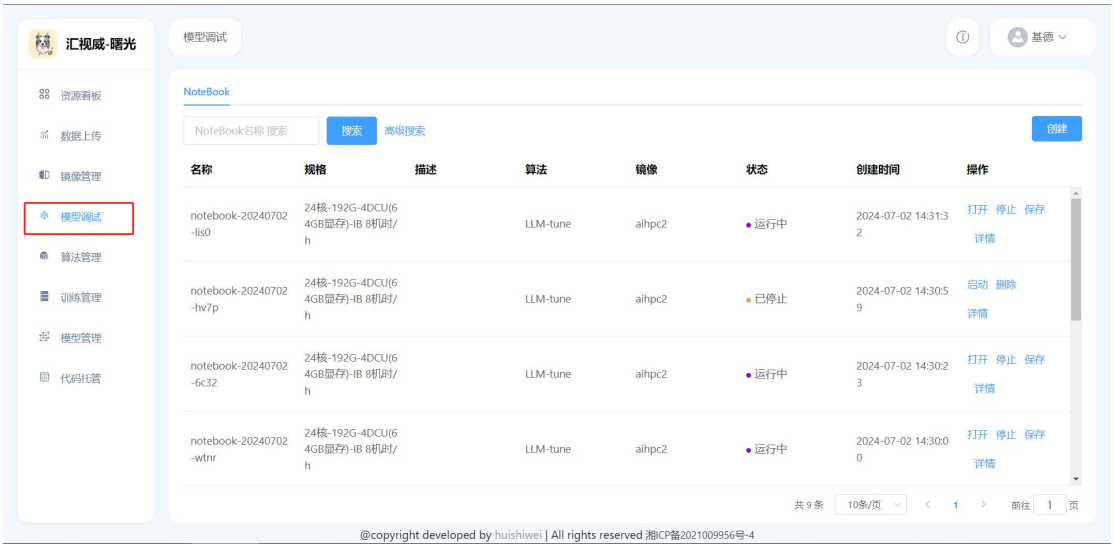
4. 使用平台内置的【公共算法】需要先复制一份到自己的空间。点击【复制算法】进行代码复制。



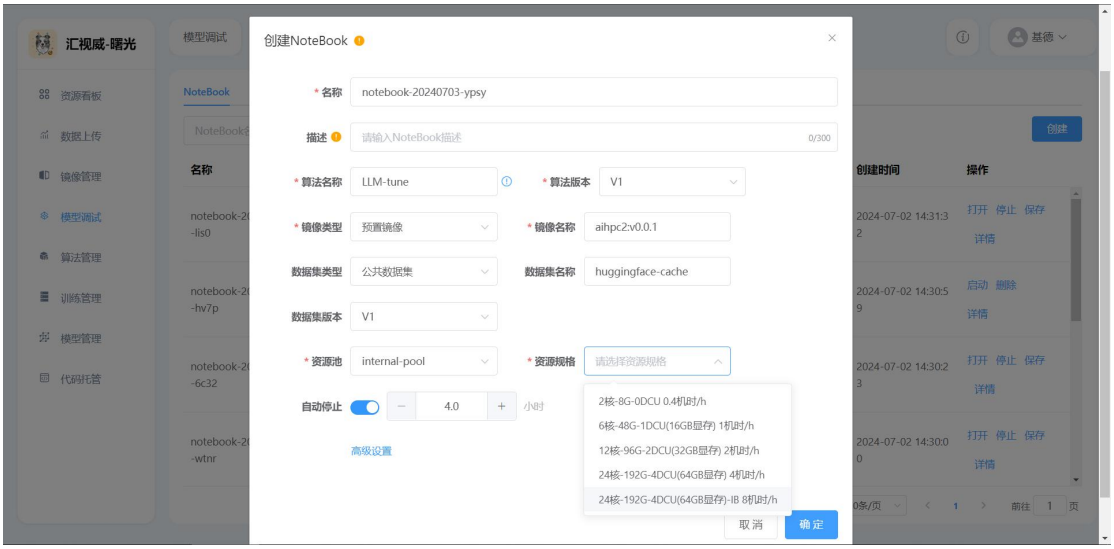


4. 创建 Notebook

1. 点击平台左侧的【模型调试】按钮，进入【模型调试】模块，点击创建，开始创建 Notebook 进行模型微调。

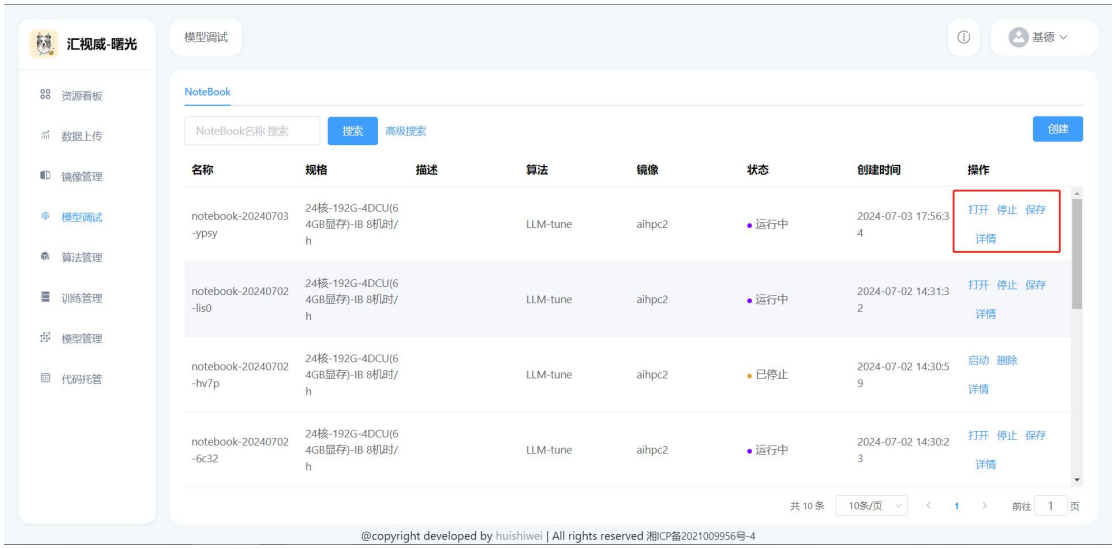


2. 创建 Notebook 时，需要选择合适的运行环境，对于 LLM 微调任务，平台已内置运行环境。在【镜像类型】中选择【预置镜像】,【镜像名称】选择【aihpc2:v0.0.1】



3. 【资源池】选择说明，平台使用的是海光 DCU 国产芯片，如果模型比较小可以选择 1DCU（16G 显存），如果模型比较大，使用不超过 4DCU，可以选择 2DCU 或 4DCU。如果模型参数超过 7B，需要用到多机多卡分布式训练，此时建议选择 4DCU-IB 的规格。
4. 平台还提供了两个高级选项，如果模型训练预估的时间比较准确，可以设置自动停止的时间，平台默认 4 小时后会主动关停 Notebook。可以在创建的时候关闭这个功能。此外，Notebook 提供了多机多卡的调试能力，在【任务数】里可以选择 2，此时平台会创建两个 Notebook 供用户使用。

5. 点击【提交】，平台将会创建 Notebook，创建好后，Notebook 的状态会显示【运行中】

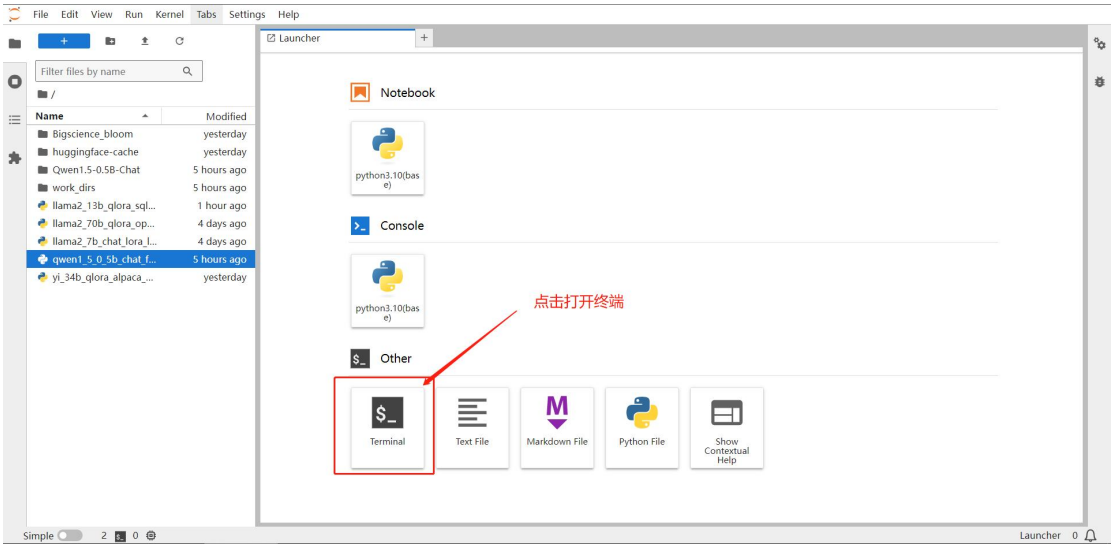


5. 训练模型

按照上面的步骤创建好 Notebook 后，我们可以开始调试我们的训练任务。
微调任务目前有很多开源工具可以使用，这里已 xtuner 为例。

单卡微调 Qwen1.5-0.5B 模型

创建 Notebook 时选择 LLM 算法以及 huggingface-cache 数据集后即可一键启动训练任务，打开终端



终端内输入如下命令即可开启训练：

```
cd /code/
NPROC_PER_NODE=1 xtuner train
qwen1_5_0_5b_chat_full_alpaca_e3_copy.py --work-dir
/userhome/xtuner-workdir --deepspeed
deepspeed_zero3_offload
```

这里所使用的训练配置文件均已配置好，可以根据自己的需求（数据/模型大小）进行修改。

4 卡微调 Llama-2-7B 模型

打开终端后在终端中输入如下命令：

```
cd /code/
NPROC_PER_NODE=4 xtuner train
llama2_7b_chat_lora_lawyer_e3_copy.py --work-dir
/userhome/xtuner-workdir --deepspeed deepspeed_zero3
```

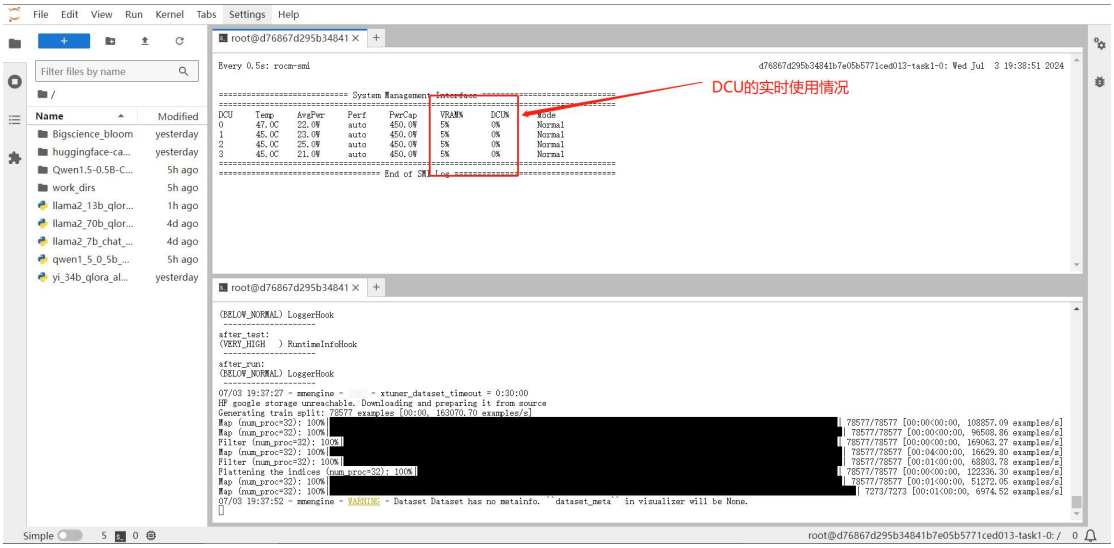
8 卡微调 Llama-2-13B 模型

这里需要注意的是，由于我们的机器单台有 4 块 DCU，8 卡 DCU 需要进行机器间通信，在创建 Notebook 的时候，使用资源规格为【24 核-192G-4DCU(64GB 显存)-IB】会让模型训练的更快（大约 4 倍速度提升）。

本例中有两个 Notebook，需要分别打开各自的 Notebook 终端并设置环境变量来启用 IB，如下

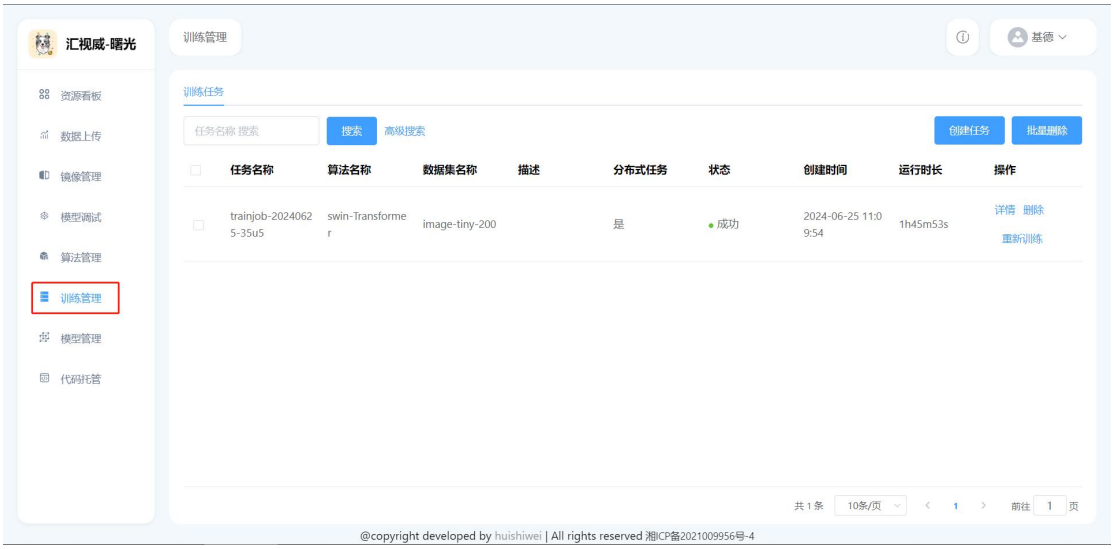
```
# 设置环境变量
export NCCL_DEBUG=INFO
export NCCL_IB_DISABLE=0
export NCCL_IB_HCA=mlx5
export NCCL_SOCKET_IFNAME=eth0
export GLOO_SOCKET_IFNAME=eth0
export HF_HOME=/code/huggingface-cache/
# 开始训练
cd /code/
# 在 Notebook 0 中执行
# 需要先在 Notebook0 中执行 ifconfig 查看 IP 地址，比如这里得到
10.244.88.180
NPROC_PER_NODE=4 NNODES=2 PORT=12345 ADDR=10.244.88.180
NODE_RANK=0 xtuner train llama2_13b_qlora_sql_e3_copy.py --
work-dir /userhome/xtuner-workdir --deepspeed
deepspeed_zero3_offload
# 在 Notebook 1 中执行
NPROC_PER_NODE=4 NNODES=2 PORT=12345 ADDR=10.244.88.180
NODE_RANK=1 xtuner train llama2_13b_qlora_sql_e3_copy.py --
work-dir /userhome/xtuner-workdir --deepspeed
deepspeed_zero3_offload
```

如果想在训练过程中观察 DCU 的使用情况，也可以再打开一个终端，输入`watch -n 1 rocm-smi` 来查看，如下所示



6. 提交训练任务

当我们想提交一个训练任务，等训练任务结束自动停止“收费”时，可以使用【训练管理】功能。



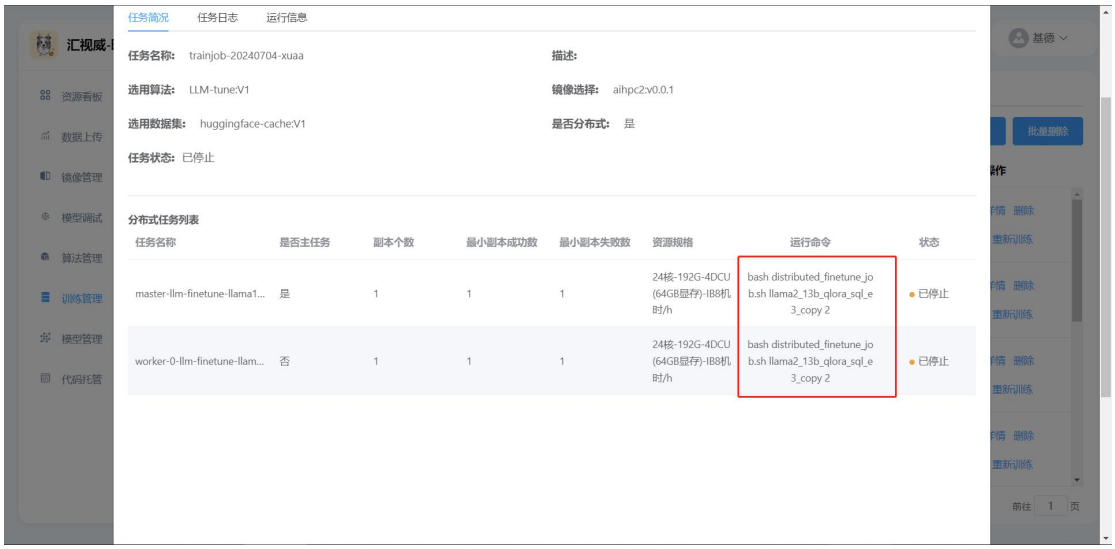
配置信息如下所示，以下展示的是一个 2 机 8 卡的分布式训练任务。

为了方便使用，我们可以编写如下脚本 distributed_finetune_job.sh

```
#!/bin/bash

export
LD_LIBRARY_PATH=/opt/dtk/hip/lib:/opt/dtk/llvm/lib:/opt/dtk
/lib:/opt/dtk/lib64:/opt/hyhal/lib:/opt/hyhal/lib64:/opt/dt
k/.hyhal/lib:/opt/dtk/.hyhal/lib64:/opt/dtk-
24.04/hip/lib:/opt/dtk-24.04/llvm/lib:/opt/dtk-
24.04/lib:/opt/dtk-
24.04/lib64:/opt/hyhal/lib:/opt/hyhal/lib64:/opt/dtk-
24.04/.hyhal/lib:/opt/dtk-
24.04/.hyhal/lib64:/usr/local/lib/:/usr/local/lib64/:/opt/m
pi/lib:/opt/hwloc/lib:/opt/dtk/hip/lib:/opt/dtk/llvm/lib:/o
pt/dtk/lib:/opt/dtk/lib64:/opt/hyhal/lib:/opt/hyhal/lib64:/
opt/mpi/lib:/opt/hwloc/lib:/usr/local/lib/:/usr/local/lib64
/:$LD_LIBRARY_PATH

# 设置环境变量
export NCCL_DEBUG=INFO
export NCCL_IB_DISABLE=0
export NCCL_IB_HCA=mlx5
export NCCL_SOCKET_IFNAME=eth0
export GLOO_SOCKET_IFNAME=eth0
export HF_HOME=/userhome/huggingface-cache/

# 设置训练参数
export TRAIN_CONFIG=${1:-"llama2_13b_qlora_sql_e3_copy"}
export WORLD_SIZE=${2:-1}
export GPU=`rocm-smi| grep "auto" | wc -l`
export MASTER_ADDR=${TASKSET_NAME}-task0-0.${TASKSET_NAME}
export MASTER_PORT=12345
export RANK=`echo ${VC_TASK_NAME} | grep -o '[0-9]'`

# 训练模型
NPROC_PER_NODE=${GPU} NNODES=${WORLD_SIZE}
PORT=${MASTER_PORT} ADDR=${MASTER_ADDR} NODE_RANK=${RANK}
xtuner train /code/${TRAIN_CONFIG}.py --work-dir
/userhome/xtuner-workdir --deepspeed
deepspeed_zero3_offload
```

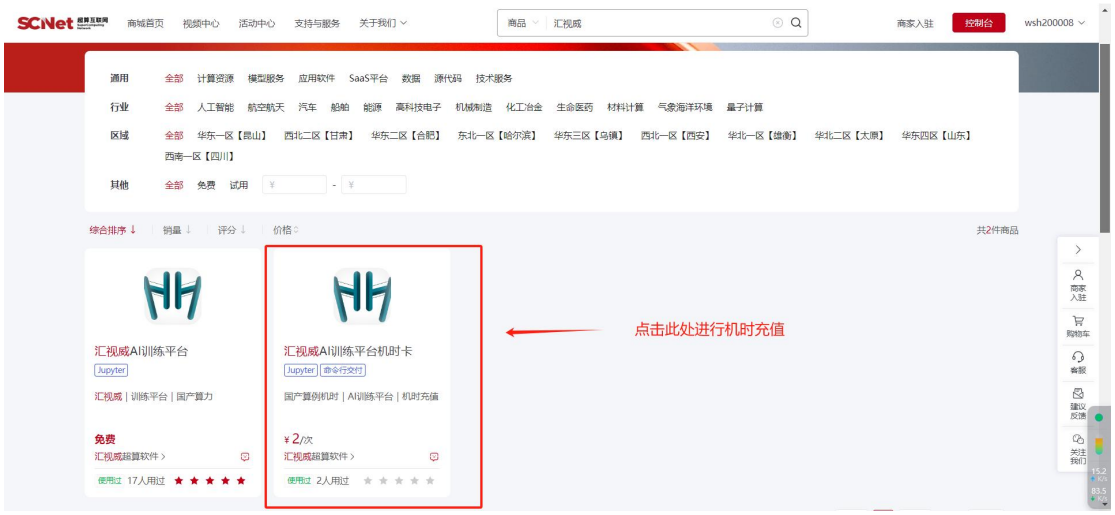
此时我们在提交任务时，仅需要填入使用多少台机器的参数（WORLD_SIZE）即可。

运行命令：bash distributed_finetune_job.sh llama2_13b_qlora_sql_e3_copy 2



7. 机时充值

1.进入超算互联网平台，搜索“汇视威 AI 训练平台机时卡”。



2.根据需要选择要充值的机时卡规格

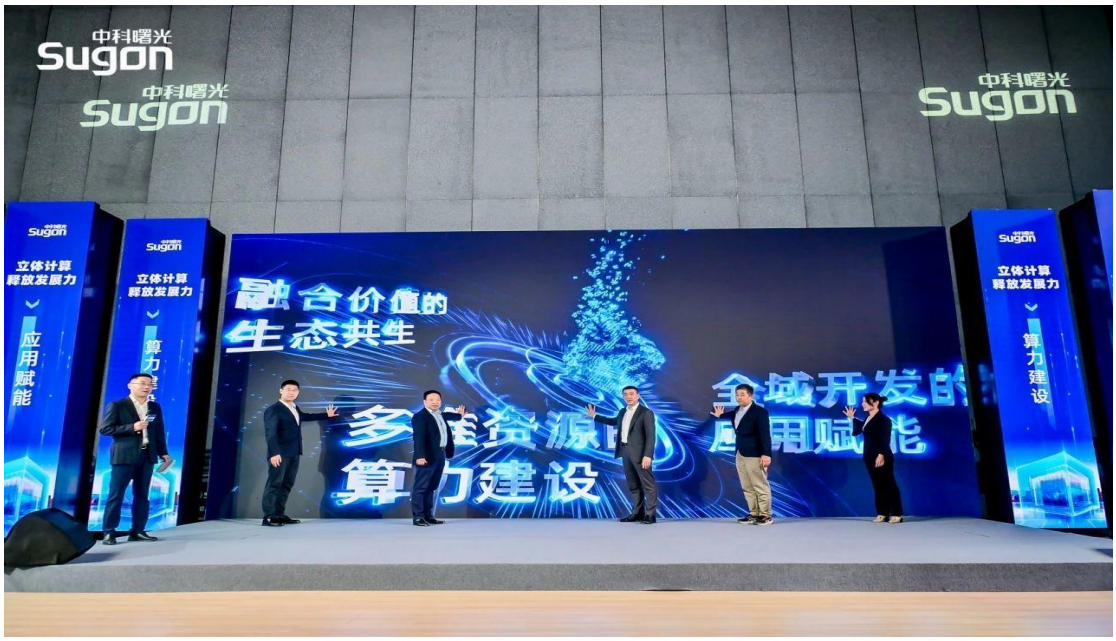

SCNet
商城首页
视频中心
活动中心
支持与服务
关于我们
商家入驻
控制台
wsh200008
商品
汇视威
搜索
我的购物车
返回
商品广场 > 技术服务 > 汇视威AI训练平台机时卡
汇视威超算软件
立即咨询
欢迎使用
Star AI平台
汇视威AI训练平台机时卡
该商品为汇视威AI训练平台机时充值专用商品，您购买商品后，您对应的训练平台账号会自动增加相应的机时。
规格
1机时卡
2机时卡
5机时卡
10机时卡
100机时卡
付费模式
次
单价
¥2 /次 自购买之日起 12个月 有效
总计
¥2
选择所需的机时卡进行充值
您已购买此商品，可直接 去使用
再次购买
加入购物车
商品详情
服务与支持
商品评价

3.机时到账

相关新闻

新闻 1

2024 年 4 月 2 日，中科曙光“立体计算湖南行”启动仪式在长沙成功举办。面对“加快发展新质生产力”的新要求，中科曙光提出“立体计算”新思路，旨在打造一种全新的计算体系构建与运营模式。汇视威作为曙光战略合作伙伴参与启动仪式。



汇视威开启发布仪式（左一）

汇视威联合启动立体计算仪式（左一）

汇视威现场演讲立体计算

新闻 2：

2024 年 4 月 11 日，首届超算互联网峰会暨国家超算互联网平台上线仪式在天津顺利举办。汇视威携国产算力产品参展。

新闻 3：

2024 年 5 月 29 日，国家超算互联网平台在湖南省长沙经济技术开发区管委会举办首场生态沙龙 -“超”话坊 ·“工业数智化创新发展探讨”。汇视威选取了智慧化工、车间等方向的工业数智化案例，分享了依据不同场景特点形成的推动行业智慧化升级的可服用的行业垂直 AI 一体机解决方案。

汇视威现场汇报

国家超算互联网生态沙龙

新闻 4：

2023 年 4 月 15 日，“凝心聚力，全芯超越——2023 光合行动·创芯中国行”长沙站成功举办。作为光合组织集智计划落地成果，会上进行长沙智算应用合作伙伴签约仪式。北大研究院、科创信息、汇视威等机构、企业，宣布入驻长沙人工智能创新中心。

汇视威（右一）签约智算应用

光合组织汇视威展台