- 华为昇腾 910B 部署 DeepSeek-R1 蒸馏系列模型详细指南 - 华为昇腾910B部署DeepSeek-R1蒸馏系列模型详细指南

## 华为昇腾 910B 部署 DeepSeek-R1 蒸馏系列模型详细指南

本文全面记录了在华为昇腾 910B (65GB) \* 8 上部署 DeepSeekR1 蒸馏系列模型（14B、32B）的全过程及测试结果。涵盖模型下载、权重转换、环境依赖安装、镜像拉取与启动、环境确认以及模型服务启动配置设置等环节。文中详细介绍了每一步的操作命令和注意事项，并对可能出现的问题提供了相应的解决方法和建议，为相关技术人员提供了完整且实用的部署参考。

本文记录 在 华为昇腾 910B(65GB) \* 8 上 部署 DeepSeekR1 蒸馏系列模型（14B、32B）全过程与测试结果。

NPU：910B3 (65GB) \* 8 （910B 有三个版本 910B1、2、3）

模型：DeepSeek-R1-Distill-Qwen-14B、DeepSeek-R1-Distill-Qwen-32B

部署方法：镜像部署 [1.0.0-800I-A2-py311-openeuler24.03-lts](https://link.csdn.net/?target=https%3A%2F%2Fwww.hiascend.com%2Fdeveloper%2Fascendhub%2Fdetail%2Faf85b724a7e5469ebd7ea13c3439d48f) （需要申请下载权限，审核需要2天左右）

本文基础环境如下：

----------------  
aarch64  
910B(65GB) \* 8  
CANN 7.0  
npu-smi 23.0.2.1  
----------------

### 模型下载

[DeepSeek-R1-Distill-Qwen-14B · 模型库](https://link.csdn.net/?target=https%3A%2F%2Fwww.modelscope.cn%2Fmodels%2Fdeepseek-ai%2FDeepSeek-R1-Distill-Qwen-14B)

[DeepSeek-R1-Distill-Qwen-32B · 模型库](https://link.csdn.net/?target=https%3A%2F%2Fwww.modelscope.cn%2Fmodels%2Fdeepseek-ai%2FDeepSeek-R1-Distill-Qwen-32B)

[modelscope 魔搭社区模型下载](https://link.csdn.net/?target=https%3A%2F%2Fblog.csdn.net%2FMnivL%2Farticle%2Fdetails%2F145465800%3Fspm%3D1001.2014.3001.5501)

本文将模型下载到服务器的 /data1/apps/models 路径下

* 例如 /data1/apps/models/deepseek-ai/DeepSeek-R1-Distill-Qwen-14B

启动镜像时，将路径挂载：-v /data1/apps/models:/storage/llm

那么在镜像容器内模型地址：

/storage/llm/deepseek-ai/DeepSeek-R1-Distill-Qwen-14B

#### 权重转换

14B、32B 是.safetensor权重，无需转换，可以直接使用。

### 环境依赖

##### 拉取镜像

下载地址： [昇腾镜像仓库详情](https://link.csdn.net/?target=https%3A%2F%2Fwww.hiascend.com%2Fdeveloper%2Fascendhub%2Fdetail%2Faf85b724a7e5469ebd7ea13c3439d48f)

登陆账号，申请下载权限 – 点击立即下载 — 弹出一个窗口 – 按照指示拉取镜像

docker pull --platform=linux/arm64 swr.cn-south-1.myhuaweicloud.com/ascendhub/mindie:1.0.0-800I-A2-py311-openeuler24.03-lts

--platform=linux/arm64 指定拉去内核为 arm 架构版本的镜像

由于本文的910B是纯内网机器，无法直接访问下载

于是 找了台可以访问公网的机器（x86的），拉取镜像、导出、传输到内网机器、导入

如果你的机器可以访问公网，直接拉去即可

查看拉取的镜像版本

docker inspect 25ba5f455ae3| grep Architecture

导出镜像

docker save -o 1.0.0-800I-A2-py311-openeuler24.03-lts.tar swr.cn-south-1.myhuaweicloud.com/ascendhub/mindie:1.0.0-800I-A2-py311-openeuler24.03-lts

* 加载

docker load -i 1.0.0-800I-A2-py311-openeuler24.03-lts.tar1.0.0-300I-Duo-py311-openeuler24.03-lts.tar

* 环境
  + python 3.11
  + torch 2.1
* [root@pm-a813-005 DeepSeek-R1-Distill-Qwen-14B]# python --version  
  Python 3.11.6  
  [root@pm-a813-005 DeepSeek-R1-Distill-Qwen-14B]# pip show torch  
  Name: torch  
  Version: 2.1.0  
  Summary: Tensors and Dynamic neural networks in Python with strong GPU acceleration  
  Home-page: https://pytorch.org/  
  Author: PyTorch Team  
  Author-email: packages@pytorch.org  
  License: BSD-3  
  Location: /usr/local/lib64/python3.11/site-packages  
  Requires: filelock, fsspec, jinja2, networkx, sympy, typing-extensions  
  Required-by: accelerate, torch-npu, torchvision
  + mindie\_llm 1.0.0
  + mindiebenchmark 1.0.0
  + mindieclient 1.0.0
  + mindiesd 1.0.0
  + mindietorch 1.0.0+torch2.1.0.abi0

### 启动镜像

#### root 特权模型

docker run -it -d --net=host --shm-size=10g \  
 --privileged \  
 --name deepseek-r1-distill-root-test \  
 -v /usr/local/Ascend/driver:/usr/local/Ascend/driver:ro \  
 -v /usr/local/sbin:/usr/local/sbin:ro \  
 -v /data1/apps/models:/storage/llm \  
 swr.cn-south-1.myhuaweicloud.com/ascendhub/mindie:1.0.0-800I-A2-py311-openeuler24.03-lts \  
 bash

docker exec -it deepseek-r1-distill-root-test bash

#### 普通用户

docker run -it -d --net=host --shm-size=10g \  
 --name deepseek-r1-distill-test1 \  
 --device=/dev/davinci\_manager \  
 --device=/dev/hisi\_hdc \  
 --device=/dev/devmm\_svm \  
 --device=/dev/davinci0 \  
 --device=/dev/davinci1 \  
 --device=/dev/davinci2 \  
 --device=/dev/davinci3 \  
 --device=/dev/davinci4 \  
 --device=/dev/davinci5 \  
 --device=/dev/davinci6 \  
 --device=/dev/davinci7 \  
 -v /usr/local/Ascend/driver:/usr/local/Ascend/driver:ro \  
 -v /usr/local/sbin:/usr/local/sbin:ro \  
 -v /data1/apps/models:/storage/llm \  
 -w /storage/llm \  
 swr.cn-south-1.myhuaweicloud.com/ascendhub/mindie:1.0.0-800I-A2-py311-openeuler24.03-lts \  
 bash

注意，以上启动命令仅供参考，请根据需求自行修改再启动容器，尤其需要注意：

1. --user，如果您的环境中HDK是通过普通用户安装（例如默认的HwHiAiUser，可以通过id HwHiAiUser命令查看该用户组ID），请设置好对应的用户组，例如用户组1001可以使用HDK，则–user mindieuser:1001，镜像中默认使用的是用户组1000。如果您的HDK是由root用户安装，且指定了–install-for-all参数，则无需指定–user参数。
2. 设定容器名称–name与镜像名称，例如mindie:1.0.0-800I-A2-py311-openeuler24.03-lts。
3. 如果不使用–priviliged参数，则需要设置各设备，包括设置想要使用的卡号–device：

...  
--name <container-name> \  
--device=/dev/davinci\_manager \  
--device=/dev/hisi\_hdc \  
--device=/dev/davinci0 \  
...

1. 设定权重挂载的路径，-v /path-to-weights:/path-to-weights:ro，注意，权重路径权限应当设置为750。如果使用普通用户镜像，权重路径所属应为镜像内默认的1000用户。可参考以下命令进行修改：

chmod -R 755 /path-to-weights  
chown -R 1000:1000 /path-to-weights  
# 进入容器后执行  
chmod -R 755 /storage/llm  
chown -R 1000:1000 /storage/llm

1. 在普通用户镜像中，注意所有文件均在 /home/mindieuser 下，请勿直接挂载 /home 目录，以免宿主机上存在相同目录，将容器内文件覆盖清除。

* 进入容器

docker exec -it deepseek-r1-distill-test1 bash

### 确认环境

#### 检验HDK是否可用

输入以下命令，应当正确显示设备信息：

npu-smi info

如果出现以下信息：

bash: npu-smi: command not found

说明宿主机上的 npu-smi 工具不在 /usr/local/sbin 路径中，可能是由于HDK版本过旧或其他原因导致，可以使用以下命令找到该工具，并在启动容器时将其挂载到容器内：

find / -name npu-smi

一般来说，可能出现在 /usr/local/bin/npu-smi 路径下。

#### 检验Torch是否可用

启动Python，并输入以下命令：

import torch  
import torch\_npu

若无报错信息，则说明Torch组件正常。

#### 检查MindIE各组件

输入以下命令：

pip list | grep mindie

应出现类似如下输出：

mindie\_llm 1.0.0  
mindiebenchmark 1.0.0  
mindieclient 1.0.0  
mindiesd 1.0.0  
mindietorch 1.0.0+torch2.1.0.abi0

或者输入以下命令：

cat /home/mindieuser/Ascend/mindie/latest/version.info

应出现类似如下输出：

Ascend-mindie : MindIE 1.0.0  
mindie-rt: 1.0.0  
mindie-torch: 1.0.0  
mindie-service: 1.0.0  
mindie-llm: 1.0.0  
mindie-sd:1.0.0  
Platform : aarch64

说明各组件正常。

#### 确认模型地址正确

cd /storage/llm/deepseek-ai/DeepSeek-R1-Distill-Qwen-7B  
cd /storage/llm/deepseek-ai/DeepSeek-R1-Distill-Qwen-14B  
cd /storage/llm/deepseek-ai/DeepSeek-R1-Distill-Qwen-32B

确认权限

chmod -R 750 /storage/llm

### 设置模型服务启动配置

#### 打开配置文件

vi /usr/local/Ascend/mindie/latest/mindie-service/conf/config.json

#### 修改建议

一般只需要修改以下配置（单实例）

{  
 ...  
 "ServerConfig" :  
 {  
 ...  
 "port" : 1040, #自定义  
 "managementPort" : 1041, #自定义  
 "metricsPort" : 1042, #自定义  
 ...  
 "httpsEnabled" : false, # 取消https协议启动服务  
 ...  
 },  
  
 "BackendConfig": {  
 ...  
 "npuDeviceIds" : [[0,1]],  
 ...  
 "ModelDeployConfig":  
 {  
 "truncation" : false,  
 "ModelConfig" : [  
 {  
 ...  
 "modelName" : "deepseek-14b",  
 "modelWeightPath" : "/storage/llm/deepseek-ai/DeepSeek-R1-Distill-Qwen-14B",  
 "worldSize" : 2,  
 ...  
 }  
 ]  
 },  
 }  
}

* （多实例）以 14B 为例， 一张卡一个实例， 八张就是八个实例，并发要求200

### ServerConfig  
- \*\*\`httpsEnabled\`\*\*：取消https协议 设为 false  
  
### 3. \`ModelDeployParam\` 部分  
#### 整体配置  
- \*\*\`modelInstanceNumber\`\*\*：由于单卡能跑一个实例，机器有 8 张卡，可设置为 8。  
- \*\*\`tokenizerProcessNumber\`\*\*：可使用默认值 8，也可根据实际性能情况进行调整。  
- \*\*\`maxSeqLen\`\*\*：根据需求，最大上下文为 8192，设置为 8192 + 4096 = 12288（输入长度 + 输出长度）。  
- \*\*\`npuDeviceIds\`\*\*：由于是单机 8 卡，设置为 \`[[0], [1], [2], [3], [4], [5], [6], [7]]\`。  
- \*\*\`multiNodesInferEnabled\`\*\*：设置为 \`false\`，因为是单机推理。  
  
#### \`ModelParam\`  
- \*\*\`worldSize\`\*\*：由于使用 8 张卡, 8 个实例，一个实例一张卡，设置为 1。  
- \*\*\`cpuMemSize\`\*\*：CPU 内存有 1.4T，可适当增大，例如设置为 100（单位：GB）。  
- \*\*\`npuMemSize\`\*\*：使用快速计算公式计算：  
 - 假设单卡总空闲显存为 60GB，模型权重占用 40GB（根据实际情况调整），后处理占用 1GB，系数取 0.8。  
 - 则 \`npuMemSize = (60 - 40/1 - 1) \* 0.8 ≈ 15\`，可设置为 15（单位：GB）。  
  
### 4. \`ScheduleParam\` 部分  
- \*\*\`maxPrefillBatchSize\`\*\*：可根据实际性能测试进行调整，初始可设置为 200。  
- \*\*\`maxPrefillTokens\`\*\*：设置为大于等于 \`maxSeqLen\` 的值，例如设置为 16384。  
- \*\*\`prefillTimeMsPerReq\`\*\*：可根据实际情况调整，使用默认值 150。  
- \*\*\`prefillPolicyType\`\*\*：可使用默认值 0（FCFS，先来先服务）。  
- \*\*\`decodeTimeMsPerReq\`\*\*：可根据实际情况调整，使用默认值 50。  
- \*\*\`decodePolicyType\`\*\*：可使用默认值 0（FCFS，先来先服务）。  
- \*\*\`maxBatchSize\`\*\*：根据 \`npuMemSize\` 和 \`cacheBlockSize\` 等参数重新计算，初始可设置为 200。  
- \*\*\`maxIterTimes\`\*\*：最大输出为 4096，设置为 4096。  
- \*\*\`maxPreemptCount\`\*\*：可根据实际情况设置，初始可设置为 0。  
- \*\*\`supportSelectBatch\`\*\*：可根据实际情况设置，初始可使用默认值 \`false\`。  
- \*\*\`maxQueueDelayMicroseconds\`\*\*：使用默认值 5000。  
  
### 预估最大并发量  
最大并发量受多种因素影响，包括模型复杂度、硬件性能、参数配置等。  
上面的配置，理论上最大并发量可达到 200 左右，但实际并发量需要通过性能测试来确定。可以逐步增加并发请求，观察系统的响应时间、资源利用率等指标，找到系统的性能瓶颈，从而确定最大并发量。

#### 配置示例

##### 14B

{  
 "Version" : "1.1.0",  
 "LogConfig" :  
 {  
 "logLevel" : "Info",  
 "logFileSize" : 20,  
 "logFileNum" : 20,  
 "logPath" : "logs/mindservice.log"  
 },  
  
 "ServerConfig" :  
 {  
 "ipAddress" : "127.0.0.1",  
 "managementIpAddress" : "127.0.0.2",  
 "port" : 1025,  
 "managementPort" : 1026,  
 "metricsPort" : 1027,  
 "allowAllZeroIpListening" : false,  
 "maxLinkNum" : 1000,  
 "httpsEnabled" : false,  
 "fullTextEnabled" : false,  
 "tlsCaPath" : "security/ca/",  
 "tlsCaFile" : ["ca.pem"],  
 "tlsCert" : "security/certs/server.pem",  
 "tlsPk" : "security/keys/server.key.pem",  
 "tlsPkPwd" : "security/pass/key\_pwd.txt",  
 "tlsCrlPath" : "security/certs/",  
 "tlsCrlFiles" : ["server\_crl.pem"],  
 "managementTlsCaFile" : ["management\_ca.pem"],  
 "managementTlsCert" : "security/certs/management/server.pem",  
 "managementTlsPk" : "security/keys/management/server.key.pem",  
 "managementTlsPkPwd" : "security/pass/management/key\_pwd.txt",  
 "managementTlsCrlPath" : "security/management/certs/",  
 "managementTlsCrlFiles" : ["server\_crl.pem"],  
 "kmcKsfMaster" : "tools/pmt/master/ksfa",  
 "kmcKsfStandby" : "tools/pmt/standby/ksfb",  
 "inferMode" : "standard",  
 "interCommTLSEnabled" : true,  
 "interCommPort" : 1121,  
 "interCommTlsCaPath" : "security/grpc/ca/",  
 "interCommTlsCaFiles" : ["ca.pem"],  
 "interCommTlsCert" : "security/grpc/certs/server.pem",  
 "interCommPk" : "security/grpc/keys/server.key.pem",  
 "interCommPkPwd" : "security/grpc/pass/key\_pwd.txt",  
 "interCommTlsCrlPath" : "security/grpc/certs/",  
 "interCommTlsCrlFiles" : ["server\_crl.pem"],  
 "openAiSupport" : "vllm"  
 },  
  
 "BackendConfig" : {  
 "backendName" : "mindieservice\_llm\_engine",  
 "modelInstanceNumber" : 8,  
 "npuDeviceIds" : [[0], [1], [2], [3], [4], [5], [6], [7]],  
 "tokenizerProcessNumber" : 8,  
 "multiNodesInferEnabled" : false,  
 "multiNodesInferPort" : 1120,  
 "interNodeTLSEnabled" : true,  
 "interNodeTlsCaPath" : "security/grpc/ca/",  
 "interNodeTlsCaFiles" : ["ca.pem"],  
 "interNodeTlsCert" : "security/grpc/certs/server.pem",  
 "interNodeTlsPk" : "security/grpc/keys/server.key.pem",  
 "interNodeTlsPkPwd" : "security/grpc/pass/mindie\_server\_key\_pwd.txt",  
 "interNodeTlsCrlPath" : "security/grpc/certs/",  
 "interNodeTlsCrlFiles" : ["server\_crl.pem"],  
 "interNodeKmcKsfMaster" : "tools/pmt/master/ksfa",  
 "interNodeKmcKsfStandby" : "tools/pmt/standby/ksfb",  
 "ModelDeployConfig" :  
 {  
 "maxSeqLen" : 12288,  
 "maxInputTokenLen" : 8192,  
 "truncation" : false,  
 "ModelConfig" : [  
 {  
 "modelInstanceType" : "Standard",  
 "modelName" : "deepseek-14b",  
 "modelWeightPath" : "/storage/llm/deepseek-ai/DeepSeek-R1-Distill-Qwen-14B",  
 "worldSize" : 1,  
 "cpuMemSize" : 50,  
 "npuMemSize" : -1,  
 "backendType" : "atb",  
 "trustRemoteCode" : false  
 }  
 ]  
 },  
  
 "ScheduleConfig" :  
 {  
 "templateType" : "Standard",  
 "templateName" : "Standard\_LLM",  
 "cacheBlockSize" : 128,  
  
 "maxPrefillBatchSize" : 200,  
 "maxPrefillTokens" : 16384,  
 "prefillTimeMsPerReq" : 150,  
 "prefillPolicyType" : 0,  
  
 "decodeTimeMsPerReq" : 50,  
 "decodePolicyType" : 0,  
  
 "maxBatchSize" : 200,  
 "maxIterTimes" : 4096,  
 "maxPreemptCount" : 0,  
 "supportSelectBatch" : false,  
 "maxQueueDelayMicroseconds" : 5000  
 }  
 }  
}

##### 32B

{  
 "Version" : "1.1.0",  
 "LogConfig" :  
 {  
 "logLevel" : "Verbose",  
 "logFileSize" : 200,  
 "logFileNum" : 64,  
 "logPath" : "logs/mindservice.log"  
 },  
  
 "ServerConfig" :  
 {  
 "ipAddress" : "127.0.0.1",  
 "managementIpAddress" : "127.0.0.2",  
 "port" : 1025,  
 "managementPort" : 1026,  
 "metricsPort" : 1027,  
 "allowAllZeroIpListening" : false,  
 "maxLinkNum" : 1000,  
 "httpsEnabled" : false,  
 "fullTextEnabled" : false,  
 "tlsCaPath" : "security/ca/",  
 "tlsCaFile" : ["ca.pem"],  
 "tlsCert" : "security/certs/server.pem",  
 "tlsPk" : "security/keys/server.key.pem",  
 "tlsPkPwd" : "security/pass/key\_pwd.txt",  
 "tlsCrlPath" : "security/certs/",  
 "tlsCrlFiles" : ["server\_crl.pem"],  
 "managementTlsCaFile" : ["management\_ca.pem"],  
 "managementTlsCert" : "security/certs/management/server.pem",  
 "managementTlsPk" : "security/keys/management/server.key.pem",  
 "managementTlsPkPwd" : "security/pass/management/key\_pwd.txt",  
 "managementTlsCrlPath" : "security/management/certs/",  
 "managementTlsCrlFiles" : ["server\_crl.pem"],  
 "kmcKsfMaster" : "tools/pmt/master/ksfa",  
 "kmcKsfStandby" : "tools/pmt/standby/ksfb",  
 "inferMode" : "standard",  
 "interCommTLSEnabled" : true,  
 "interCommPort" : 1121,  
 "interCommTlsCaPath" : "security/grpc/ca/",  
 "interCommTlsCaFiles" : ["ca.pem"],  
 "interCommTlsCert" : "security/grpc/certs/server.pem",  
 "interCommPk" : "security/grpc/keys/server.key.pem",  
 "interCommPkPwd" : "security/grpc/pass/key\_pwd.txt",  
 "interCommTlsCrlPath" : "security/grpc/certs/",  
 "interCommTlsCrlFiles" : ["server\_crl.pem"],  
 "openAiSupport" : "vllm"  
 },  
  
 "BackendConfig" : {  
 "backendName" : "mindieservice\_llm\_engine",  
 "modelInstanceNumber" : 4,  
 "npuDeviceIds" : [[0,1], [2,3], [4,5], [6,7]],  
 "tokenizerProcessNumber" : 8,  
 "multiNodesInferEnabled" : false,  
 "multiNodesInferPort" : 1120,  
 "interNodeTLSEnabled" : true,  
 "interNodeTlsCaPath" : "security/grpc/ca/",  
 "interNodeTlsCaFiles" : ["ca.pem"],  
 "interNodeTlsCert" : "security/grpc/certs/server.pem",  
 "interNodeTlsPk" : "security/grpc/keys/server.key.pem",  
 "interNodeTlsPkPwd" : "security/grpc/pass/mindie\_server\_key\_pwd.txt",  
 "interNodeTlsCrlPath" : "security/grpc/certs/",  
 "interNodeTlsCrlFiles" : ["server\_crl.pem"],  
 "interNodeKmcKsfMaster" : "tools/pmt/master/ksfa",  
 "interNodeKmcKsfStandby" : "tools/pmt/standby/ksfb",  
 "ModelDeployConfig" :  
 {  
 "maxSeqLen" : 13000,  
 "maxInputTokenLen" : 4096,  
 "truncation" : false,  
 "ModelConfig" : [  
 {  
 "modelInstanceType" : "Standard",  
 "modelName" : "deepseek-32b",  
 "modelWeightPath" : "/storage/llm/deepseek-ai/DeepSeek-R1-Distill-Qwen-32B",  
 "worldSize" : 2,  
 "cpuMemSize" : 100,  
 "npuMemSize" : 10,  
 "backendType" : "atb",  
 "trustRemoteCode" : false  
 }  
 ]  
 },  
  
 "ScheduleConfig" :  
 {  
 "templateType" : "Standard",  
 "templateName" : "Standard\_LLM",  
 "cacheBlockSize" : 128,  
  
 "maxPrefillBatchSize" : 50,  
 "maxPrefillTokens" : 16384,  
 "prefillTimeMsPerReq" : 150,  
 "prefillPolicyType" : 0,  
  
 "decodeTimeMsPerReq" : 50,  
 "decodePolicyType" : 0,  
  
 "maxBatchSize" : 200,  
 "maxIterTimes" : 4096,  
 "maxPreemptCount" : 0,  
 "supportSelectBatch" : false,  
 "maxQueueDelayMicroseconds" : 5000  
 }  
 }  
}

### 服务启动项参数说明

详细查看官网  
[配置参数说明-快速开始-MindIE Service开发指南-服务化集成部署-MindIE1.0.RC2开发文档-昇腾社区](https://link.csdn.net/?target=https%3A%2F%2Fwww.hiascend.com%2Fdocument%2Fdetail%2Fzh%2Fmindie%2F10RC2%2Fmindieservice%2Fservicedev%2Fmindie_service0004.html)

#### OtherParam参数

| 配置项 | 取值类型 | 取值范围 | 配置说明 |
| --- | --- | --- | --- |
| ResourceParam |  |  |  |
| cacheBlockSize | uint32\_t | [1, 128] | kvcache block的size大小。必填，默认值：128；建议值：128，其他值建议取为2的n次幂。 |
| LogParam |  |  |  |
| logLevel | string | “Verbose”“Info”“Warning”“Error”“None” | “Verbose”：打印Verbose、Info、Warning和Error级别的日志。“Info”：打印Info、Warning和Error级别的日志。“Warning”：打印Warning和Error级别的日志。“Error”：打印Error级别的日志。“None”：不打印日志。必填，默认值：“Info”。 |
| logPath | string | 日志文件路径，长度<=4096。 | 支持绝对和相对路径。如果配置为相对路径，则代码中会取工程目录，最后拼接而成。例如，假设MindIE Service的安装路径为“/opt/Ascend-mindie-service{version}linux-x86\_64/”，则默认的日志绝对路径为“/opt/Ascend-mindie-service{version}linux-x86\_64/logs/mindservice.log”。若配置路径不满足要求，则使用默认路径：“工程路径/logs/mindservice.log”。必填，默认值：“logs/mindservice.log”。 |
| ServeParam |  |  |  |
| ipAddress | string | IPv4地址。 | EndPoint提供的业务面RESTful接口绑定的IP地址。全零侦听会导致三面隔离失效，不满足安全配置要求，禁止绑定IP地址为0.0.0.0。如果存在环境变量MIES\_CONTAINER\_IP，则优先取环境变量值作为业务面IP地址。如果不存在环境变量MIES\_CONTAINER\_IP，则取该配置值。必填，默认值：“127.0.0.1”。 |
| managementIpAddress | string | IPv4地址。 | EndPoint提供的管理面RESTful接口绑定的IP地址。全零侦听会导致三面隔离失效，不满足安全配置要求，禁止绑定IP地址为0.0.0.0。如果该环境变量MIES\_CONTAINER\_MANAGEMENT\_IP存在，则直取环境变量值作为管理面IP地址。如果“managementIpAddress”字段存在，则取字段本身值；否则取“ipAddress”字段的值作为管理面IP地址。如果采用多IP地址的方案，对“ipAddress”和“managementAddress”的初始值都需要做相应的修改。选填，默认值：“127.0.0.2”。 |
| port | int32\_t | [1024, 65535] | EndPoint提供的业务面RESTful接口绑定的端口号。如果采用物理机/宿主机IP地址通信，请自行保证端口号无冲突。必填，默认值：1025。 |
| managementPort | int32\_t | [1024, 65535] | EndPoint提供的管理面（管理面接口参考 [表1](https://link.csdn.net/?target=https%3A%2F%2Fwww.hiascend.com%2Fdocument%2Fdetail%2Fzh%2Fmindie%2F10RC2%2Fmindieservice%2Fservicedev%2Fmindie_service0035.html%23ZH-CN_TOPIC_0000001960000104__zh-cn_topic_0000001909743885_table1755073435611) ）接口绑定的端口号。业务面与管理面可采用四种方案：单IP地址单端口号（推荐）单IP地址多端口号多IP地址单端口号多IP地址多端口号在单卡节点中，不能使用多IP地址单端口号的方案，会因端口号占用而无法启动。选填，默认值：1026。 |
| maxLinkNum | uint32\_t | [1, 1000] | RESTful接口请求并发数，EndPoint支持的最大并发请求数。必填，默认值：1000。 |
| httpsEnabled | bool | truefalse | 是否开启https通信。true：开启https通信。false：关闭https通信。必填，默认值：true，建议值：true，取值为false时，忽略后续https通信相关参数。 |
| tlsCaPath | string | 建议tlsCaPath+tlsCaFile路径长度<=4096。实际路径为工程路径+tlsCaPath，上限限制与操作系统有关，最小值为1。 | 根证书路径，只支持软件包安装路径下的相对路径。“httpsEnabled”=true生效，生效后必填，默认值：“security/ca/”。 |
| tlsCaFile | set– [string](https://link.csdn.net/?target=string) | 建议tlsCaPath+tlsCaFile路径长度<=4096。不可为空，并且tlsCaPath+tlsCaFile路径长度上限与操作系统有关，最小值为1。 | 业务面根证书名称列表。“httpsEnabled”=true生效，生效后必填，默认值：[“ca.pem”]。 |
| tlsCert | string | 建议文件路径长度<=4096。实际路径为工程路径+tlsCert，上限限制与操作系统有关，最小值为1。 | 业务面服务证书文件路径，只支持软件包安装路径下的相对路径。“httpsEnabled”=true生效，生效后必填，默认值：“security/certs/server.pem”。 |
| tlsPk | string | 建议文件路径长度<=4096。实际路径为工程路径+tlsPk，上限限制与操作系统有关，最小值为1。 | 业务面服务证书私钥文件路径，证书私钥的长度要求>=3072，只支持软件包安装路径下的相对路径。“httpsEnabled”=true生效，生效后必填，默认值：“security/keys/server.key.pem”。 |
| tlsPkPwd | string | 文件路径长度<=4096。支持为空；若非空，则实际路径为工程路径+tlsPkPwd，上限限制与操作系统有关，最小值为1。 | 业务面服务证书私钥加密密钥文件路径，只支持软件包安装路径下的相对路径。“httpsEnabled”=true生效，生效后选填，默认值：“security/pass/key\_pwd.txt”。若私钥经过加密但是未提供此文件，系统启动时会要求用户在交互窗口输入私钥加密口令。 |
| tlsCrl | string | 建议文件路径长度<=4096。支持为空；若非空，则实际路径为工程路径+tlsCrl，上限限制与操作系统有关，最小值为1。 | 业务面服务证书吊销列表文件路径，只支持软件包安装路径下的相对路径。“httpsEnabled”=true生效，生效后必填，默认值：“security/certs/server\_crl.pem”。“httpsEnabled”=false不启用吊销列表。“tlsCrl”的值只能配套“tlsCaFile”文件列表中的第一个CA文件。 |
| managementTlsCaFile | set– [string](https://link.csdn.net/?target=string) | 建议tlsCaPath+managementTlsCaFile路径长度<=4096。不可为空，并且tlsCaPath+managementTlsCaFile路径长度上限与操作系统有关，最小值为1。 | 管理面根证书名称列表，当前管理面证书和业务面证书放在同一个路径（tlsCaPath）下。“httpsEnabled”=true且“ipAddress”!=“managementIpAddress”生效，生效后必填，默认值：[“management\_ca.pem”]。 |
| managementTlsCert | string | 建议文件路径长度<=4096。实际路径为工程路径+managementTlsCert，上限限制与操作系统有关，最小值为1。 | 管理面服务证书文件路径，只支持软件包安装路径下的相对路径。“httpsEnabled”=true且“ipAddress”!=“managementIpAddress”生效，生效后必填，默认值：“security/certs/management\_server.pem”。 |
| managementTlsPk | string | 建议文件路径长度<=4096。实际路径为工程路径+managementTlsPk，上限限制与操作系统有关，最小值为1。 | 管理面服务证书私钥文件路径，证书私钥的长度要求>=3072，只支持软件包安装路径下的相对路径。“httpsEnabled”=true且“ipAddress”!=“managementIpAddress”生效，生效后必填，默认值：“security/keys/management\_server.key.pem”。 |
| managementTlsPkPwd | string | 文件路径长度<=4096。支持为空；若非空，则实际路径为工程路径+managementTlsPkPwd，上限限制与操作系统有关，最小值为1 | 管理面服务证书私钥加密密钥文件路径，只支持软件包安装路径下的相对路径。“httpsEnabled”=true且“ipAddress”!=“managementIpAddress”生效，生效后选填，默认值：“security/pass/management/key\_pwd.txt”。若私钥经过加密但是未提供此文件，系统启动时会要求用户在交互窗口输入私钥加密口令。 |
| managementTlsCrl | string | 建议文件路径长度<=4096。支持为空；若非空，则实际路径为工程路径+managementTlsCrl，上限限制与操作系统有关，最小值为1。 | 管理面证书吊销列表文件路径，只支持软件包安装路径下的相对路径。“httpsEnabled”=true且“ipAddress”!=“managementIpAddress”生效，生效后必填，默认值：“security/certs/management\_server\_crl.pem”。“httpsEnabled”=false不启用吊销列表。“managementTlsCrl”的值只能配套“managementTlsCaFile”文件列表中的第一个CA文件。 |
| kmcKsMaster | string | 建议文件路径长度<=4096。实际路径为工程路径+kmcKsMaster，上限限制与操作系统有关，最小值为1。 | KMC密钥库文件路径，只支持软件包安装路径下的相对路径。“httpsEnabled”=true生效，生效后必填，默认值：“tools/pmt/master/ksfa”。 |
| kmcKsStandby | string | 建议文件路径长度<=4096。实际路径为工程路径+kmcKsStandby，上限限制与操作系统有关，最小值为1。 | KMC密钥库备份文件路径，只支持软件包安装路径下的相对路径。“httpsEnabled”=true生效，生效后必填，默认值：“tools/pmt/standby/ksfb”。 |
| multiNodesInferPort | uint32\_t | [1024, 65535] | 跨机通信的端口号，多机推理场景使用。选填，默认值：1120。 |
| interNodeTLSEnabled | bool | truefalse | 多机推理时，跨机通信是否开启证书安全认证。true：开启证书安全认证。false：关闭证书安全认证。选填，默认值：true。取值为false时，忽略后续参数。 |
| interNodeTlsCaFile | string | 建议文件路径长度<=4096。实际路径为工程路径+interNodeTlsCaFile，上限限制与操作系统有关，最小值为1。 | 根证书名称路径，只支持软件包安装路径下的相对路径。“interNodeTLSEnabled”=true生效，生效后必填，默认值：“security/ca/ca.pem”。 |
| interNodeTlsCert | string | 建议文件路径长度<=4096。实际路径为工程路径+interNodeTlsCert，上限限制与操作系统有关，最小值为1。 | 服务证书文件路径，只支持软件包安装路径下的相对路径。“interNodeTLSEnabled”=true生效，生效后必填，默认值：“security/certs/server.pem”。 |
| interNodeTlsPk | string | 建议文件路径长度<=4096。实际路径为工程路径+interNodeTlsPk，上限限制与操作系统有关，最小值为1。 | 服务证书私钥文件路径，只支持软件包安装路径下的相对路径。“interNodeTLSEnabled”=true生效，生效后必填，默认值：“security/keys/server.key.pem”。 |
| interNodeTlsPkPwd | string | 建议文件路径长度<=4096。支持为空；若非空，则实际路径为工程路径+interNodeTlsPkPwd，上限限制与操作系统有关，最小值为1。 | 服务证书私钥加密密钥文件路径，只支持软件包安装路径下的相对路径。“interNodeTLSEnabled”=true生效，生效后必填，默认值：“security/pass/mindie\_server\_key\_pwd.txt”。 |
| interNodeKmcKsfMaster | string | 建议文件路径长度<=4096。实际路径为工程路径+interNodeKmcKsfMaster，上限限制与操作系统有关，最小值为1。 | KMC密钥库文件路径，只支持软件包安装路径下的相对路径。“interNodeTLSEnabled”=true生效，生效后必填，默认值：“tools/pmt/master/ksfa”。 |
| interNodeKmcKsfStandby | string | 建议文件路径长度<=4096。实际路径为工程路径+interNodeKmcKsfStandby，上限限制与操作系统有关，最小值为1。 | KMC密钥库备份文件路径，只支持软件包安装路径下的相对路径。“interNodeTLSEnabled”=true生效，生效后必填，默认值：“tools/pmt/standby/ksfb”。 |

说明

* 如果网络环境不安全，不开启https通信，即“httpsEnabled”=“false”时，会存在较高的网络安全风险。
* 如果推理服务所在的计算节点的网络为跨公网和局域网，绑定0.0.0.0的IP地址可能导致网络隔离失效，存在较大安全风险。故该场景下禁止EndPoint的IP地址绑定为0.0.0.0。
* 如果配置了相同的管理面和业务面的IP地址，会导致隔离失效。

#### WorkFlowParam参数

| 配置项 | 取值类型 | 取值范围 | 配置说明 |
| --- | --- | --- | --- |
| TemplateParam |  |  |  |
| templateType | string | 当前取值只能为：“Standard” | 普通推理。必填，默认值：“Standard”。 |
| templateName | string | 由大写字母、小写字母和下划线组成，且不以下划线作为开头和结尾，字符串长度小于或等于256。 | 工作流名称。必填，默认值：“Standard\_llama”。 |

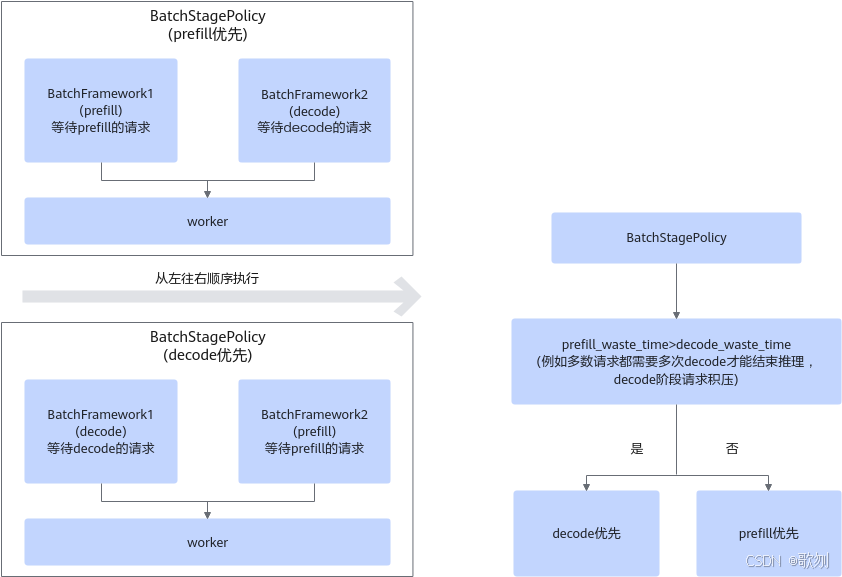
#### ModelDeployParam参数

| 配置项 | 取值类型 | 取值范围 | 配置说明 |
| --- | --- | --- | --- |
| engineName | string | 长度1~50，只支持小写字母加下划线。且不以下划线作为开头和结尾。 | 根据engineName找对应的so文件。必填，默认值：“mindieservice\_llm\_engine”。 |
| modelInstanceNumber | uint32\_t | [1, 10] | 模型实例个数。必填，默认值：1。 |
| tokenizerProcessNumber | uint32\_t | [1, 32] | tokenizer进程数。选填，默认值：8。 |
| maxSeqLen | uint32\_t | 上限根据显存和用户需求来决定，最小值需大于0。 | 最大序列长度。输入的长度+输出的长度<=maxSeqLen，用户根据自己的推理场景选择maxSeqLen。如果maxSeqLen大于模型支持的最大序列长度，可能会影响推理精度。必填，默认值：2560。 |
| npuDeviceIds | set- [set<size\_t](https://link.csdn.net/?target=set%253Csize_t) > | 根据模型和环境的实际情况来决定。 | 表示启用哪几张卡。对于每个模型实例分配的npuIds。多机推理场景下该值无效，每个节点上使用的npuDeviceIds根据ranktable计算获得。必填，默认值：[[0,1,2,3]]。 |
| multiNodesInferEnabled | bool | truefalse | false：单机推理。true：多机推理。选填，默认值：false。 |
| ModelParam |  |  |  |
| modelInstanceType | string | “Standard”“StandardMock” | 模型类型。“Standard”：普通推理。“StandardMock”：假模型。选填，默认值：“Standard”。 |
| modelName | string | 由大写字母、小写字母、数字、中划线、点和下划线组成，且不以中划线、点和下划线作为开头和结尾，字符串长度小于或等于256。 | 模型名称。必填，默认值：“llama\_65b”。 |
| modelWeightPath | string | 文件绝对路径长度的上限与操作系统有关，最小值为1。 | 模型权重路径。程序会读取该路径下的config.json中torch\_dtype和vocab\_size字段的值，需保证路径和相关字段存在。必填，默认值：“/data/atb\_testdata/weights/llama1-65b-safetensors”。该路径会进行安全校验，必须使用绝对路径，且和执行用户的属组和权限保持一致。 |
| worldSize | uint32\_t | 根据模型实际情况来决定。每一套模型参数中worldSize必须与使用的NPU数量相等。 | 启用几张卡推理。目前llama-65b至少启用四张NPU卡。多机推理场景下该值无效，worldSize根据ranktable计算获得。必填，默认值：4。 |
| cpuMemSize | uint32\_t | 上限根据显存和用户需求来决定。只有当maxPreemptCount为0时，才可以取值为0。 | CPU中可以用来申请kv cache的size上限。必填，默认值：5，建议值：5，单位：GB。 |
| npuMemSize | uint32\_t | 上限根据显存和用户需求来决定，下限大于0。 | NPU中可以用来申请kv cache的size上限。必填，默认值：8，建议值：8，单位：GB。快速计算公式：npuMemSize=(单卡总空闲-权重/NPU卡数-后处理占用)\*系数，其中系数取0.8。 |
| backendType | string | “atb”“ms” | 对接的后端类型。必填，默认值：“atb”。 |
| pluginParams | string | 根据并行解码实际所需填写一个json字符串。 | 选填，默认值：““。 |

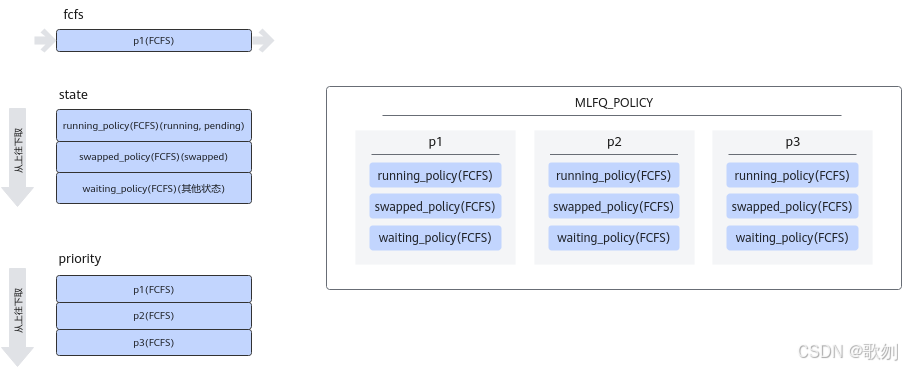
#### ScheduleParam参数

| 配置项 | 取值类型 | 取值范围 | 配置说明 |
| --- | --- | --- | --- |
| maxPrefillBatchSize | uint32\_t | [1, maxBatchSize] | 最大prefill batch size。maxPrefillBatchSize和maxPrefillTokens谁先达到各自的取值就完成本次组batch。该参数主要是在明确需要限制prefill阶段batch size的场景下使用，否则可以设置为0（此时引擎将默认取maxBatchSize值）或与maxBatchSize值相同。必填，默认值：50。 |
| maxPrefillTokens | uint32\_t | [5120, 512000]，且必须大于或等于maxSeqLen的取值。 | 每次prefill时，当前batch中所有input token总数，不能超过maxPrefillTokens。maxPrefillTokens和maxPrefillBatchSize谁先达到各自的取值就完成本次组batch。必填，默认值：8192。 |
| prefillTimeMsPerReq | uint32\_t | [0, 1000] | 与decodeTimeMsPerReq比较，计算当前应该选择prefill还是decode。单位：ms，当“supportSelectBatch”=true时有效。其调度策略流程图请参见 [图1](https://link.csdn.net/?target=https%3A%2F%2Fwww.hiascend.com%2Fdocument%2Fdetail%2Fzh%2Fmindie%2F10RC2%2Fmindieservice%2Fservicedev%2Fmindie_service0004.html%23ZH-CN_TOPIC_0000001996361197__fig178111591121) 。必填，默认值：150。 |
| prefillPolicyType | uint32\_t | 013 | prefill阶段的调度策略，其调度策略流程图请参见 [图2](https://link.csdn.net/?target=https%3A%2F%2Fwww.hiascend.com%2Fdocument%2Fdetail%2Fzh%2Fmindie%2F10RC2%2Fmindieservice%2Fservicedev%2Fmindie_service0004.html%23ZH-CN_TOPIC_0000001996361197__fig186221198317) 。0：FCFS，先来先服务。1：STATE，prefill阶段等同于FCFS策略。3：MLFQ，多级反馈队列。其中，3是0/1的组合。必填，默认值：0。 |
| decodeTimeMsPerReq | uint32\_t | [0, 1000] | 与prefillTimeMsPerReq比较，计算当前应该选择prefill还是decode。单位：ms，当“supportSelectBatch”=true时有效。其调度策略流程图请参见 [图1](https://link.csdn.net/?target=https%3A%2F%2Fwww.hiascend.com%2Fdocument%2Fdetail%2Fzh%2Fmindie%2F10RC2%2Fmindieservice%2Fservicedev%2Fmindie_service0004.html%23ZH-CN_TOPIC_0000001996361197__fig178111591121) 。必填，默认值：50。 |
| decodePolicyType | uint32\_t | 013 | decode阶段的调度策略。其调度策略流程图请参见 [图2](https://link.csdn.net/?target=https%3A%2F%2Fwww.hiascend.com%2Fdocument%2Fdetail%2Fzh%2Fmindie%2F10RC2%2Fmindieservice%2Fservicedev%2Fmindie_service0004.html%23ZH-CN_TOPIC_0000001996361197__fig186221198317) 。0：FCFS，先来先服务。1：STATE，decode阶段优先执行未被抢占和换出的请求。3：MLFQ，多级反馈队列。其中，3是0/1的组合。必填，默认值：0。 |
| maxBatchSize | uint32\_t | [1, 5000]，且必须大于或等于maxPreemptCount的取值。 | 最大decode batch size。首先计算block\_num：Total Block Num = Floor(NPU显存/(模型网络数cacheBlockSize模型注意力头数注意力头大小Cache类型字节数Cache数))，其中，Cache数=2；在tensor并行的情况下，block\_numworld\_size为实际的分配block数。如果是多卡，公式中的模型注意力头数注意力大小的值需要均摊在每张卡上，即“模型注意力头数注意力大小/卡数”。公式中的Floor表示计算结果向下取整。为每个请求申请的block数量Block Num=Ceil(输入Token数/Block Size)+Ceil(最大输出Token数/Block Size)。输入Token数：输入（字符串）做完tokenizer后的tokenID个数；最大输出Token数：模型推理最大迭代次数和最大输出长度之间取较小值。公式中的Ceil表示计算结果向上取整。maxBatchSize=Total Block Num/Block Num。必填，默认值：200。 |
| maxIterTimes | uint32\_t | [1, maxSeqLen-1] | 迭代次数，即一句话最大可生成长度。与允许推理生成的最大token个数max\_tokens（或max\_new\_tokens）取较小值作为最大可生成长度。必填，默认值：512。 |
| maxPreemptCount | int32\_t | [0, maxBatchSize]，当取值大于0时，cpuMemSize取值不可为0。 | 每一批次最大可抢占请求的上限，即限制一轮调度最多抢占请求的数量，最大上限为maxBatchSize，取值大于0则表示开启可抢占功能。必填，默认值：0。 |
| supportSelectBatch | bool | truefalse | batch选择策略。false：表示每一轮调度时，优先调度和执行prefill阶段的请求。true：表示每一轮调度时，根据当前prefill与decode请求的数量，自适应调整prefill和decode阶段请求调度和执行的先后顺序。必填，默认值：false。 |
| maxQueueDelayMicroseconds | uint32\_t | [500, 1000000] | 队列等待时间，单位：us。必填，默认值：5000。 |

#### 图1 调度策略和执行先后顺序流程图



#### 图2 prefill和decode阶段的调度策略流程图



### 启动模型

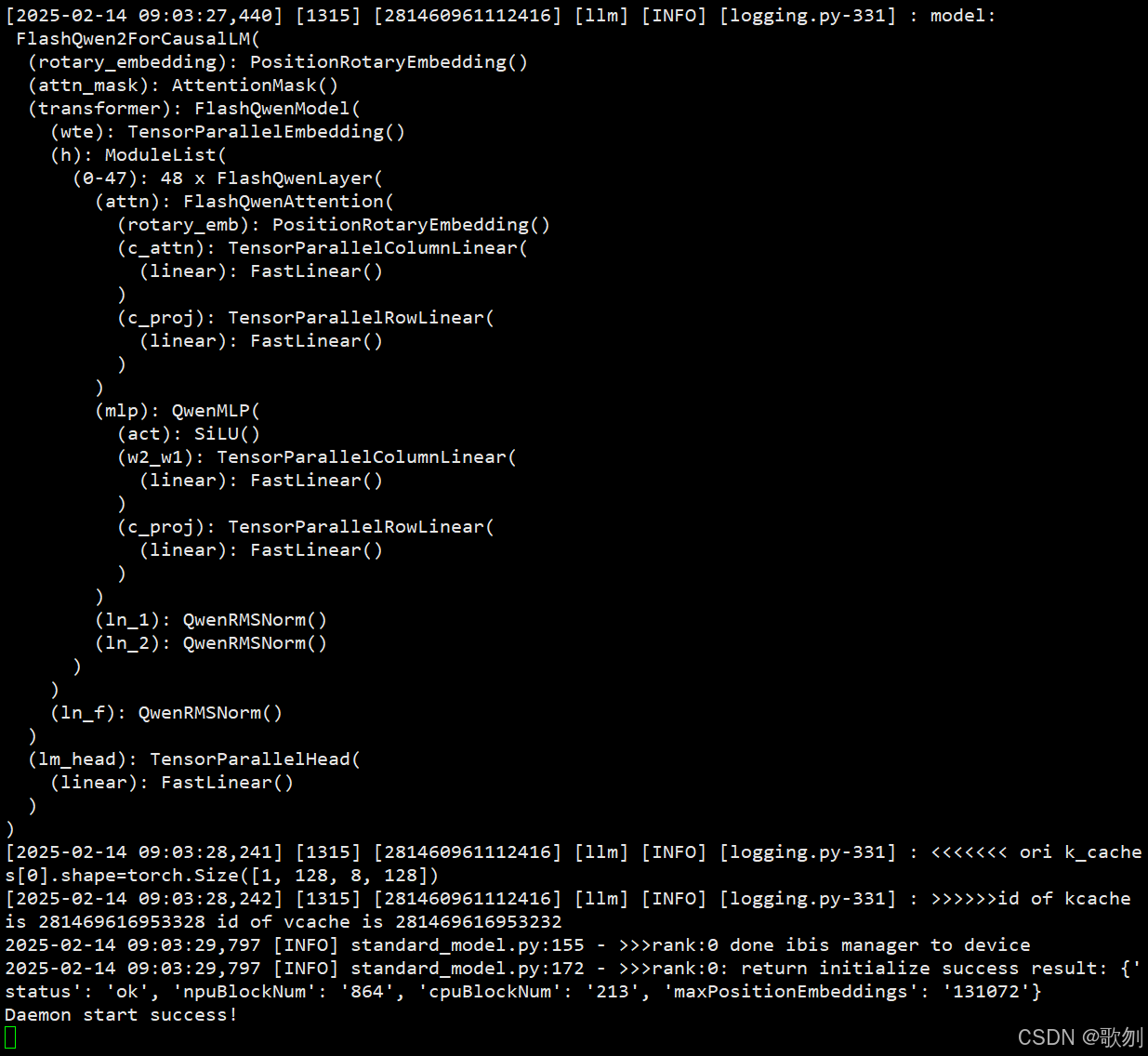
#### 拉起服务化接口

cd /usr/local/Ascend/mindie/latest/mindie-service/bin  
./mindieservice\_daemon

后台启动

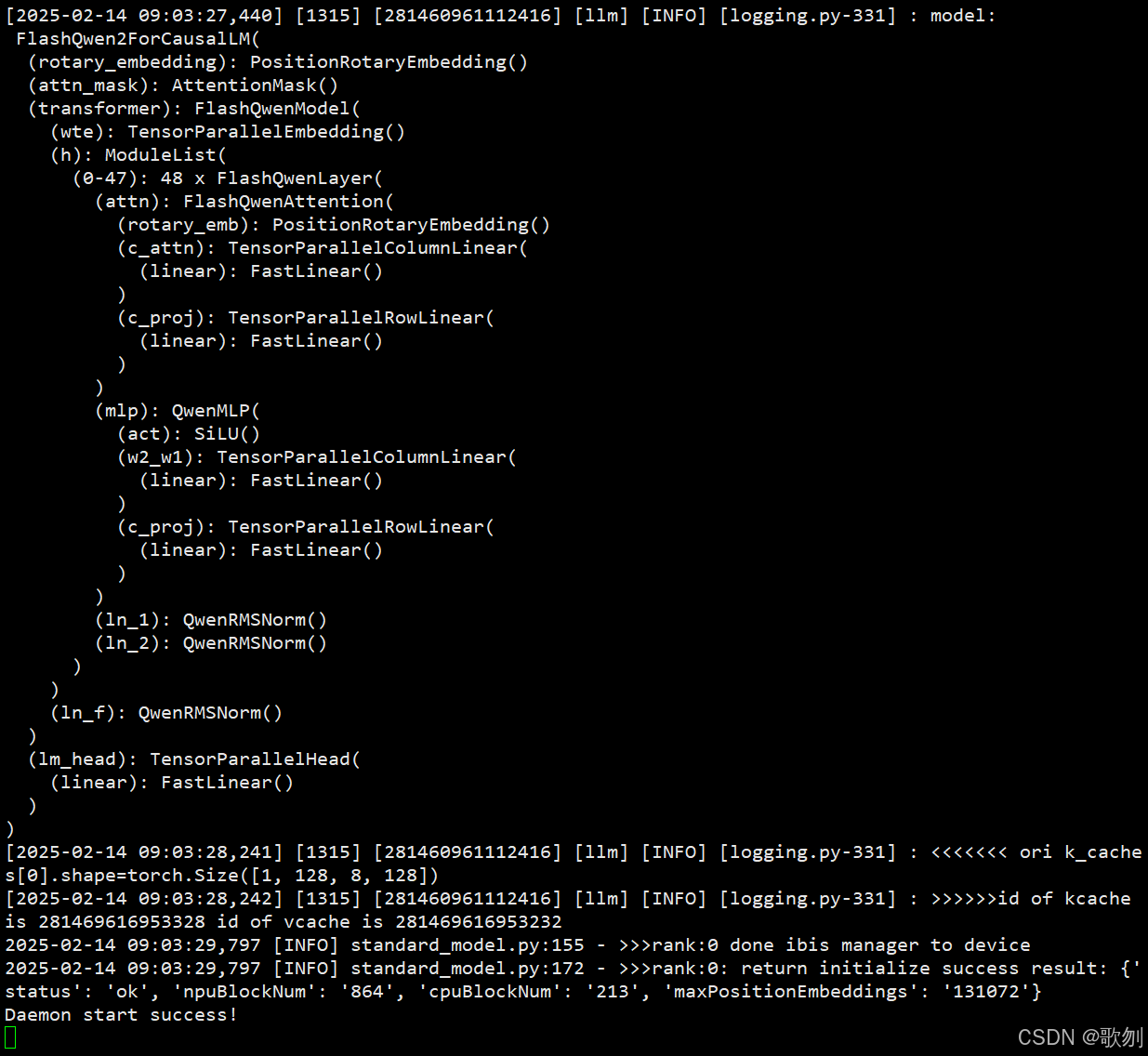
cd $MIES\_INSTALL\_PATH  
nohup ./bin/mindieservice\_daemon > output.log 2>&1 &  
tail -f output.log  
  
# nuhup 开启一个后台进程  
[1] 107  
# 杀死进程 kill 107

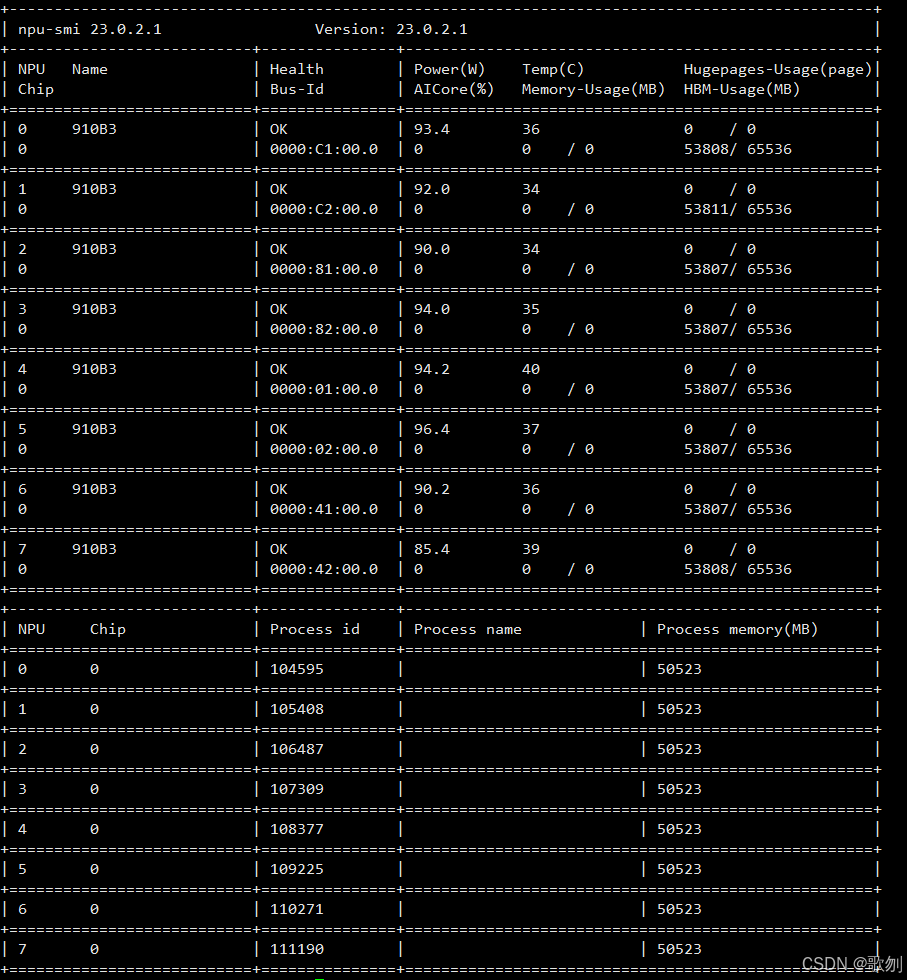
Daemon start success! 则为启动成功



#### 接口测试

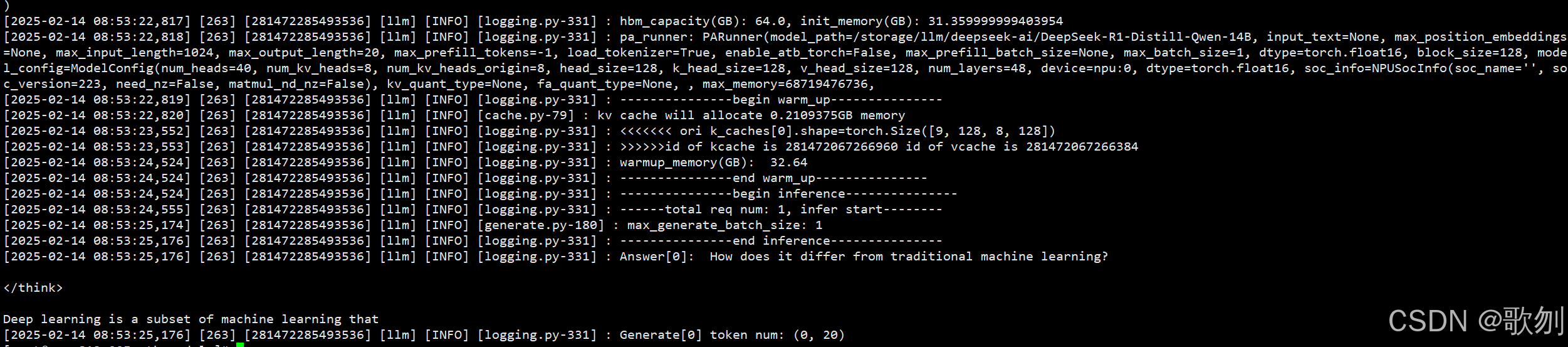
time curl -X POST http://127.0.0.1:1025/v1/chat/completions \  
 -H "Accept: application/json" \  
 -H "Content-type: application/json" \  
 -d '{  
 "model": "deepseek-14b",  
 "messages": [{  
 "role": "user",  
 "content": "我有五天假期，我想去海南玩，请给我一个攻略"  
 }],  
 "max\_tokens": 2048,  
 "presence\_penalty": 1.03,  
 "frequency\_penalty": 1.0,  
 "seed": null,  
 "temperature": 0.5,  
 "top\_p": 0.95,  
 "stream": false  
}'





#### 脚本测试(可选)

cd $ATB\_SPEED\_HOME\_PATH  
python examples/run\_pa.py --model\_path /storage/llm/deepseek-ai/DeepSeek-R1-Distill-Qwen-14B



### 并发测试

#### 14B - 单卡运行

八张卡 八个实例

| 设备 | 模型 | 上下文（输出+输出长度） | 并发 | 循环次数 | 并发请求总输出速率（tokens/s） | 单个请求速率的平均 | 请求超时个数（超过60s的请求） |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 910B3 \* 8 | deepseek-14b | 2048 | 1 | 1 | 33.03 | 33.03 | 0 |
| 910B3 \* 8 | deepseek-14b | 4096 | 1 | 1 | 31.91 | 31.91 | 0 |
| 910B3 \* 8 | deepseek-14b | 8192 | 1 | 1 | 31.08 | 31.08 | 0 |
| 910B3 \* 8 | deepseek-14b | 2048 | 4 | 1 | 101.63 | 32.93 | 0 |
| 910B3 \* 8 | deepseek-14b | 4096 | 4 | 1 | 113.43 | 31.98 | 0 |
| 910B3 \* 8 | deepseek-14b | 8192 | 4 | 1 | 94.86 | 31.11 | 0 |
| 910B3 \* 8 | deepseek-14b | 2048 | 8 | 1 | 213.05 | 31.99 | 0 |
| 910B3 \* 8 | deepseek-14b | 4096 | 8 | 1 | 185.08 | 30.09 | 0 |
| 910B3 \* 8 | deepseek-14b | 8192 | 8 | 1 | 154.08 | 29.45 | 0 |
| 910B3 \* 8 | deepseek-14b | 2048 | 16 | 1 | 284.99 | 31.00 | 0 |
| 910B3 \* 8 | deepseek-14b | 4096 | 16 | 1 | 279.60 | 29.44 | 0 |
| 910B3 \* 8 | deepseek-14b | 8192 | 16 | 1 | 346.78 | 27.92 | 0 |
| 910B3 \* 8 | deepseek-14b | 2048 | 32 | 1 | 579.98 | 29.46 | 0 |
| 910B3 \* 8 | deepseek-14b | 4096 | 32 | 1 | 575.08 | 26.80 | 0 |
| 910B3 \* 8 | deepseek-14b | 8192 | 32 | 1 | 560.29 | 24.92 | 0 |
| 910B3 \* 8 | deepseek-14b | 2048 | 64 | 1 | 932.15 | 24.59 | 0 |
| 910B3 \* 8 | deepseek-14b | 4096 | 64 | 1 | 1118.59 | 24.52 | 0 |
| 910B3 \* 8 | deepseek-14b | 8192 | 64 | 1 | 816.21 | 21.98 | 2 |
| 910B3 \* 8 | deepseek-14b | 2048 | 96 | 1 | 1294.45 | 25.58 | 0 |
| 910B3 \* 8 | deepseek-14b | 4096 | 96 | 1 | 1437.15 | 21.76 | 2 |
| 910B3 \* 8 | deepseek-14b | 8192 | 96 | 1 | 1291.17 | 18.78 | 5 |
| 910B3 \* 8 | deepseek-14b | 2048 | 128 | 1 | 1307.53 | 20.16 | 5 |
| 910B3 \* 8 | deepseek-14b | 4096 | 128 | 1 | 1560.00 | 16.81 | 28 |
| 910B3 \* 8 | deepseek-14b | 8192 | 128 | 1 | 1348.41 | 13.06 | 37 |
| 910B3 \* 8 | deepseek-14b | 2048 | 196 | 1 | 1417.76 | 12.30 | 77 |
| 910B3 \* 8 | deepseek-14b | 4096 | 196 | 1 | 404.82 | 2.84 | 171 |
| 910B3 \* 8 | deepseek-14b | 8192 | 196 | 1 | 521.18 | 2.94 | 162 |

#### 32B - 双卡并行

八张卡 四个实例

| 设备 | 模型 | 上下文（输出+输出长度） | 并发 | 循环次数 | 并发请求总输出速率（tokens/s） | 单个请求速率的平均 | 请求超时个数（超过60s的请求） |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 910B3 \* 8 | deepseek-32b | 2048 | 1 | 1 | 27.64 | 27.64 | 0 |
| 910B3 \* 8 | deepseek-32b | 4096 | 1 | 1 | 26.43 | 26.43 | 0 |
| 910B3 \* 8 | deepseek-32b | 8192 | 1 | 1 | 25.09 | 25.09 | 0 |
| 910B3 \* 8 | deepseek-32b | 2048 | 4 | 1 | 80.31 | 26.23 | 0 |
| 910B3 \* 8 | deepseek-32b | 4096 | 4 | 1 | 67.86 | 23.03 | 0 |
| 910B3 \* 8 | deepseek-32b | 8192 | 4 | 1 | 81.54 | 23.39 | 0 |
| 910B3 \* 8 | deepseek-32b | 2048 | 8 | 1 | 147.15 | 23.17 | 0 |
| 910B3 \* 8 | deepseek-32b | 4096 | 8 | 1 | 131.06 | 22.09 | 0 |
| 910B3 \* 8 | deepseek-32b | 8192 | 8 | 1 | 123.23 | 20.39 | 0 |
| 910B3 \* 8 | deepseek-32b | 2048 | 16 | 1 | 279.69 | 21.08 | 0 |
| 910B3 \* 8 | deepseek-32b | 4096 | 16 | 1 | 161.08 | 19.83 | 2 |
| 910B3 \* 8 | deepseek-32b | 8192 | 16 | 1 | 223.36 | 19.38 | 0 |
| 910B3 \* 8 | deepseek-32b | 2048 | 32 | 1 | 312.54 | 21.06 | 0 |
| 910B3 \* 8 | deepseek-32b | 4096 | 32 | 1 | 367.03 | 18.93 | 1 |
| 910B3 \* 8 | deepseek-32b | 8192 | 32 | 1 | 273.43 | 18.20 | 5 |
| 910B3 \* 8 | deepseek-32b | 2048 | 64 | 1 | 762.26 | 20.52 | 0 |
| 910B3 \* 8 | deepseek-32b | 4096 | 64 | 1 | 521.32 | 16.75 | 6 |
| 910B3 \* 8 | deepseek-32b | 8192 | 64 | 1 | 442.43 | 14.48 | 15 |
| 910B3 \* 8 | deepseek-32b | 2048 | 96 | 1 | 866.97 | 18.46 | 1 |
| 910B3 \* 8 | deepseek-32b | 4096 | 96 | 1 | 905.75 | 11.65 | 20 |
| 910B3 \* 8 | deepseek-32b | 8192 | 96 | 1 | 471.27 | 5.59 | 50 |
| 910B3 \* 8 | deepseek-32b | 2048 | 128 | 1 | 522.60 | 7.00 | 75 |
| 910B3 \* 8 | deepseek-32b | 4096 | 128 | 1 | 117.51 | 0.99 | 118 |
| 910B3 \* 8 | deepseek-32b | 8192 | 128 | 1 | 0.00 | 0.00 | 128 |
| 910B3 \* 8 | deepseek-32b | 2048 | 196 | 1 | 1345.32 | 14.69 | 91 |
| 910B3 \* 8 | deepseek-32b | 4096 | 196 | 1 | 925.30 | 11.43 | 146 |
| 910B3 \* 8 | deepseek-32b | 8192 | 196 | 1 | 755.33 | 9.04 | 166 |

#### 并发测试脚本代码

找一个文件夹，创建、执行并发请求脚本

1. python model\_request\_test.py
2. python statistic.py

* 新建请求脚本

vi model\_request\_test.py

# -\*- coding: utf-8 -\*-  
# @Time : 2025/2/14 14:29  
  
import os.path  
  
import asyncio  
import aiohttp  
import time  
import json  
import logging  
  
# 配置日志  
logging.basicConfig(level=logging.INFO, format='%(asctime)s - %(levelname)s - %(message)s')  
  
test\_context = """  
我们过了江，进了车站。我买票，他忙着照看行李。行李太多了，得向脚夫⑾行些小费才可过去。他便又忙着和他们讲价钱。我那时真是聪明过分，总觉他说话不大漂亮，非自己插嘴不可，但他终于讲定了价钱；就送我上车。他给我拣定了靠车门的一张椅子；我将他给我做的紫毛大衣铺好座位。他嘱我路上小心，夜里要警醒些，不要受凉。又嘱托茶房好好照应我。我心里暗笑他的迂；他们只认得钱，托他们只是白托！而且我这样大年纪的人，难道还不能料理自己么？我现在想想，我那时真是太聪明了。  
我说道：“爸爸，你走吧。”他往车外看了看，说：“我买几个橘子去。你就在此地，不要走动。”我看那边月台的栅栏外有几个卖东西的等着顾客。走到那边月台，须穿过铁道，须跳下去又爬上去。父亲是一个胖子，走过去自然要费事些。我本来要去的，他不肯，只好让他去。我看见他戴着黑布小帽，穿着黑布大马褂⑿，深青布棉袍，蹒跚⒀地走到铁道边，慢慢探身下去，尚不大难。可是他穿过铁道，要爬上那边月台，就不容易了。他用两手攀着上面，两脚再向上缩；他肥胖的身子向左微倾，显出努力的样子。这时我看见他的背影，我的泪很快地流下来了。我赶紧拭干了泪。怕他看见，也怕别人看见。我再向外看时，他已抱了朱红的橘子往回走了。过铁道时，他先将橘子散放在地上，自己慢慢爬下，再抱起橘子走。到这边时，我赶紧去搀他。他和我走到车上，将橘子一股脑儿放在我的皮大衣上。于是扑扑衣上的泥土，心里很轻松似的。过一会儿说：“我走了，到那边来信！”我望着他走出去。他走了几步，回过头看见我，说：“进去吧，里边没人。”等他的背影混入来来往往的人里，再找不着了，我便进来坐下，我的眼泪又来了。  
"""  
  
# 输入文本列表 此处为示例，实际测试请使用长度在 1800、3500、7000 字左右的文本作为输入 使得上下文长度在 2048、4096、8192 字左右  
input\_texts = [  
 test\_context \* (2048 // len(test\_context)) + "\n====\n总结以上文本为字数200字的摘要。",  
 test\_context \* (4096 // len(test\_context)) + "\n====\n总结以上文本为字数500字的摘要。",  
 test\_context \* (8192 // len(test\_context)) + "\n====\n总结以上文本为字数1000字的摘要。",  
]  
  
# 并发请求列表  
concurrency\_levels = [1, 4, 8, 16, 32, 64, 96, 128, 196]  
# concurrency\_levels = [96, 128, 196]  
# concurrency\_levels = [1]  
  
# 循环次数  
loop\_count = 1  
  
# 请求接口地址  
url = "http://127.0.0.1:1025/v1/chat/completions"  
  
# 设备和模型信息  
device = "910B3 \* 8"  
# model = "DeepSeek-R1-Distill-32B"  
model = "deepseek-14b" # 此处对应配置文件中的 ModelDeployConfig.ModelConfig.modelName  
  
if not os.path.exists(model):  
 os.mkdir(model)  
  
async def make\_request(session, input\_text):  
 # logging.info("开始单个请求")  
 headers = {  
 "Accept": "application/json",  
 "Content-type": "application/json"  
 }  
 data = {  
 "model": model,  
 "messages": [{"role": "user", "content": input\_text}],  
 "max\_tokens": 2048,  
 "presence\_penalty": 1.03,  
 "frequency\_penalty": 1.0,  
 "seed": None,  
 "temperature": 0.5,  
 "top\_p": 0.95,  
 "stream": True  
 }  
 start\_time = time.time()  
 try:  
 async with session.post(url, headers=headers, json=data, timeout=60) as response:  
 output\_tokens = 0  
 async for chunk in response.content.iter\_chunked(65535):  
 try:  
 chunk\_str = chunk.decode('utf-8').strip()  
 if chunk\_str.startswith("data: "):  
 chunk\_str = chunk\_str[len("data: "):]  
 chunk\_data = json.loads(chunk\_str)  
 """  
 data: {"id":"endpoint\_common\_34","object":"chat.completion.chunk","created":1739519727,  
 "model":"deepseek-32b",  
 "usage":{"prompt\_tokens":6,"completion\_tokens":27,"total\_tokens":33},"choices":[{"index":0,"delta":{"role":"assistant","content":""},"finish\_reason":"stop"}]}  
  
 """  
 output\_tokens += 1  
 except (json.JSONDecodeError, UnicodeDecodeError):  
 continue  
 end\_time = time.time()  
 elapsed\_time = end\_time - start\_time  
 print(f"elapsed\_time: 0.0614")  
 output\_rate = output\_tokens / elapsed\_time if elapsed\_time > 0 else 0  
 logging.info(  
 f"单个请求完成，输出 tokens: {output\_tokens}，耗时: {elapsed\_time:.2f}s，输出速率: {output\_rate:.2f} tokens/s")  
 return output\_tokens, elapsed\_time, output\_rate  
 except asyncio.TimeoutError:  
 logging.warning("单个请求超时")  
 return 0, 60, 0  
 except Exception as e:  
 print(f"ERROR: {e}")  
 return 0, 60, 0  
  
async def run\_concurrent\_tests(concurrency, input\_text):  
 logging.info(f"开始并发数为 {concurrency} 的测试")  
 async with aiohttp.ClientSession() as session:  
 tasks = [make\_request(session, input\_text) for \_ in range(concurrency)]  
 results = await asyncio.gather(\*tasks)  
 total\_output\_tokens = sum([result[0] for result in results])  
 total\_elapsed\_time = max([result[1] for result in results])  
 total\_output\_rate = total\_output\_tokens / total\_elapsed\_time if total\_elapsed\_time > 0 else 0  
 average\_single\_rate = sum([result[2] for result in results]) / concurrency  
 timeout\_count = sum([1 for result in results if result[1] >= 60])  
 logging.info(  
 f"并发数为 {concurrency} 的测试完成，总输出 tokens: {total\_output\_tokens}，总耗时: {total\_elapsed\_time:.2f}s，"  
 f"并发请求总输出速率: {total\_output\_rate:.2f} tokens/s，单个请求速率平均: {average\_single\_rate:.2f} tokens/s，超时个数: {timeout\_count}")  
 return total\_output\_rate, average\_single\_rate, timeout\_count  
  
async def main():  
 print(  
 "|设备|模型|上下文（输出+输出长度）|并发|循环次数|并发请求总输出速率（tokens/s）|单个请求速率的平均|请求超时个数（超过60s的请求）|")  
 print("| ------| ------| --------| ------| ----------| ------------------| ----------| --------------|")  
  
 for concurrency in concurrency\_levels:  
 all\_results = []  
 for i, input\_text in enumerate(input\_texts):  
 input\_length = len(input\_text)  
 total\_output\_rate, average\_single\_rate, timeout\_count = await run\_concurrent\_tests(concurrency, input\_text)  
 context = 2048 \* (2 \*\* i)  
 print(f"测试 {i + 1}/{len(input\_texts)} 完成，并发数为 {concurrency}，循环次数为 {loop\_count}")  
 result = {  
 "设备": device,  
 "模型": model,  
 "上下文（输出+输出长度）": context,  
 "并发": concurrency,  
 "循环次数": loop\_count,  
 "并发请求总输出速率（tokens/s）": total\_output\_rate,  
 "单个请求速率的平均": average\_single\_rate,  
 "请求超时个数（超过60s的请求）": timeout\_count  
 }  
 all\_results.append(result)  
 print(  
 f"|{device}|{model}|{context}|{concurrency}|{loop\_count}|{total\_output\_rate:.2f}|{average\_single\_rate:.2f}|{timeout\_count}|")  
  
 # 按并发数保存到 JSON 文件  
 filename = f'{model}/test\_results\_concurrency\_{concurrency}.json'  
 with open(filename, 'w', encoding='utf-8') as f:  
 json.dump(all\_results, f, ensure\_ascii=False, indent=4)  
 logging.info(f"并发数为 {concurrency} 的测试结果已保存到 {filename}")  
  
if \_\_name\_\_ == "\_\_main\_\_":  
 asyncio.run(main())

* 新建统计脚本

vi statistic.py

# -\*- coding: utf-8 -\*-  
# @Time : 2025/2/17 8:32  
import json  
  
# 并发请求列表  
concurrency\_levels = [1, 4, 8, 16, 32, 64, 96, 128, 196]  
version = 'deepseek-14b' # 此处对应配置文件中的 ModelDeployConfig.ModelConfig.modelName  
  
# 汇总所有结果  
all\_results = []  
for concurrency in concurrency\_levels:  
 filename = f'{version}/test\_results\_concurrency\_{concurrency}.json'  
 try:  
 with open(filename, 'r', encoding='utf-8') as f:  
 results = json.load(f)  
 all\_results.extend(results)  
 except FileNotFoundError:  
 print(f"未找到文件 {filename}，请确保之前的测试已成功保存结果。")  
  
# 生成 Markdown 表格表头  
markdown\_table = "|设备|模型|上下文（输出+输出长度）|并发|循环次数|并发请求总输出速率（tokens/s）|单个请求速率的平均|请求超时个数（超过60s的请求）|\n"  
markdown\_table += "| ------| ------| --------| ------| ----------| ------------------| ----------| --------------|\n"  
  
# 填充表格内容  
for result in all\_results:  
 markdown\_table += f"|{result['设备']}|{result['模型']}|{result['上下文（输出+输出长度）']}|{result['并发']}|{result['循环次数']}|{result['并发请求总输出速率（tokens/s）']:.2f}|{result['单个请求速率的平均']:.2f}|{result['请求超时个数（超过60s的请求）']}|\n"  
  
# 输出 Markdown 表格  
print(markdown\_table)  
  
# 保存 Markdown 表格到文件  
with open(f'{version}/summary\_table.md', 'w', encoding='utf-8') as f:  
 f.write(markdown\_table)

### 报错

#### Operation not permitted

[root@pm-a813-005 bin]# ./mindieservice\_daemon  
terminate called after throwing an instance of 'system\_error'  
 what(): Operation not permitted

模型路径的权限设置错误

chmod -R 755 /path-to-weights

参考 [mindie/README.md · Ascend/ascend-docker-image - Gitee.com](https://link.csdn.net/?target=https%3A%2F%2Fgitee.com%2Fascend%2Fascend-docker-image%2Fblob%2Fdev%2Fmindie%2FREADME.md%23%25E5%2590%25AF%25E5%258A%25A8%25E5%2591%25BD%25E4%25BB%25A4)

#### ConnectionRefusedError

[root@pm-a813-005 /]# vi /usr/local/Ascend/mindie/latest/mindie-service/conf/config.json  
[root@pm-a813-005 /]# cd /usr/local/Ascend/mindie/latest/mindie-service/bin  
[root@pm-a813-005 bin]# ./mindieservice\_daemon  
...  
Traceback (most recent call last):  
 File "/usr/lib64/python3.11/multiprocessing/process.py", line 314, in \_bootstrap  
 self.run()  
 File "/usr/lib64/python3.11/multiprocessing/process.py", line 108, in run  
 self.\_target(\*self.\_args, \*\*self.\_kwargs)  
 File "/usr/local/Ascend/ascend-toolkit/latest/python/site-packages/tbe/common/repository\_manager/route.py", line 71, in wrapper  
 raise exp  
 File "/usr/local/Ascend/ascend-toolkit/latest/python/site-packages/tbe/common/repository\_manager/route.py", line 63, in wrapper  
 func(\*args, \*\*kwargs)  
 File "/usr/local/Ascend/ascend-toolkit/latest/python/site-packages/tbe/common/repository\_manager/route.py", line 268, in task\_distribute  
 key, func\_name, detail = resource\_proxy[TASK\_QUEUE].get()  
 ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^  
 File "<string>", line 2, in get  
 File "/usr/lib64/python3.11/multiprocessing/managers.py", line 822, in \_callmethod  
 kind, result = conn.recv()  
 ^^^^^^^^^^^  
 File "/usr/lib64/python3.11/multiprocessing/connection.py", line 250, in recv  
 buf = self.\_recv\_bytes()  
 ^^^^^^^^^^^^^^^^^^  
 File "/usr/lib64/python3.11/multiprocessing/connection.py", line 430, in \_recv\_bytes  
 buf = self.\_recv(4)  
 ^^^^^^^^^^^^^  
 File "/usr/lib64/python3.11/multiprocessing/connection.py", line 395, in \_recv  
 chunk = read(handle, remaining)  
 ^^^^^^^^^^^^^^^^^^^^^^^  
ConnectionResetError: [Errno 104] Connection reset by peer  
/usr/lib64/python3.11/multiprocessing/resource\_tracker.py:254: UserWarning: resource\_tracker: There appear to be 30 leaked semaphore objects to clean up at shutdown  
 warnings.warn('resource\_tracker: There appear to be %d '  
/usr/lib64/python3.11/multiprocessing/resource\_tracker.py:254: UserWarning: resource\_tracker: There appear to be 30 leaked semaphore objects to clean up at shutdown  
 warnings.warn('resource\_tracker: There appear to be %d '  
Daemon is killing...  
Killed

当前镜像 和 宿主机服务器的 驱动版本不对应，前往官网换个镜像

#### 引用pytorch

警告而已，影响不大

[root@pm-a813-005 atb-models]# python  
Python 3.11.6 (main, Nov 27 2024, 18:16:08) [GCC 12.3.1 (openEuler 12.3.1-38.oe2403)] on linux  
Type "help", "copyright", "credits" or "license" for more information.  
>>> import torch  
  
>>> import torch\_npu  
/usr/local/lib64/python3.11/site-packages/torch\_npu/\_\_init\_\_.py:248: UserWarning: On the interactive interface, the value of TASK\_QUEUE\_ENABLE is set to 0 by default. Do not set it to 1 to prevent some unknown errors  
 warnings.warn("On the interactive interface, the value of TASK\_QUEUE\_ENABLE is set to 0 by default. \  
>>>

#### the size of npuDeviceIds (subset) does not equal to worldSize

the size of npuDeviceIds (subset) does not equal to worldSize  
ERR: Failed to init endpoint! Please check the service log or console output.  
Killed

此错误表明 npuDeviceIds（可能是 NPU 设备 ID 的子集）的数量与 worldSize 不匹配。在分布式计算的场景下，worldSize 通常代表参与计算的所有进程或设备的总数，而 npuDeviceIds 则是指定要使用的 NPU 设备的 ID 列表。当这两者的数量不一致时，就会触发该错误。

vi /usr/local/Ascend/mindie/latest/mindie-service/conf/config.json

...  
"BackendConfig" : {  
 "backendName" : "mindieservice\_llm\_engine",  
 "modelInstanceNumber" : 1,  
 "npuDeviceIds" : [[0]],  
  
....  
"ModelConfig" : [  
 {  
 "modelInstanceType" : "Standard",  
 "modelName" : "deepseek-14b",  
 "modelWeightPath" : "/storage/llm/deepseek-ai/DeepSeek-R1-Distill-Qwen-14B",  
 "worldSize" : 1, # 此处数量要与 npuDeviceIds 一致  
....

#### Failed to init endpoint

The serverConfig.kmcKsfMaster path is invalid by: The input file: ksfa is not a regular file or not exists  
The serverConfig.kmcKsfStandby path is invalid by: The input file: ksfb is not a regular file or not exists  
The serverConfig\_.tlsCert path is invalid by: The input file: server.pem is not a regular file or not exists  
ERR: serverConfig\_.tlsCrlFiles file not exit .  
The serverConfig\_.tlsCaFile path is invalid by: The input file: ca.pem is not a regular file or not exists  
The serverConfig\_.tlsPk path is invalid by: The input file: server.key.pem is not a regular file or not exists  
The serverConfig\_.tlsPkPwd path is invalid by: The input file: key\_pwd.txt is not a regular file or not exists  
The ServerConfig.managementTlsCert path is invalid by: The input file: server.pem is not a regular file or not exists  
The ServerConfig.managementTlsCrlPath path is not a dir by:   
ERR: serverConfig\_.managementTlsCrlFiles file not exit .  
ERR: serverConfig\_.managementTlsCaFile file not exit .  
The ServerConfig.managementTlsPk path is invalid by: The input file: server.key.pem is not a regular file or not exists  
The ServerConfig.managementTlsPkPwd path is invalid by: The input file: key\_pwd.txt is not a regular file or not exists  
ERR: Failed to init endpoint! Please check the service log or console output.  
Killed

解决方法就是取消https 启动服务

vi /usr/local/Ascend/mindie/latest/mindie-service/conf/config.json

"ServerConfig" :  
 {  
 "ipAddress" : "127.0.0.1",  
 "managementIpAddress" : "127.0.0.2",  
 "port" : 1025,  
 "managementPort" : 1026,  
 "metricsPort" : 1027,  
 "allowAllZeroIpListening" : false,  
 "maxLinkNum" : 1000,  
 "httpsEnabled" : false, # 设置为 false 不是用https  
。。。

ERR: Failed to init endpoint! Please check the service log or console output. Killed

实例启动太多 共享内存不够，导致日志写入失败

解决： 减少实例数量 或者增加 增加共享内存空间  
docker run…. –shm-size=10g

### 其他查询指令

#### 系统架构

uname -m

#### NPU 信息

npu-smi info

#### CANN 版本

* x86

cat /usr/local/Ascend/ascend-toolkit/latest/x86\_64-linux/ascend\_toolkit\_install.info

* arm

cat /usr/local/Ascend/ascend-toolkit/latest/arm64-linux/ascend\_toolkit\_install.info

[root@pm-a813-005 /]# cat /usr/local/Ascend/ascend-toolkit/latest/arm64-linux/ascend\_toolkit\_install.info  
package\_name=Ascend-cann-toolkit  
version=8.0.0  
innerversion=V100R001C20SPC001B251  
compatible\_version=[V100R001C15],[V100R001C17],[V100R001C18],[V100R001C19],[V100R001C20]  
arch=aarch64  
os=linux  
path=/usr/local/Ascend/ascend-toolkit/8.0.0/aarch64-linux

### 基础环境搭建

#### 获取CANN&MindIE安装包&环境准备

* [Atlas 800I A2/Atlas 300I Duo/Atlas 300 V](https://link.csdn.net/?target=https%3A%2F%2Fwww.hiascend.com%2Fdeveloper%2Fdownload%2Fcommunity%2Fresult%3Fmodule%3Dpt%2Bie%2Bcann)
* [环境准备指导](https://link.csdn.net/?target=https%3A%2F%2Fwww.hiascend.com%2Fdocument%2Fdetail%2Fzh%2FCANNCommunityEdition%2F80RC2alpha002%2Fsoftwareinst%2Finstg%2Finstg_0001.html)

#### CANN安装

# 增加软件包可执行权限，{version}表示软件版本号，{arch}表示CPU架构，{soc}表示昇腾AI处理器的版本。  
chmod +x ./Ascend-cann-toolkit\_{version}\_linux-{arch}.run  
chmod +x ./Ascend-cann-kernels-{soc}\_{version}\_linux.run  
# 校验软件包安装文件的一致性和完整性  
./Ascend-cann-toolkit\_{version}\_linux-{arch}.run --check  
./Ascend-cann-kernels-{soc}\_{version}\_linux.run --check  
# 安装  
./Ascend-cann-toolkit\_{version}\_linux-{arch}.run --install  
./Ascend-cann-kernels-{soc}\_{version}\_linux.run --install  
  
# 设置环境变量  
source /usr/local/Ascend/ascend-toolkit/set\_env.sh

#### MindIE安装

# 增加软件包可执行权限，{version}表示软件版本号，{arch}表示CPU架构。  
chmod +x ./Ascend-mindie\_${version}\_linux-${arch}.run  
./Ascend-mindie\_${version}\_linux-${arch}.run --check  
  
# 方式一：默认路径安装  
./Ascend-mindie\_${version}\_linux-${arch}.run --install  
# 设置环境变量  
cd /usr/local/Ascend/mindie && source set\_env.sh  
  
# 方式二：指定路径安装  
./Ascend-mindie\_${version}\_linux-${arch}.run --install-path=${AieInstallPath}  
# 设置环境变量  
cd ${AieInstallPath}/mindie && source set\_env.sh

#### Torch\_npu安装

下载 pytorch\_v{pytorchversion}\_py{pythonversion}.tar.gz

tar -xzvf pytorch\_v{pytorchversion}\_py{pythonversion}.tar.gz  
# 解压后，会有whl包  
pip install torch\_npu-{pytorchversion}.xxxx.{arch}.whl

### 相关链接

* [模型库-ModelZoo-昇腾社区](https://link.csdn.net/?target=https%3A%2F%2Fwww.hiascend.com%2Fsoftware%2Fmodelzoo%2Fmodels%3Fmodel%3Dr1)
* [模型库-魔搭社区](https://link.csdn.net/?target=https%3A%2F%2Fwww.modelscope.cn%2Fmodels%3Fname%3DDeepSeek-R1-Distill%26amp%3Bpage%3D1)
* [https://modelers.cn/MindIE](https://link.csdn.net/?target=%E6%A8%A1%E5%9E%8B%E5%BA%93-%E7%94%B5%E4%BF%A1%E9%AD%94%E4%B9%90%E7%A4%BE%E5%8C%BA)
* [modelscope 魔搭社区模型下载](https://link.csdn.net/?target=https%3A%2F%2Fblog.csdn.net%2FMnivL%2Farticle%2Fdetails%2F145465800%3Fspm%3D1001.2014.3001.5501)
* [mindie/README.md · Ascend/ascend-docker-image - Gitee.com](https://link.csdn.net/?target=https%3A%2F%2Fgitee.com%2Fascend%2Fascend-docker-image%2Fblob%2Fdev%2Fmindie%2FREADME.md%23%25E5%2590%25AF%25E5%258A%25A8%25E5%2591%25BD%25E4%25BB%25A4)
* [配置参数说明-快速开始-MindIE Service开发指南-服务化集成部署-MindIE1.0.RC2开发文档-昇腾社区](https://link.csdn.net/?target=https%3A%2F%2Fwww.hiascend.com%2Fdocument%2Fdetail%2Fzh%2Fmindie%2F10RC2%2Fmindieservice%2Fservicedev%2Fmindie_service0004.html%23ZH-CN_TOPIC_0000001996361197__fig178111591121)
* [单机推理-配置MindIE Server-配置MindIE-MindIE安装指南-环境准备-MindIE1.0.RC2开发文档-昇腾社区](https://link.csdn.net/?target=https%3A%2F%2Fwww.hiascend.com%2Fdocument%2Fdetail%2Fzh%2Fmindie%2F10RC2%2Fenvdeployment%2Finstg%2Fmindie_instg_0025.html%23ZH-CN_TOPIC_0000001996360273__zh-cn_topic_0000001909624101_section660017365231)
* [DeepSeek-R1-Distill-Qwen-32B-模型库-ModelZoo-昇腾社区](https://link.csdn.net/?target=https%3A%2F%2Fwww.hiascend.com%2Fsoftware%2Fmodelzoo%2Fmodels%2Fdetail%2Fed4a8911f1a74613a82dc50b1661208c)
* [mindie镜像版本下载](https://link.csdn.net/?target=https%3A%2F%2Fwww.hiascend.com%2Fdeveloper%2Fascendhub%2Fdetail%2Faf85b724a7e5469ebd7ea13c3439d48f)
* [Altas产品查询CANN软件包版本的方法 - 华为](https://link.csdn.net/?target=https%3A%2F%2Fsupport.huawei.com%2Fenterprise%2Fzh%2Fknowledge%2FEKB1100100366)
* [npu-smi命令介绍（适用于1.0.11-1.0.15版本） - Atlas 300I 推理卡 用户指南（型号 3010）31 - 华为](https://link.csdn.net/?target=https%3A%2F%2Fsupport.huawei.com%2Fenterprise%2Fzh%2Fdoc%2FEDOC1100079287%2F10dcd668)
* [昇腾模型库](https://link.csdn.net/?target=https%3A%2F%2Fwww.hiascend.com%2Fsoftware%2Fmodelzoo)
* [MindIE官方文档](https://link.csdn.net/?target=https%3A%2F%2Fwww.hiascend.com%2Fdocument%2Fdetail%2Fzh%2Fmindie)
* [MindIE Service开发指南](https://link.csdn.net/?target=https%3A%2F%2Fwww.hiascend.com%2Fdocument%2Fdetail%2Fzh%2Fmindie%2F10RC3%2Fmindieservice%2Fservicedev%2Fmindie_service0001.html)
* [DeepSeek-R1模型卡片](https://link.csdn.net/?target=https%3A%2F%2Fhuggingface.co%2Fdeepseek-ai%2FDeepSeek-R1)
* [DeepSeek模型量化方法介绍](https://link.csdn.net/?target=https%3A%2F%2Fgitee.com%2Fascend%2Fmsit%2Ftree%2Fbr_noncom_MindStudio_8.0.0_POC_20251231%2Fmsmodelslim%2Fexample%2FDeepSeek)
* [昇腾镜像仓库](https://link.csdn.net/?target=https%3A%2F%2Fascend.huawei.com%2Fascendhub)
* [量化技术白皮书](https://link.csdn.net/?target=https%3A%2F%2Fgitee.com%2Fascend%2Fmsit)
* [Ascend/ModelZoo-PyTorch](https://link.csdn.net/?target=https%3A%2F%2Fgitee.com%2Fascend%2FModelZoo-PyTorch.git)
* [在线推理过程中可使用的环境变量配置](https://link.csdn.net/?target=https%3A%2F%2Fwww.hiascend.com%2Fdocument%2Fdetail%2Fzh%2FPytorch%2F600%2Fapiref%2FEnvvariables%2FEnvir_009.html)
* [昇腾社区资源下载](https://link.csdn.net/?target=https%3A%2F%2Fwww.hiascend.com%2Fdeveloper%2Fdownload%2Fcommunity%2Fresult%3Fmodule%3Dpt%2Bie%2Bcann)
* [CANN环境准备指导](https://link.csdn.net/?target=https%3A%2F%2Fwww.hiascend.com%2Fdocument%2Fdetail%2Fzh%2FCANNCommunityEdition%2F80RC2alpha002%2Fsoftwareinst%2Finstg%2Finstg_0001.html)
* DeepSeek 系列模型 华为昇腾官方教程 | 模型名称 | 安装教程 | | — | — | | DeepSeek V3 | [DeepSeek-V3-模型库-ModelZoo-昇腾社区](https://link.csdn.net/?target=https%3A%2F%2Fwww.hiascend.com%2Fsoftware%2Fmodelzoo%2Fmodels%2Fdetail%2F678bdeb4e1a64c9dae51d353d84ddd15) | | DeepSeek R1 | [DeepSeek-R1-模型库-ModelZoo-昇腾社区](https://link.csdn.net/?target=https%3A%2F%2Fwww.hiascend.com%2Fsoftware%2Fmodelzoo%2Fmodels%2Fdetail%2F68457b8a51324310aad9a0f55c3e56e3) | | DeepSeek-R1-Distill-Qwen-1.5B | [DeepSeek-R1-Distill-Qwen-1.5B-模型库-ModelZoo-昇腾社区](https://link.csdn.net/?target=https%3A%2F%2Fwww.hiascend.com%2Fsoftware%2Fmodelzoo%2Fmodels%2Fdetail%2Fb54367ef746f484188576c39f5706899) | | DeepSeek-R1-Distill-Qwen-7B | [DeepSeek-R1-Distill-Qwen-7B-模型库-ModelZoo-昇腾社区](https://link.csdn.net/?target=https%3A%2F%2Fwww.hiascend.com%2Fsoftware%2Fmodelzoo%2Fmodels%2Fdetail%2F11aa2a48479d4d229a9830b8e41fc011) | | DeepSeek-R1-Distill-Llama-8B | [DeepSeek-R1-Distill-Llama-8B-模型库-ModelZoo-昇腾社区](https://link.csdn.net/?target=https%3A%2F%2Fwww.hiascend.com%2Fsoftware%2Fmodelzoo%2Fmodels%2Fdetail%2F8625241d99c64b528f1ff4dba7defb26) | | DeepSeek-R1-Distill-Qwen-14B | [DeepSeek-R1-Distill-Qwen-14B-模型库-ModelZoo-昇腾社区](https://link.csdn.net/?target=https%3A%2F%2Fwww.hiascend.com%2Fsoftware%2Fmodelzoo%2Fmodels%2Fdetail%2F2a0d0cb1bc644eee8318a12429de67b6) | | DeepSeek-R1-Distill-Qwen-32B | [DeepSeek-R1-Distill-Qwen-32B-模型库-ModelZoo-昇腾社区](https://link.csdn.net/?target=https%3A%2F%2Fwww.hiascend.com%2Fsoftware%2Fmodelzoo%2Fmodels%2Fdetail%2Fed4a8911f1a74613a82dc50b1661208c) | | DeepSeek-R1-Distill-Llama-70B | [DeepSeek-R1-Distill-Llama-70B-模型库-ModelZoo-昇腾社区](https://link.csdn.net/?target=https%3A%2F%2Fwww.hiascend.com%2Fsoftware%2Fmodelzoo%2Fmodels%2Fdetail%2Fee3f9897743a4341b43710f8d204733a) | | Janus-Pro-7B | [Janus-Pro-模型库-ModelZoo-昇腾社区](https://link.csdn.net/?target=https%3A%2F%2Fwww.hiascend.com%2Fsoftware%2Fmodelzoo%2Fmodels%2Fdetail%2Fffe1a0f4e8ba43aeb989251a3f0308e9) |

[松山湖开发者村综合服务平台](https://community.sslcode.com.cn/)

助力广东及东莞地区开发者，代码托管、在线学习与竞赛、技术交流与分享、资源共享、职业发展，成为松山湖开发者首选的工作与学习平台

* 9805
* 33
* 0
* 分享

回到  
顶部