解决方案实践

昇腾云按需算力部署 AI 模型

文档版本	1.0.0
发布日期	2026-09-04

华为云计算技术有限公司

商标声明


和其他华为商标均为华为技术有限公司的商标。
本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受华为云计算技术有限公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为云计算技术有限公司对本文档内容不做任何明示或暗示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

目 录

1 方案概述

应用场景

该解决方案旨在帮助您快速搭建使用**魔坊（ModelArts）模型训推平台**在线推理的基础环境。项目利用按需计算资源，依次完成模型权重获取、镜像准备、部署脚本脚本配置。

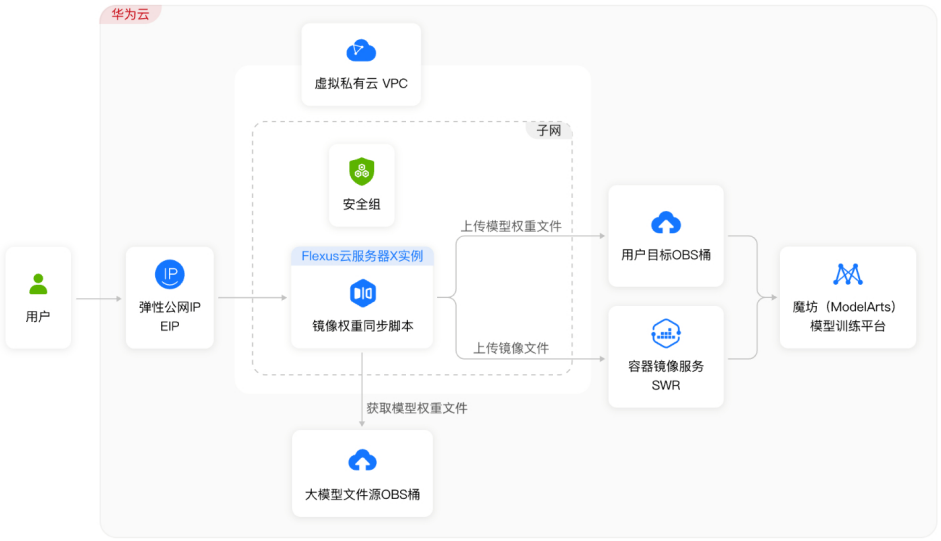
适用于如下场景：

1. **智能客服：**自动应答常见问题，处理退换货、物流查询，降低人工成本。
2. **内容创作：**协助撰写邮件、报告、文案、稿件等，提供写作灵感。
3. **信息总结**：快速提炼文档、会议记录、合同的核心要点，提升处理效率。
4. **代码开发：**生成代码片段、解释逻辑、辅助排查错误。
5. **语言翻译**：支持多语种互译，并润色译文使其更地道。
6. **教育培训**：解释概念、解答疑问、生成练习题，提供个性化辅导。

方案架构

该解决方案的云服务器作为中转服务器，**下载模型权重并上传至OBS，并下载推理镜像并推送至SWR**。后续通过魔坊（ModelArts）模型训推平台部署推理服务。

图 1-1 方案架构图



该解决方案将会部署如下资源：

- 创建一台**Flexus云服务器X实例**，用于模型权重同步、推理镜像拉取和推送至SWR。
- 创建一个**弹性公网IP EIP**，提供访问公网和被公网访问能力。
- 创建一个私有**对象存储服务 OBS桶**，用于保存模型权重和脚本。
- 创建一个**容器镜像服务 SWR**组织，用于保存推理镜像。

方案优势

- 一键部署，免人工干预

  一键轻松部署，快速完成华为云资源创建。部署完成后，即可直接获取模型部署所需的权重、镜像及启动脚本。
- OBS高速同步权重

  基于华为OBS高速传输自动拉取模型权重，相比手动上传显著提升同步效率，模型准备时间从小时级压缩至分钟级。
- 端到端自动化闭环

  覆盖模型上传、镜像准备到环境配置全流程，无需人工介入，无需精通底层技术。

约束与限制

- 支持Region：西南-贵阳一、华北-乌兰察布一。
- 该解决方案部署前，需注册华为账号并开通华为云，完成实名认证，且账号不能处于欠费或冻结状态。如果计费模式选择“包年包月”，请确保账户余额充足以便一键部署资源的时候可以自动支付；或者在一键部署的过程进入**费用中心**，找到“待支付订单”并手动完成支付。

2 资源规划

该解决方案主要部署如下资源，不同产品的花费仅供参考，实际以收费账单为准，具体请参考华为云**官网价格**。“请注意，此处列出的资源成本规划仅涵盖‘快速部署’阶段产生的资源及相关费用。后续在‘魔坊（ModelArts）模型训推平台’上实际使用服务时，将产生额外的计费项目。”

表 2-1 资源和成本规划(按需计费)

华为云服务	配置示例	数量	一次完整流程预估（1小时）花费
Flexus云服务器X实例	• 区域：西南-贵阳一、华北-乌兰察布一 • 计费模式：按需计费 • 规格：Flexus云服务器X实例 | 性能模式（开启）| kx1.4u.8g | 4核 | 8 GB • 镜像：Huawei Cloud EulerOS 2.0 Standard 64 bit for ARM(10GiB) • 系统盘：通用型SSD | 100GB	1台	西南-贵阳一：0.387元 华北-乌兰察布一：0.401元
弹性公网IP EIP	• 按需计费：0.8元/GB • 区域：西南-贵阳一、华北-乌兰察布一 • 计费模式：按需计费 • 线路：动态BGP • 公网带宽：按流量计费 • 带宽大小：300Mbit/s	1个	0.8元/GB

华为云服务	配置示例	数量	一次完整流程预估（1小时）花费
对象存储服务 OBS	• 桶策略：私有 • 数据冗余存储策略：单AZ存储 • 区域：西南-贵阳一、华北-乌兰察布一	1个	费用包括存储空间、请求费用、流量费用两部分，详细请参考每月账单。
合计	-		西南-贵阳一：0.387元+ 弹性公网IP EIP费用+OBS实际使用费用 华北-乌兰察布一：0.401元+ 弹性公网IP EIP费用+OBS实际使用费用

该解决方案支持一键部署的模型列表：

表 2-2 支持的模型及其最小卡数和最大序列（6.5.908 版本）

序号	模型分类	模型名称	是否支持fp16/bf16推理	是否支持W8A8量化	v0/v1后端	最小卡数（64G显存）	最大序列(K) max-model-len	开源权重获取地址
1	大语言模型	Qwen3-14B	√	x	v1	1	32	https://huggingface.co/Qwen/Qwen3-14B
2		Qwen3-30B-A3B-Instruct-2507	√	x	v1	2	32	https://huggingface.co/Qwen/Qwen3-30B-A3B
3		Qwen3-32B	√	x	v1	2	32	https://huggingface.co/Qwen/Qwen3-32B
4		DeepSeek-R1-Distill-Llama-70B	√	x	v1	4	32	https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-70B

序号	模型分类	模型名称	是否支持fp16/bf16推理	是否支持W8A8量化	v0/v1后端	最小卡数（64G显存）	最大序列(K) max-model-len	开源权重获取地址
5	Embedding&Rerank	Qwen3-Embedding-8B	√	x	v0	1	40	https://huggingface.co/Qwen/Qwen3-Embedding-8B
6		Qwen3-Reranker-8B	√	x	v0	1	40	https://huggingface.co/Qwen/Qwen3-Reranker-8B
7		bge-reranker-v2-m3	√	x	v0	1	8	https://huggingface.co/BAAI/bge-reranker-v2-m3
8		bge-large-zh-v1.5	√	x	v0	1	0.5	https://huggingface.co/BAAI/bge-large-zh-v1.5
9		bge-m3	√	x	v0	1	8	https://huggingface.co/BAAI/bge-m3
10	多模态	Qwen3-VL-30B-A3B-Instruct	√	x	v1	2	32	https://huggingface.co/Qwen/Qwen3-VL-30B-A3B-Instruct
11		Qwen3-VL-32B-Instruct	√	x	v1	2	32	https://huggingface.co/Qwen/Qwen3-VL-32B-Instruct

表 2-3 支持的模型及其最小卡数和最大序列（开源版本）

序号	模型名称	是否支持fp16/bf16推理	是否支持W8A8量化	是否支持W4A8量化	v0/v1后端	最小卡数（64G显存）	最大序列(K) max-model-len	开源权重获取地址
1	Qwen3.6-35B-A3B	√	x	x	v1	2	128	https://modelscope.cn/models/Qwen/Qwen3.6-35B-A3B/files
2	Qwen3.6-27B	√	x	x	v1	2	128	https://modelscope.cn/models/Qwen/Qwen3.6-27B/files
3	Qwen3-Next-80B-A3B-Instruct	√	x	x	v1	4	128	https://modelscope.cn/models/Qwen/Qwen3-Next-80B-A3B-Instruct/files
4	GLM-5.1-w4a8	x	x	√	v1	8	36	/
5	Owen3.5-122B-A10B	√	x	x	v1	8	256	/
6	DeepSeek-V4-Flash-w8a8-mtp	x	√	x	x	8	128	/
7	DeepSeek-V4-Flash-0731-w8a8	x	√	x	x	8	128	/

3 实施步骤

3.1 准备工作

创建 rf_admin_trust 委托（可选）

说明

当您使用新注册的华为云账号时，则无需执行该准备工作，如果您使用的是IAM用户账户，请确认您是否在admin用户组中，如果您不在admin组中，则需要为您的账号**授予相关权限**，并完成以下准备工作。

步骤1 进入华为云官网，打开**控制台管理**界面，鼠标移动至个人账号处，打开“统一身份认证”菜单。

图 3-1 控制台管理界面

图 3-2 统一身份认证菜单

步骤2 进入“委托”菜单，搜索“rf_admin_trust”委托。

图 3-3 委托列表

- 如果委托存在，则不用执行接下来的创建委托的步骤
- 如果委托不存在时执行接下来的步骤创建委托

步骤3 单击步骤2界面中的“创建委托”按钮，在委托名称中输入“rf_admin_trust”，委托类型选择“云服务”，输入“RFS”，单击“下一步”。

图 3-4 创建委托

步骤4 在搜索框中输入”Tenant Administrator”并勾选搜索结果，单击“下一步”。

图 3-5 选择策略

步骤5 选择“所有资源”，并单击下一步完成配置。

图 3-6 设置授权范围

步骤6 “委托”列表中出现“rf_admin_trust”委托则创建成功。

图 3-7 委托列表

----结束

3.2 快速部署

本章节帮助用户快速完成模型权重同步、推理镜像准备工作。一键部署该解决方案时，参照本章节中的步骤和说明进行操作，即可完成快速部署。

步骤1 在“昇腾云按需算力部署AI模型”解决方案页面，单击“开始部署”，进入解决方案部署页。勾选“我已阅读”，单击“开始部署”。

图 3-8 资源栈创建页面

图 3-9 进入到创建资源栈页面，保持默认配置即可。

步骤2 进入解决方案购买页面，设置基础配置。

图 3-10 基础配置

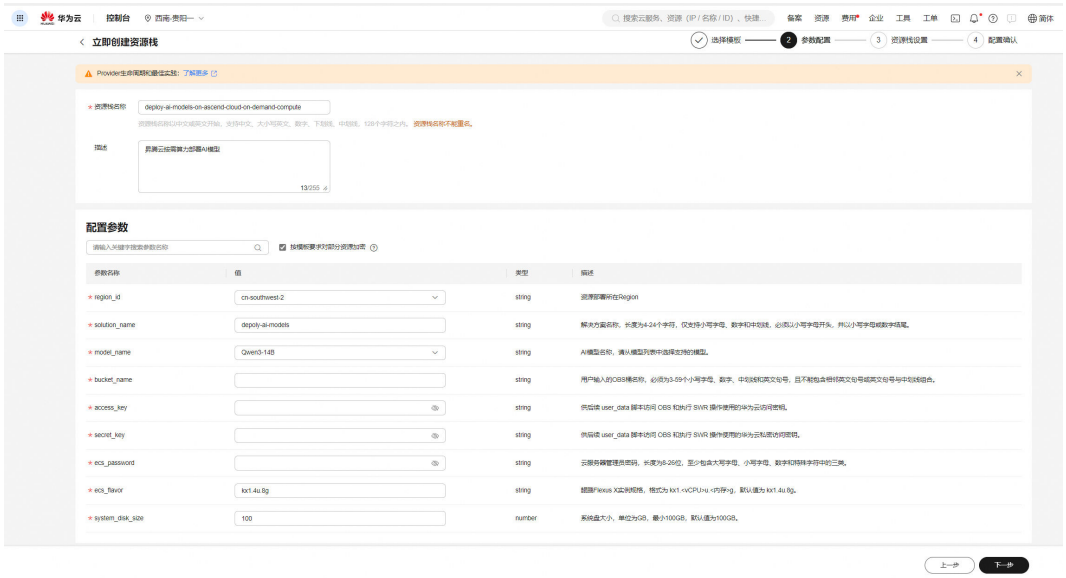
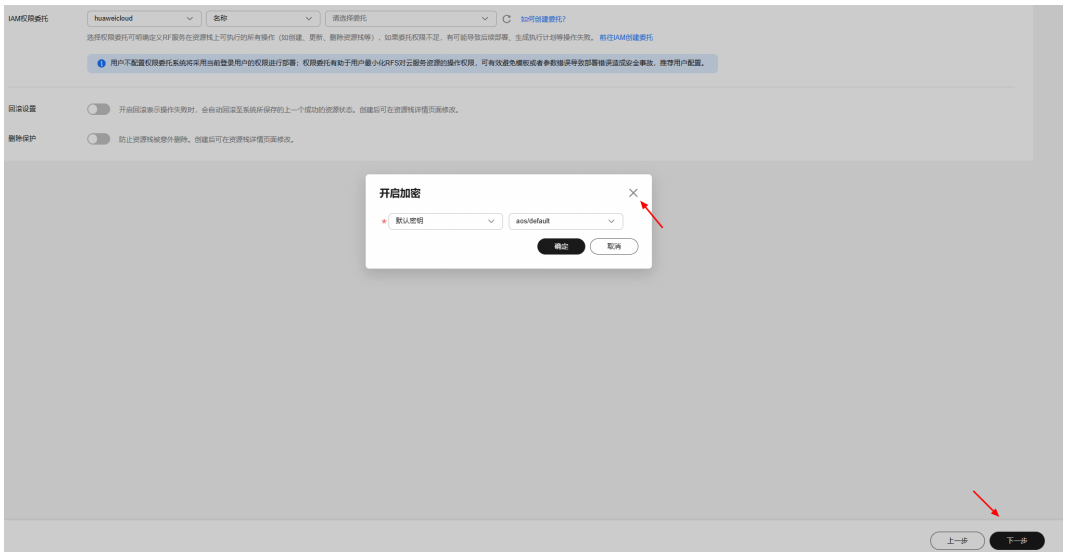


表 3-1 基础配置说明

参数	说明
资源栈名称	资源栈名称以中文或英文开始，支持中文、大小写英文、数字、下划线、中划线，128个字符之内。 资源栈名称不能重名。
region_id	资源部署所在Region。
solution_name	解决方案名称，长度为4-49个字符，仅支持小写字母、数字和中划线，必须以小写字母开头，并以小写字母或数字结尾。
model_name	AI模型名称，请从模型列表中选择支持的模型。
bucket_name	用户输入的OBS桶名称，必须为3-59个小写字母、数字、中划线和英文句号，且不能包含相邻英文句号或英文句号与中划线组合。
access_key	供后续 user_data 脚本访问 OBS 和执行 SWR 操作使用的华为云访问密钥。
secret_key	供后续 user_data 脚本访问 OBS 和执行 SWR 操作使用的华为云私密访问密钥。
ecs_password	云服务器管理员密码，长度为8-26位，至少包含大写字母、小写字母、数字和特殊字符中的三类。
ecs_flavor	鲲鹏Flexus X实例规格，格式为 kx1.<vCPU>u.<内存>g，默认值为 kx1.4u.8g。
system_disk_size	系统盘大小，单位为GB，最小100GB，默认值为100GB。

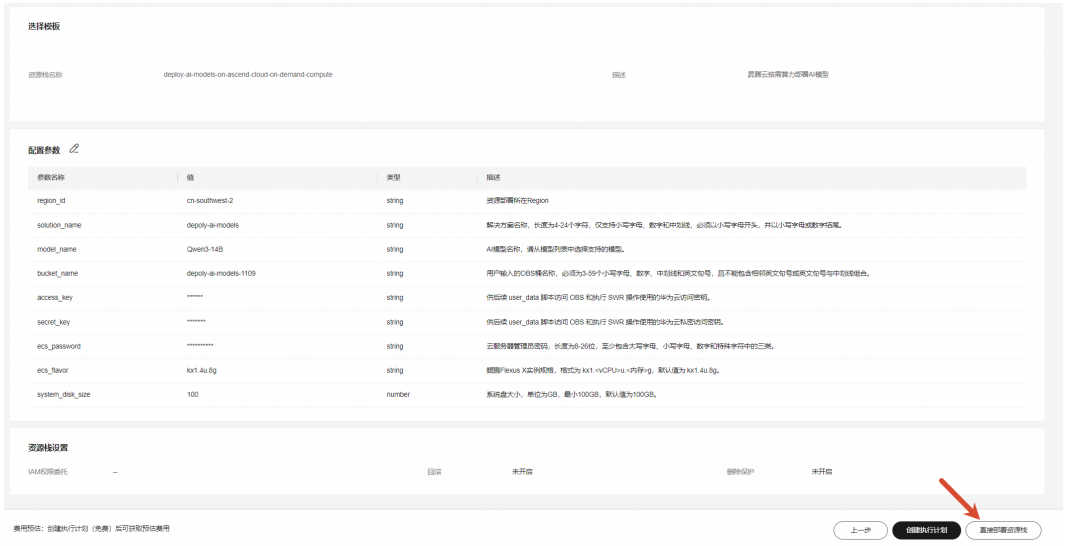
步骤3 开启加密页面关闭后单击“下一步”即可。

图 3-11 关闭加密配置页面



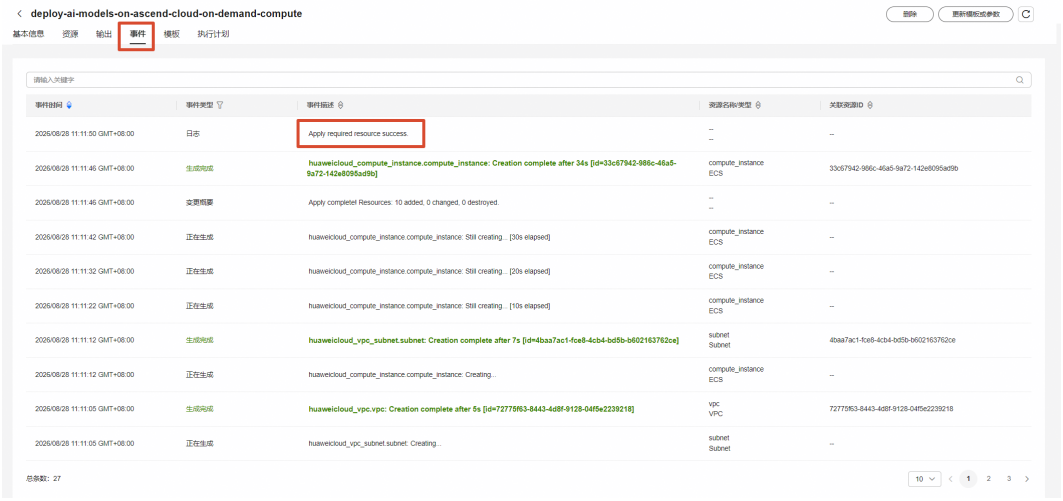
步骤4 部署资源栈。

图 3-12 直接部署资源栈



步骤5 待“事件”页面出现“Apply required resource success”，表示该解决方案资源已经创建完成。

图 3-13 部署完成



步骤6 刷新页面，在“输出”页面查看访问链接。

图 3-14 查看访问链接

----结束

3.3 开始使用

说明

如果您要在魔坊（ModelArts）模型训推平台创建专属资源池资源，请联系您所在企业的华为技术工程师开通白名单，具体购买流程请参考“**创建专属资源池**”文档。

如果您购买的是专属资源池，为了提升部署效率，建议提前在专属资源池中**模型预热（可选）**。首次部署推理服务时，系统需从 OBS 下载模型权重至节点，大模型加载耗时较长；启用预热功能可提前加载模型，从而显著缩短部署时间。

部署在线服务

步骤1 选择"模型推理 > 在线推理"，进入在线服务管理页面。

图 3-15 在线服务管理页面

步骤2 单击右上角的"部署"，进入"部署在线服务"页面。

步骤3 在部署在线服务页面，基本信息可参考如下**表1 服务信息**填写完成后，单击"下一步"，进行部署配置。

图 3-16 基本信息配置

图 3-17 网络信息配置

图 3-18 高可用配置

表 3-2 服务信息

参数类别	参数名称	参数说明	取值示例
基础信息	服务名称	自定义服务名称，建议加上模型名称。	Qwen3-14B
网络配置	服务接入方式	选择"默认"，使用ModelArts平台能力。	默认
网络配置	服务协议	选择"HTTP"，请求超时按需修改。	HTTP
网络配置	认证方式	按需选择服务认证方式	API KEY 认证
高可用配置	流控策略	流控策略由“策略类型”与“流控方式”组成，相同的策略不支持重复添加，多个策略将作为规则集组合生效。	保持默认
高可用配置	请求超时时间（秒）	发送请求后等待系统返回首Token结果的最长等待时间，输入值必须在1到1200之间，单位：秒，超过该时间仍未收到回复，请求将被自动终止。流式传输场景下,每次收到请求响应时，会重新刷新该超时时间，但系统端到端响应超时时间固定为 3600 秒。	根据具体使用调节

步骤4 在部署配置页面填写配置参数。

- 资源配置：根据需要选择公共资源池或者专属资源池，本文档以公共资源池为例。

图 3-19 资源配置

- 模型配置：选择自定义模型，存储地址为**3.2 快速部署**创建对象存储 OBS桶，桶名称为**表3-1**中**bucket_name**的名称，挂载路径默认为：**/home/models-file**。

图 3-20 模型配置

- 单元配置：选择基础模型，单元实例规格为部署模型需要的最小卡数，参考**表2-2**中的最小卡数（64G显存），镜像类型为选择自定义镜像，镜像源地址为**lm_inference**文件下的镜像。如果选择专属资源池也可以**参考推理单元参数说明**的自定义规格分配。

图 3-21 单元配置

- 启动命令，根据实际部署模型替换对应的变量。

  a. 如果是**表2-2**中的模型，启动命令如下：
  ```
  sh /home/models-file/code/run_vllm_908.sh ${model_name} ${port} ${required_cards}
  ```

  b. 如果是**表2-3**中的模型，启动命令如下：
  ```
  sh /home/models-file/code/run_vllm_open_source.sh ${model_name} ${port} ${required_cards}
  ```

  变量解释：

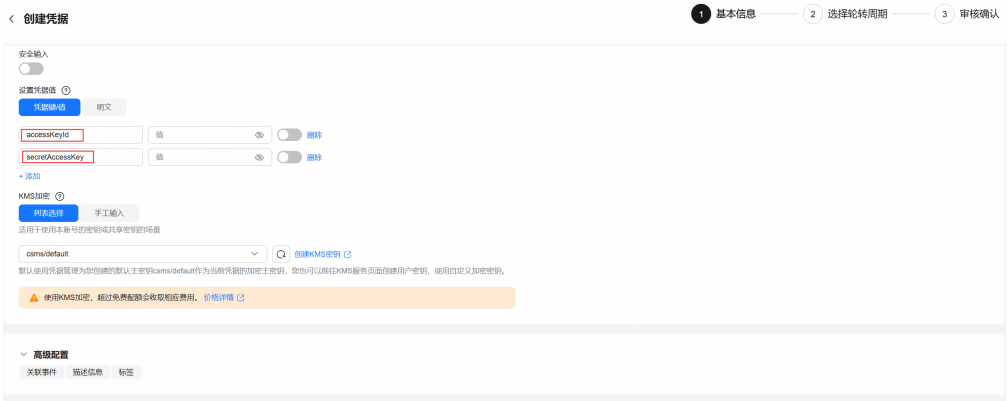
  - **model_name**：模型名称，和**表2表3 支持的模型及其最小卡数和最大序列**中的模型名称一致
  - **required_cards**：NPU卡的数量，参考**表2表3 支持的模型及其最小卡数和最大序列**中的最小卡数

- **port**：服务部署的端口

图 3-22 启动命令

- 单击创建凭据，自定义填写凭据名称，企业项目，参考**访问密钥**获取账号的AK,SK。键为accessKeyId和secretAccessKey，值分别对应用户的AK、SK信息。

图 3-23 创建凭据



- 部署管理配置中容器协议为http，容器端口和**图8启动命令**中的port变量值保持一致，填写完配置参数单击下一步，启动部署任务。

图 3-24 部署管理配置

----结束

测试在线服务

通过API接口在魔坊（ModelArts）模型训推平台的在线推理页面验证推理服务可用性。本文章只是示例，详细可参考**测试在线服务**。

步骤1 进入**在线推理**服务列表，选择部署成功的在线推理服务，单击服务调用，复制调用地址。

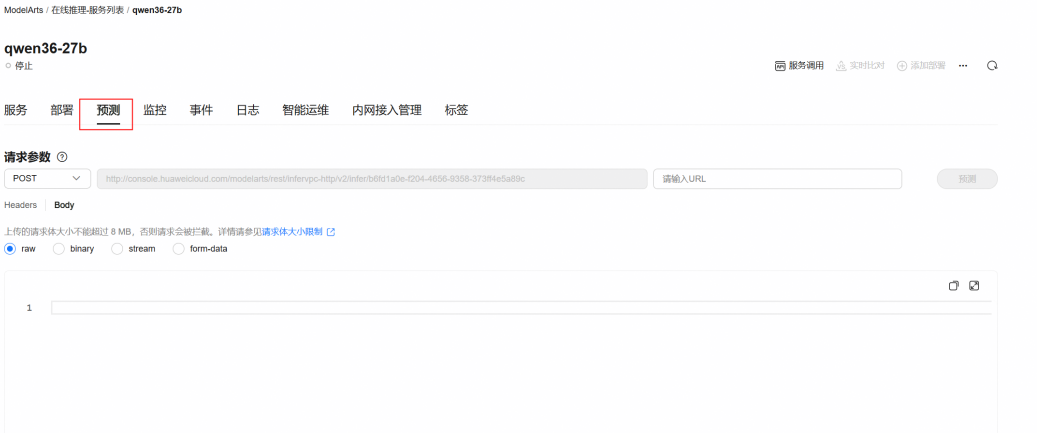
图 3-25 获取调用地址

步骤2 进入**在线推理**服务列表，单击API Key管理。

图 3-26 创建 API Key

步骤3 单机要预测的服务名称，进入详情页，选择"预测"页签。

图 3-27 预测页签



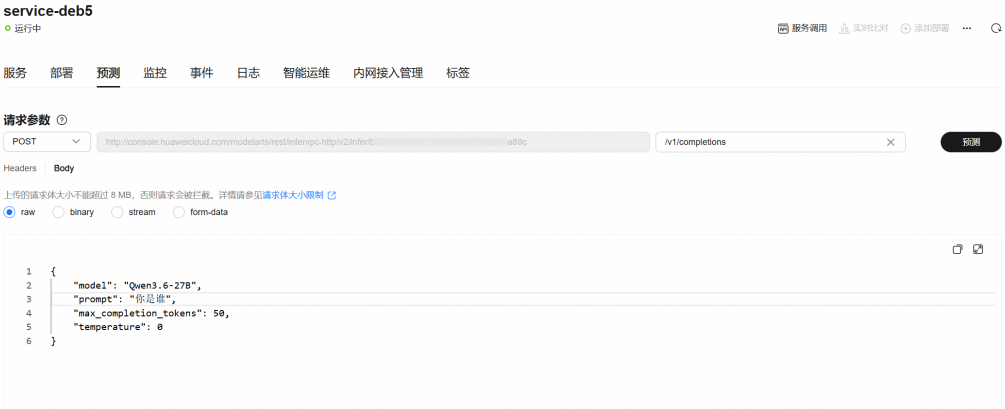
步骤4 设置请求参数。 示例：

- 请求方法：POST
- 请求URL：**步骤1**复制的URL+/v1/completions
- Headers：填写Header信息，比如api-key鉴权信息。例如：Authorization：'Bearer ' + api_key
- 数据格式：stream
- 请求Body：模型名称需要和**表2表3 支持的模型及其最小卡数和最大序列**中模型名称保持一致

```
{   "model": ${name},
    "prompt": "The future of AI is",
    "max_completion_tokens": 50,
    "temperature": 0
}
```

步骤5 单击"预测"，在预测结果处查看返回结果，服务正常时返回模型名称、模型信息和回答内容。

图 3-28 预测模型



----结束

通过 curl 命令预测

大语言模型

通过OpenAI服务API接口启动服务使用以下推理测试命令。

- ${url}：进入**在线推理**服务列表，选择部署成功的在线推理服务，单击服务调用，复制调用地址。
- ${container_model_path}：和**表2 表3支持的模型及其最小卡数和最大序列**模型名称一致。
- ${API-Key}：**步骤2创建的API Key**，无认证该参数无需填写。

OpenAI Completions API with vLLM

```
curl -X POST ${url}/v1/completions \
-H "Content-Type: application/json" \
-H 'Authorization: Bearer ${API-Key}' \
-d '{
      "model": "${container_model_path}",
      "prompt": "hello",
      "max_tokens": 32,
      "temperature": 0
}'
```

OpenAI Chat Completions API with vLLM

```
curl -X POST ${url}/v1/chat/completions \
-H "Content-Type: application/json" \
-H 'Authorization: Bearer ${API-Key}' \
-d '{
    "model": "${container_model_path}",
    "messages": [
        {
            "role": "user",
            "content": "hello"
        }
    ],
    "max_tokens": 32,
    "temperature": 0
}'
```

服务的API与vLLM官网相同，此处介绍关键参数。详细参数解释请参见官网**https://docs.vllm.ai/en/stable/api/vllm/vllm.sampling_params.html**

OpenAI服务相关请求参数说明请参照**表3-3**。

表 3-3 OpenAI 服务请求参数说明

参数	是否必选	默认值	参数类型	描述
model	是	无	Str	通过OpenAI服务API接口启动服务时，推理请求必须填写此参数。取值必须和启动推理服务时的model ${container_model_path} 参数保持一致。 通过vLLM服务API接口启动服务时，推理请求不涉及此参数。
prompt	是	-	Str	请求输入的问题。

参数	是否必选	默认值	参数类型	描述
max_tokens	否	16	Int	每个输出序列要生成的最大tokens数量。
top_k	否	-1	Int	控制要考虑的前几个tokens数量的整数。设置为-1表示考虑所有tokens。 适当降低该值可以减少采样时间。
top_p	否	1.0	Float	控制要考虑的前几个tokens的累积概率的浮点数。必须在(0, 1]范围内。设置为1表示考虑所有tokens。
temperature	否	1.0	Float	控制采样的随机性的浮点数。较低的值使模型更加确定性，较高的值使模型更加随机。0表示贪婪采样。
stop	否	None	None/Str/List	用于停止生成的字符串列表。返回的输出将不包含停止字符串。 例如：["你", "好"]，生成文本时遇到"你"或者"好"将停止文本生成。
stream	否	False	Bool	是否开启流式推理。默认为False，表示不开启流式推理。
n	否	1	Int	返回多条正常结果。 约束与限制： 不使用beam_search场景下，n取值建议为1≤n≤10。如果n>1时，必须确保不使用greedy_sample采样。也就是top_k > 1; temperature > 0。 使用beam_search场景下，n取值建议为1<n≤10。如果n=1，会导致推理请求失败。 **说明** n建议取值不超过10，n值过大会导致性能劣化，显存不足时，推理请求会失败。
use_beam_search	否	False	Bool	是否使用beam_search替换采样。 约束与限制：使用该参数时，如下参数需按要求设置： n>1 top_p = 1.0 top_k = -1 temperature = 0.0 **警告** 使用 beam_search 时，需显式设置 max_tokens，避免请求无法按预期停止。

参数	是否必选	默认值	参数类型	描述
presence_penalty	否	0.0	Float	presence_penalty表示会根据当前生成的文本中新出现的词语进行奖惩。取值范围[-2.0，2.0]。
frequency_penalty	否	0.0	Float	frequency_penalty会根据当前生成的文本中各个词语的出现频率进行奖惩。取值范围[-2.0，2.0]。
length_penalty	否	1.0	Float	length_penalty表示在beam search过程中，对于较长的序列，模型会给予较大的惩罚。 如果要使用length_penalty，必须添加如下三个参数，并且需将use_beam_search参数设置为true，best_of参数设置大于1，top_k固定为-1。 "top_k": -1 "use_beam_search":true "best_of":2
ignore_eos	否	False	Bool	ignore_eos表示是否忽略EOS并且继续生成token。
guided_json	否	None	Union[str, dict, BaseModel]	使用openai启动服务，如果需要使用JSON Schema时要配置guided_json参数。

调用 Embedding&Rerank

通过OpenAI服务API接口启动服务使用以下推理测试命令。

- ${url}：进入**在线推理**服务列表，选择部署成功的在线推理服务，单击服务调用，复制调用地址。
- ${container_model_path}：和**表2 表3支持的模型及其最小卡数和最大序列**模型名称一致。
- ${API-Key}：**步骤2创建的API Key**，无认证该参数无需填写。

rerank接口示例如下:

```
curl -X POST ${url}/v1/rerank \
    -H "Content-Type: application/json" \
    -H 'Authorization: Bearer ${API-Key}' \
    -d '{
        "model": "${container_model_path}",
        "query": "What is the capital of France?",
        "documents": [
            "The capital of France is Paris",
            "Reranking is fun!",
            "vLLM is an open-source framework for fast AI serving"
        ]
    }'
```

表 3-4 表 1 Rerank 服务请求参数说明

参数	是否必选	默认值	参数类型	描述
model	是	无	Str	${container_model_path} 的值和**表2表3 支持的模型及其最小卡数和最大序列**模型名称一致；
query	是	无	Str	用户查询文本
documents	是	无	Str	待排序文档列表（通常为Embedding召回的Top-K结果）

使用OpenAI启动服务（仅支持V0启动），embeddings接口示例如下:

```
curl -X POST ${url}/v1/embeddings \
   -H "Content-Type: application/json" \
   -H 'Authorization:Bearer${API-Key}' \
   -d '{
      "model": "${container_model_path}",
      "input":"I love shanghai"
   }'
```

表 3-5 表 2 Embedding 服务请求参数说明

参数	是否必选	默认值	参数类型	描述
model	是	无	Str	${container_model_path} 的值和**表2表3 支持的模型及其最小卡数和最大序列**模型名称一致；
input	是	无	Str	支持字符串或字符串列表

调用多模型模型

通过OpenAI服务API接口启动服务使用以下推理测试命令。

- ${url}：进入**在线推理**服务列表，选择部署成功的在线推理服务，单击服务调用，复制调用地址。
- ${container_model_path}：和**表2 表3支持的模型及其最小卡数和最大序列**模型名称一致。
- ${API-Key}：**步骤2创建的API Key**，无认证该参数无需填写。
- ${image_url}值为图片地址（如https://example.com/cat.jpg）。

```
curl ${url}/v1/chat/completions \
   -H "Content-Type: application/json" \
   -H 'Authorization: Bearer ${API-Key}' \
   -d '{
   "model": "${container_model_path}",
   "messages": [
      {"role": "system", "content": "You are a helpful assistant."},
```

```
        {"role": "user", "content": [
            {"type": "image_url", "image_url": {"url": "${image_url}"}},
            {"type": "text", "text": "图片中内容"}
        ]}
    ],
    "max_tokens": 512,
    "temperature": 0.7
}'
```

模型预热（可选）

步骤1 登录**ModelArts管理控制台**。

步骤2 **预热模型权重按照如下操作：**

- 在左侧菜单栏中选择"资源管理 > 专属算力资源 > 资源池"。
- 在资源池列表，选择目标专属资源池，单击名称，进入详情页。
- 选择"预热任务"，单击"添加预热任务"，在对话框中完成任务配置。此处仅列出关键参数配置说明，详细操作请参见**在专属资源池添加模型预热**。
- 配置完成后，单击"确定"提交模型预热任务。 当"预热状态"变成"保温中"则表示模型预热已完成。

表4 添加模型预热任务的关键参数说明

参数名称	参数说明	取值示例
权重路径地址	选择待预热的模型权重文件的OBS路径。即步骤一的路径。	obs://xxx/models/
文件占用空间	根据需要预热的模型权重文件大小输入数值。选择完"权重路径地址"后，单击"获取权重文件占用空间"获取到推荐值，一般情况下，该预测值略大于模型实际目录大小。	780

步骤3 返回ModelArts主菜单栏，选择"模型推理 > 在线推理"，进入在线服务管理页面。

----结束

3.4 快速卸载

步骤1 在AI模型部署完成后，可释放OBS桶、SWR组织和Flexus云服务器X实例及相关资源。

步骤2 **卸载OBS桶**。登录**对象存储服务OBS控制台**，单击进入“桶列表”，选择此次部署的OBS桶进入。

图 3-29 选择此次部署的 OBS 桶

步骤3 全选文件，点击删除按钮，删除OBS桶内所有文件。

图 3-30 删除 OBS 桶内所有文件

步骤4 进入**容器镜像服务 SWR控制台**，删除创建的镜像。

图 3-31 删除创建的 SWR 镜像

步骤5 登录**资源编排服务 RFS**，单击进入“资源栈”，单击该方案堆栈后的“删除”。

图 3-32 一键卸载

步骤6 在弹出的删除堆栈确认框中，输入DELETE ，单击“确定”，即可卸载解决方案。

图 3-33 删除堆栈确认

----结束

4 附录

名词解释

- 魔坊（ModelArts）模型训推平台：面向开发者的一站式AI开发平台，可快速创建和部署模型，管理全周期AI工作流，助力千行百业智能升级。
- Flexus云服务器X实例：Flexus云服务器X实例是新一代面向中小企业和开发者打造的柔性算力云服务器。Flexus云服务器X实例功能接近ECS， 同时还具备独有特点，例如Flexus云服务器X实例具有更灵活的vCPU内存配比、支持热变配不中断业务变更规格、支持性能模式等。
- 虚拟私有云 VPC：是用户在华为云上申请的隔离的、私密的虚拟网络环境。用户可以基于VPC构建独立的云上网络空间，配合弹性公网IP、云连接、云专线等服务实现与Internet、云内私网、跨云私网互通，帮您打造可靠、稳定、高效的专属云上网络。
- 弹性公网IP EIP：提供独立的公网IP资源，包括公网IP地址与公网出口带宽服务。可以与弹性云服务器、裸金属服务器、虚拟IP、弹性负载均衡、NAT网关等资源灵活地绑定及解绑，提供访问公网和被公网访问能力。

5 修订记录

表 5-1 修订记录

发布日期	修订记录
2026-8-28	第一次正式发布。