

智果（AgentArts）智能体平台

Managed Agents

文档版本 01
发布日期 2026-09-24



版权所有 © 华为云计算技术有限公司 2026。保留一切权利。

非经本公司书面许可，任何单位和个人不得擅自摘抄、复制本文档内容的部分或全部，并不得以任何形式传播。

商标声明



HUAWEI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。

华为云计算技术有限公司

地址：贵州省贵安新区黔中大道交兴功路华为云数据中心 邮编：550029

网址：<https://www.huaweicloud.com/>

目 录

- 1 Managed Agents 介绍..... 1
- 2 快速使用.....4
- 3 配置智能体运行环境..... 15
 - 3.1 配置运行环境.....15
 - 3.2 存储配置..... 20
- 4 配置专属智能体..... 24
 - 4.1 设置智能体..... 24
 - 4.2 模型..... 32
 - 4.3 提示词.....47
 - 4.4 内置工具..... 49
 - 4.5 Skill..... 51
- 5 将工作交给智能体..... 53
 - 5.1 调用智能体概述..... 53
 - 5.2 调用智能体（IAM 认证）..... 54
 - 5.3 调用智能体（API Key 认证）..... 58
 - 5.4 检查结果状态.....60
 - 5.5 查看智能体运行时数据..... 61
- 6 常见问题.....62
 - 6.1 环境被智能体关联后能否修改..... 62
 - 6.2 会话空闲超时和最大存活时间有何区别..... 62
 - 6.3 会话存储、SFS Turbo、OBS 三种存储有什么区别..... 63
 - 6.4 存储挂载路径有哪些约束..... 63
 - 6.5 会话超时后数据还在吗..... 64
 - 6.6 创建智能体时，报错 AgentArts.03002006，如何处理？..... 64

1 Managed Agents 介绍

Managed Agents是云端全托管的智能体运行基础设施平台。开发者只需定义智能体模型、提示词、工具与技能、权限等，平台负责提供整套Harness（运行编排框架）底座，自动完成Agent循环、沙箱隔离、网络管控、会话持久化、日志审计、资源调度，无需自行搭建复杂底层工程。

AgentArts提供三种开发智能体的方式，具体区别如下：

表 1-1 三种开发智能体的区别

-	Managed Agents	智能体管理（低码智能体）	托管与运行
核心定位	表单化声明+全托管运行智能体	可视化编排智能体应用	用SDK编程构建智能体
工作方式	填表单配置，模型自主规划执行	拖拽画布、流程编排	写代码（Runtime/Tools/Memory/Gateway SDK、CLI）
目标用户	需要快速上线通用Agent并做企业级治理的团队	业务人员、低代码开发者	专业开发者、平台工程师

核心能力

- 快速创建智能体**
支持通过表单快速创建智能体，配置身份、工具技能、身份认证，并关联复用执行环境模板。
- 智能体生命周期监控**
支持日志采集、基线指标监控与全链路追踪，可实时观察智能体从部署、运行到调用的全过程状态。
- 智能体关联复用环境**
创建智能体并关联环境模板，多智能体复用同一套执行环境配置。

核心概念

- 智能体（Agent）**
定义功能与职责（“做什么”），涵盖模型、系统提示词、运行环境、工具与技能等配置项。关联运行环境以保障正常执行。同时支持版本管理。
- 环境（Environment）**
定义智能体运行位置（“哪里运行”），涵盖入站网关、出网网络配置、会话存储等配置项。被智能体关联后禁止修改。
- 委托**
智能体运行所需的云资源访问授权。
- Skill**
从技能库选择挂载到智能体的技能包。
- 会话**
一次完整的交互上下文，引用智能体和环境配置。

工作流程

- 创建环境：配置出网网络配置、会话存储等。
- 创建智能体：配置模型、系统提示词、工具与技能等，并关联运行环境。
- 通过API调用智能体：获取调用智能体的相关信息。
- 发送请求并获取响应：发送请求内容，并可查看响应结果信息。
- 干预或中断：指任务执行过程中，通过发送新的内容来变更原有执行方向，或是直接终止、暂停当前任务。
- 在智能体详情的“版本”、“日志”页签查看版本演进与运行日志。

支持的工具

智能体通过以下工具与运行环境交互：

- 内置工具**
智能体内置工具集包含以下工具。

表 1-2 内置工具

工具名称	描述
*	包含所有命令，可以在会话中执行bash命令或文件操作。
bash	在会话中执行bash命令。
read_file	读取文本、图片、PDF、Jupyter Notebook、Office文档。
write_file	写入文件内容（覆盖式）。
edit_file	精确字符串替换编辑已有文件。
glob	使用glob模式查找文件。
list_files	列出目录内容。
grep	在文件中搜索内容，支持正则表达式。

- Skill：特定领域能力（如翻译、摘要）。

适用场景

- 长时运行：任务持续运行数分钟至数小时，执行期间包含多次工具调用。
- 全托管：沙箱环境由平台托管运行，无需自行部署，依托受控基础设施满足合规规范及数据存储相关要求。
- 文件操作：支持对沙箱环境中的文件进行读取、写入、搜索以及编辑等相关操作。
- 有状态交互：在多轮交互过程中持续维护上下文信息与文件系统的运行状态。
- 高效开发与低运维负担：开发者可专注实现智能体核心逻辑，不必从零构建和维护智能体循环、沙箱编排以及工具执行层基础设施。

费用提示

Managed Agents如果选择API Key认证，则创建后即开始计费。API Key由凭据管理服务（DEW）托管，托管按需计费，费用说明请参考[DEW 计费说明](#)。

Managed Agents在运行过程中产生的费用，具体请参考[计费项](#)。

2 快速使用

本章节旨在帮助用户快速了解基本功能，建议在充分了解各项功能后再进行深度应用。

前提条件

- 已开通AgentArts服务，并在“管理中心 > 授权管理”完成依赖云服务授权。
- 本地已安装Python 3.10或以上（调用智能体时使用）。

步骤一：接入模型

这里以接入华为云MaaS的GLM-5.3模型为例，采用OpenAI兼容协议，区域选择“西南-贵阳”。

查看及获取 MaaS 的 GLM-5.3 模型信息

- 进入“西南-贵阳” [MaaS页面](#)，选择“模型推理 > 在线推理”，找到GLM-5.3模型，查看模型类型。

图 2-1 获取模型信息 1

MaaS

模型广场

模型推理

模型部署

模型定制

管理计划

API Key 管理

订阅管理

调用统计

预置服务 自定义接入点

预置服务 (31)

服务名称/来源 状态 类型 计费方式 推理定价 充值折扣 模型规格 调用统计 操作

gptPangu-2.0-Pro

已开通

文本生成

按Token计费

输入: ¥3.20 - ¥4.00 / 百万tokens
输出: ¥14.00 - ¥17.00 / 百万tokens

-

TPM: 1,000,000
RPM: 100

无限制

API调用

更多

gptPangu-2.0-Fast

已开通

文本生成

按Token计费

输入: ¥0.20 - ¥0.50 / 百万tokens
输出: ¥1.00 / 百万tokens

-

TPM: 1,000,000
RPM: 100

无限制

API调用

更多

GLM-5.3

已开通

文本生成

按Token计费

输入: ¥2.00 - ¥8.00 / 百万tokens
输出: ¥20.00 / 百万tokens

-

TPM: 1,000,000
RPM: 100

20200814 主推版本

无限制

API调用

更多

GLM-5.2

已开通

文本生成

按Token计费

输入: ¥0.60 - ¥8.00 / 百万tokens
输出: ¥19.00 - ¥20.00 / 百万tokens

-

TPM: 1,000,000
RPM: 100

无限制

API调用

更多

GLM-5.1

已开通

文本生成

按Token计费

输入: ¥0.01 - ¥0.00 / 百万tokens
输出: ¥15.00 - ¥20.00 / 百万tokens

-

TPM: 1,000,000
RPM: 100

无限制

API调用

更多

Kimi K2.5

已开通

文本生成

按Token计费

输入: ¥4.00 - ¥0.00 / 百万tokens
输出: ¥10.00 - ¥27.00 / 百万tokens

-

TPM: 100,000
RPM: 10

无限制

API调用

更多

- 单击“API调用 > 主推版本”，查看接口协议，获取API地址（Base_url）、model参数值（模型名称）、认证方式及位置。

图 2-2 获取模型信息 2

API调用说明

×

标准接口

OpenAI兼容接口

协议

接口信息

API地址

https://api.modelarts-maas.com/openai/v1

Base_url

model参数

glm-5.3

模型名称

步骤一：获取API Key

在调用MaaS的模型服务时，需要填写API Key用于接口的鉴权认证。请[创建 API Key](#)或使用已有API Key，前往[API Key管理](#)

步骤二：复制调用示例并替换接口信息、API Key

1.安装环境

请先按如下命令安装环境

```
pip install --upgrade "openai>=1.0"
```

2.复制调用示例并替换接口信息、API Key

- 复制 下方调用示例代码
- 替换 其中的 接口信息（API地址、model参数）为上方接口信息
- 替换 其中的 API Key为已获取的API Key

调用示例代码

Python

```
from openai import OpenAI

base_url = "https://api.modelarts-maas.com/openai/v1" # API地址
api_key = "MAAS_API_KEY" # 把MAAS_API_KEY替换成已获取的API Key
client = OpenAI(api_key=api_key, base_url=base_url)

response = client.chat.completions.create(
```

3. 获取API Key。
- 在MaaS左侧导航，选择“模型与统计 > API Key 管理”，单击“创建 API Key”，如图2-3所示，单击“确定”，复制保存API Key信息。

图 2-3 创建 API Key

创建 API Key

⌵

⚠ API Key 仅会在新建后显示一次，请及时复制并妥善保存。

基础信息

标签

maas-apikey

标签长度范围为 1 到 100 个字符，支持大小写英文字母、数字、下划线、中划线。

描述

maas-apikey

权限设置

权限范围

全部（访问所有资源不受限）

4. 查看API Key鉴权请求格式。
- 在[OpenAI兼容接口](#)页面，查看鉴权请求。

图 2-4 API Key 请求格式

约束限制

该功能仅支持“西南-贵阳一”区域。

接口信息

名称	说明	取值
API地址	调用模型服务的API地址。	https://api.modelarts-maas.com/openai/v1/chat/completions 如果使用OpenAI SDK，base_url设置为https://api.modelarts-maas.com/openai/v1。
鉴权请求	向服务器提供身份验证/授权凭证。	"Authorization": "Bearer \$MaaS_API_Key"

接入模型供应商

1. 登录[AgentArts控制台](#)，在左侧导航选择“托管与运行 > 模型”。
2. 单击“添加模型供应商”，配置模型供应商基本信息，如[图2-5](#)所示。
Base_url为模型API地址，获取方法请参考[2](#)。

图 2-5 配置模型供应商基本信息

添加模型提供商

基本信息

名称

maas

以小写字母开头，以小写字母或数字结尾，可以包含小写字母、数字和中划线，长度为2-40个字符。

描述 (可选)

接入华为云MaaS模型即服务的GLM模型

Base_url

https://modelarts.cn-southwest-2.myhuaweicloud.com

以https://开头的完整url，长度为1-2048个字符。

3. 配置权限与访问控制，如图2-6所示。

图 2-6 权限与访问控制

权限与访问控制

出站身份认证

IAM 认证

API Key 认证

maas-apiey

创建出站身份

位置

标头

参数名称 (可选)

Authorization

指定将包含API密钥的参数名称，默认为Authorization。

前缀 (可选)

Bearer

在API密钥值之前添加的可选前缀(例如，在授权标头中使用的"Bearer")。

出网网络配置

公网访问

私网访问

- 出站身份认证：访问MaaS模型的认证方式，获取请参考2。
单击“创建出站身份”，填写信息，如图2-7所示，单击“创建”。
API Key值：获取方式请参考3。

图 2-7 创建出站身份

< 创建出站身份

身份名称

maas-apiey

认证类型

OAuth 客户端

API Key

IAM 委托

API Key 是用于身份验证和授权访问资源的唯一标识符。

计费提示

身份认证凭证在凭据管理服务（CSMS）托管，托管按需计费。[了解计费详情](#)

API Key 值

.....

^ 标签（可选）

如果您需要使用同一标签标识多种云资源，即所有服务均可在标签输入框下拉选择同一标签，建议在TMS中[创建预定义标签](#)

+ 添加标签

您还可以添加20个标签。

- 位置：标头，获取请参考2。
- 参数名称（可选）：Authorization，获取请参考4。
- 前缀（可选）：Bearer，获取请参考4。

4. 高级配置。
- 保持默认。
- 模型供应商信息配置完成后，即可接入模型服务。

接入模型服务

1. 单击“添加模型服务”，配置模型信息，如图2-8所示，配置完成后，单击“确定”。
- 模型名称：查看GLM-5.3接口中的model字段值，具体请参考2。
- API接口协议：查看GLM-5.3遵循的接口协议，具体请参考2。
- 模型类型：查看GLM-5.3的模型类型，具体请参考1。

图 2-8 配置模型信息

添加模型服务

×

模型服务名称

maas

支持中英文、数字及符号：_/\-，须以中英文或数字开头结尾，长度2-64个字符。

描述 (可选)

接入华为云MaaS模型即服务的GLM-5.3模型

模型名称

glm-5.3

由英文、数字及符号：连字符、点、下划线、斜杠、问号和@，支持*的前缀通配以及后缀通配，长度为1-256个字符。

API接口协议

标准OpenAI协议

Anthropic协议

华为云MaaS标准API V1

模型类型

☒ 文本对话/图像理解

☐ 文本向量化

☐ 多模态对话

☐ 文本排序

☐ 实时语音对话

2. 单击“创建”。
- 模型供应商显示在模型列表中，接入的模型服务显示在模型供应商详情页面的模型服务列表中，即可被选择使用。

步骤二：快速创建智能体

1. 在Managed Agents页面，选择“智能体 > 创建智能体 > 快速创建”，配置智能体参数。

图 2-9 选择快速创建



图 2-10 创建智能体

创建智能体

创建时会为智能体自动创建环境。创建后，名称不可进行编辑。

名称

agent-taraqm4j

以小写字母开头，以小写字母或数字结尾，只能包含小写字母、数字和中划线，长度为2-40个字符

模型

maas / glm-5.3

管理模型

取消

确定

表 2-1 参数说明

参数	说明
名称	智能体的展示名称，供用户在列表中快速识别与检索该智能体。系统默认生成以agent开头的名称，可以自定义。创建后，不可编辑。 由2~40个字符组成，包含小写字母、数字和中划线，且必须以小写字母开头、以小写字母或数字结尾。
模型	智能体使用的模型。

2. 单击“确定”。
- 自动创建环境，入站身份认证为API Key，入网配置为defaultgw，未开启日志记录。

说明

API Key由凭据管理服务（DEW）托管，托管按需计费，费用说明请参考DEW 计费说明。
创建后，智能体显示在智能体列表中。

步骤三：调用智能体

用于调测智能体与用户本地连通性。

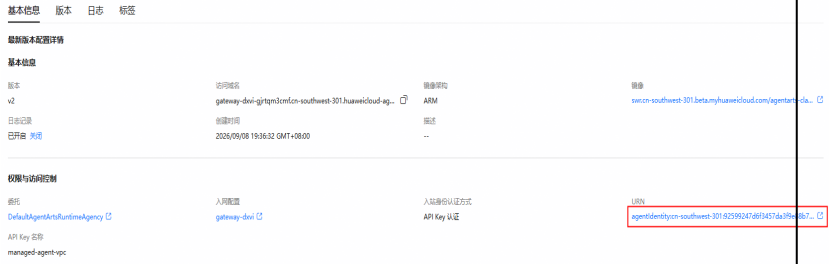
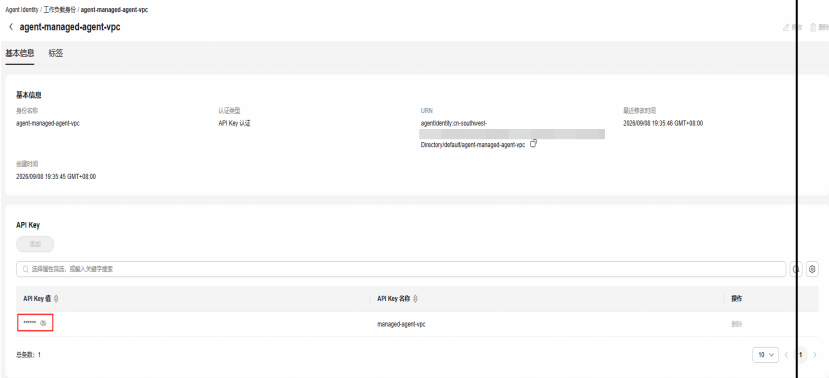
获取调用信息

调用智能体的命令格式如下，需要参考表2-2获取相应的调用信息：

```
python managed_agents_client.py --url "ws://{endpoint}/runtimes/{runtime_name}/ws" --auth "Bearer {api_key}" --session-id "{会话ID}" --message "你好"
```

表 2-2 获取调用信息

参数	操作
<code>managed_agents_client.py</code>	脚本文件名由脚本名称和文件类型后缀组成。例如，脚本名称为 <code>managed_agents_client</code> ，文件类型为Python脚本（.py），则完整文件名为 <code>managed_agents_client.py</code> 。
运行时名称 (<code>{runtime_name}</code>)	1. 在“Managed Agents > 智能体”页面，复制需要调用的智能体名称。 2. 在左侧导航选择“托管与运行 > 智能体运行时”，搜索已获取的智能体名称。 在“智能体运行时”列表，显示的名称以managed-开头，即为运行时名称。 
访问域名 (<code>{endpoint}</code>)	单击运行时名称，进入运行时详情页面，在“访问域名”区域，即可复制访问域名。 

参数	操作
API Key ({api_key})	1. 在运行时详情页面，单击URN。  2. 进入工作负载身份详情页面，即可复制API Key值。 
会话ID ({会话ID})	每个会话的唯一标识符。用户可将会话ID设置为任意字符串，由英文、数字、“-”、“_”组成，不超过64个字符。例如“123e4567e89b12d3a456426614174000”。

调用智能体

安装依赖

打开命令提示符窗口，输入如下命令：

```
pip install websockets>=10.0 --break-system-packages --break-system-packages
```

发送调用智能体命令

下载managed_agents_client.py至本地，该脚本用于通过WebSocket与Managed Agents后端交互，调测智能体与用户本地连通性。

在managed_agents_client.py脚本所在目录，使用Windows命令提示符（CMD）工具，执行如下命令，把四个占位符换成您在获取调用信息中获取的值：

```
python managed_agents_client.py --url "ws://{endpoint}/runtimes/{runtime_name}/ws" --auth "Bearer {api_key}" --session-id "{会话ID}" --message "你好"
```

显示结果

预期终端输出：

```
[2/5] 已连接 ws://{endpoint}/runtimes/{runtime_name}/ws
[3/5] 服务端就绪
[4/5] 已发送: 你好
```

```
[Frame] {"response_kind":"e2a.chunk","status":"in_progress","body":{"delta_kind":"text","delta":"你好","event_type":"chat.delta"}}
你好
[Frame] {"response_kind":"e2a.chunk","status":"in_progress","body":{"delta_kind":"text","delta":"! 有什么可以帮你的吗? ","event_type":"chat.delta"}}
! 有什么可以帮你的吗?
[Frame] {"response_kind":"e2a.complete","status":"succeeded","is_final":true,"body":{"result":{"content":"你好! 有什么可以帮你的吗? }}}

=====
回复: 你好! 有什么可以帮你的吗?
=====
[5/5] 已断开
```

权限审批和用户提问均在后台自动处理，终端只显示“Frame”原始帧和流式回复文本。

Managed Agents 执行流程

通过API发送消息后，Managed Agents执行流程：

1. 沙箱配置：沙箱将按照您设置的环境配置完成构建。
2. 智能体循环执行：在智能体循环执行过程中，Managed Agents根据您发送的消息，自主决策需要调用的工具。
3. 工具执行：所有文件写入操作、bash命令执行以及其余工具调用，都将在沙箱环境中运行。
4. 流式事件推送：当智能体执行任务时，您将收到实时状态更新。
5. 转为空闲状态：当智能体没有需要执行的任务时，会触发该状态。

后续操作

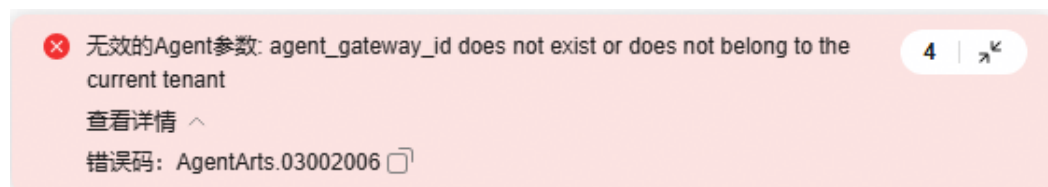
- [配置专属智能体](#)
- [配置智能体运行环境](#)
- [将工作交给智能体](#)

常见问题

Q1：报错AgentArts.03002006，如何处理？

A：创建智能体时，如果报错AgentArts.03002006，请按如下操作进行处理：

图 2-11 报错信息



1. 在[AgentArts智能体平台](#)左侧导航，选择“治理与安全 > 网络配置”。
2. 在defaultgwi详情页面，单击“编辑”，再单击“确定”即可，如[图2-12](#)和[图2-13](#)所示。

图 2-12 编辑

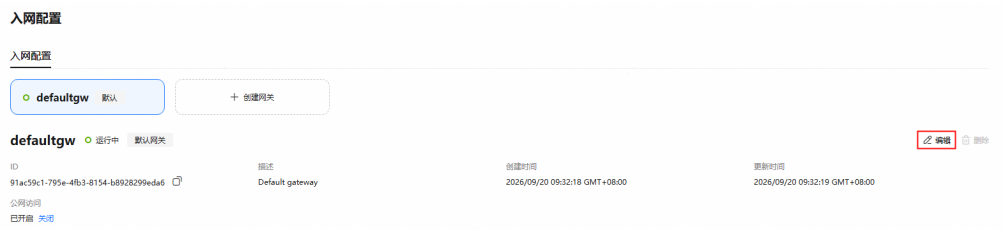
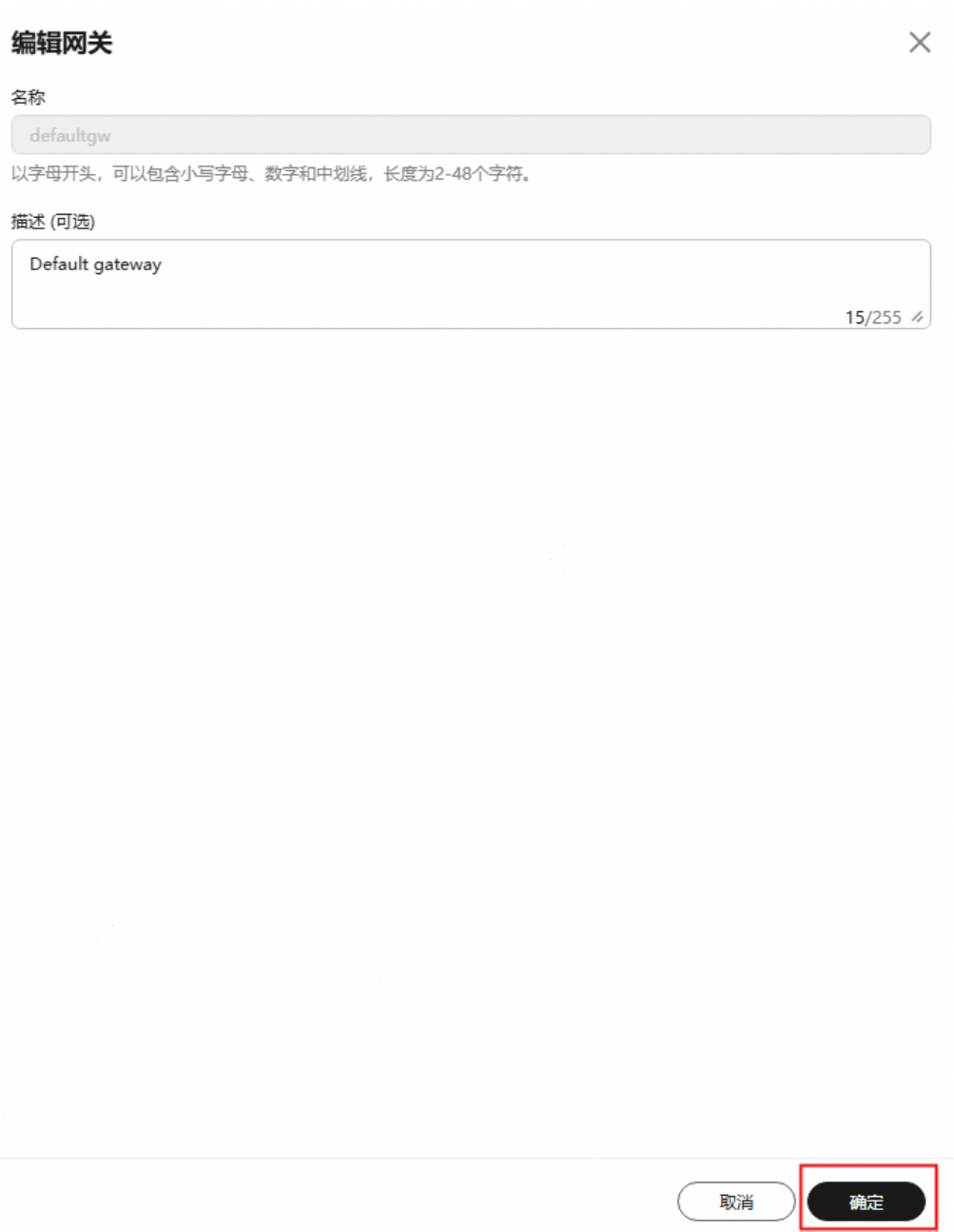


图 2-13 确定



3. 再在“Managed Agents”页面，执行创建智能体操作。

3 配置智能体运行环境

3.1 配置运行环境

为智能体自定义运行环境。

环境（Environment）定义智能体运行位置（“哪里运行”），涵盖入站网关、出网网络配置、会话存储等配置项。被智能体关联后禁止修改。

环境是独立于智能体的运行沙箱。一个环境可以被多个智能体引用。

约束与限制

每个租户可以创建1000个环境。

前提条件

- 已开通AgentArts服务，并在“管理中心 > 授权管理”完成依赖云服务授权。

创建环境

在[Managed Agents](#)页面，选择“环境”页签，单击“创建环境”，配置环境参数。创建后，环境显示在环境列表中。

图 3-1 创建环境



基本信息

环境的基本信息，便于查找、调用和管理。

图 3-2 基本信息

基本信息

名称

environment-ewso2rdu

以小写字母开头，以小写字母或数字结尾，只能包含小写字母、数字和中划线，长度为2-40个字符

描述 (可选)

请输入描述

0/4,096

表 3-1 基本信息

参数	说明
名称	环境的展示名称，供用户在列表中快速识别与检索该环境。系统默认生成以environment开头的名称，可以自定义。创建后，不可编辑。 由2~40个字符组成，包含小写字母、数字和-，且必须以小写字母开头、以小写字母或数字结尾。
描述（可选）	环境的描述信息，帮助用户快速理解该环境。 由0~4096个字符组成。

网络配置

Managed Agents环境提供出网网络配置能力。

沙箱向外发起网络请求的能力，是基础设施层的安全边界，优先级高于系统提示词、工具权限，即使提示词被注入攻击，网络层依然可以拦截非法访问。

网络配置

出网网络配置

公网访问

私网访问

出网网络配置在环境创建后，不可编辑。

- **公网访问**：通过互联网出口访问公网上的服务。公网访问时，智能体不支持添加Skill。
- **私网访问**：打通您的企业私有VPC内网，只能访问内网资源，流量不经过互联网。
当选择“私网访问”时，需配置如下参数：
 - 选择已配置的VPC，默认显示第一个。如未配置，请单击“新建VPC”，具体请参考[创建虚拟私有云和子网](#)。
 - 选择已配置的子网，默认显示第一个。如未配置，请单击“新建子网”，具体请参考[创建虚拟私有云和子网](#)。
 - 选择已配置的安全组，默认显示第一个。如未配置，请单击“新建安全组”，具体请参考[创建安全组](#)。

存储配置

存储配置决定沙箱中文件的挂载方式、数据保留时长、产物输出路径，和网络配置一样属于底层资源配置。优先级高于提示词、工具权限。

图 3-3 存储配置

存储配置

☐ 会话存储

用于保存会话上下文、中间结果、用户偏好等，存储资源由后台服务统一托管，每个会话拥有独立存储空间。配置后不支持取消配置，仅支持配置1个。

☐ SFS Turbo

用于保存会话过程中产生的文件等，存储资源为用户的SFS Turbo文件系统。该配置仅在私网访问下支持并限制数量为5个。

☐ OBS

用于保存会话过程中产生的文件等，存储资源为用户的OBS桶。该配置仅在私网访问下支持并限制数量为10个。

表 3-2 存储配置参数说明

参数	说明
会话存储	会话级别的存储持久化，用于保存会话上下文、中间结果、用户偏好等。存储资源由后台服务统一托管，每个会话拥有独立存储空间。配置后不可取消，仅支持1个。 挂载路径为/home、/mnt、/data、/tmp及其子目录，不能包含“..”，最大长度为4000字符。
SFS Turbo	当“出网网络配置”选择“私网访问”时，支持此参数。 高性能分布式共享文件存储服务，支持多云主机、容器同时挂载访问，提供低时延、弹性扩容的持久化文件存储能力。最多添加5个SFS Turbo类型的存储，且SFS Turbo仅在私网访问下支持配置。 单击“添加”，详细参数配置请参考表3-3。
OBS	当“出网网络配置”选择“私网访问”时，支持此参数。 用于保存会话过程中产生的文件等，存储资源为用户的OBS桶。最多只能选择 5 个不同的OBS桶，最多添加10个OBS类型的存储，且OBS仅在私网访问下支持配置。 单击“添加”，详细参数配置请参考表3-3。

表 3-3 添加存储参数说明

类型	参数	说明
会话存储	挂载路径	输入会话存储将被挂载的具体目录路径。配置后，每个会话在持久化存储中拥有独立的子目录，这些子目录将作为状态存储挂载到该路径。 挂载路径为/home、/mnt、/data、/tmp及其子目录，不能包含“..”，最大长度为4000字符。 存在多个挂载时（包括OBS、SFS Turbo、会话存储之间），禁止使用相同或存在包含关系的路径。

类型	参数	说明
SFS Turbo	文件系统	选择文件系统。如未创建，单击“创建文件系统”，详细操作说明请参考创建文件系统。 配置后单击  刷新。 说明 - SFS Turbo和智能体运行时中的出网网络配置必须在同一VPC内，SFS Turbo文件系统使用的安全组需放通NFS协议使用的端口号，具体信息请参考VPC的安全组是否影响高性能弹性文件服务的使用，安全组需要放通运行时出站的子网网段访问SFS Turbo。 - 选择的文件系统需为NFS协议。 - AgentArts部署智能体运行时挂载NFS文件系统的soft/hard的模式为hard模式，关于soft/hard的模式详情信息请查看挂载SFS Turbo文件系统到Linux云服务器。 - 为了您的信息安全，请勿在文件系统中存放敏感数据。
	存储路径	输入文件系统中数据的存储路径。 存储路径以“/”开头，不能包含“..”，最大长度4000字符，默认为/。
	挂载路径	输入文件系统中数据的挂载路径，使得该目录下的文件和目录可以被访问。 - 挂载路径为/home、/mnt、/data、/tmp及其子目录，不能包含相对路径，最大长度为4000字符。 - 存在多个挂载时（包括OBS、SFS Turbo、会话存储之间），禁止使用相同或存在包含关系的路径。 - 同一智能体运行时的多个版本挂载在同一目录时会自动共享。 说明 挂载路径的Linux文件系统权限，继承自SFS Turbo存储路径的原有权限（SFS Turbo根目录权限为任何用户可读写）。此权限不支持通过chmod/chown等命令修改。
	读写权限	选择对挂载文件系统的读写权限。 - 读写：可以读取、写入和修改文件系统中的数据，默认为读写。 说明 Linux文件系统权限和读写权限两者必须同时允许，读写才能成功，如果任一项禁止，读写操作将失败。 - 只读：只能读取文件系统中的数据，不能写入或修改数据。

类型	参数	说明
OBS	OBS桶	选择已创建的OBS桶。如未创建，单击“创建OBS桶”，详细操作说明请参考创建桶。 配置后单击刷新。 说明 - 挂载路径的OBS文件系统权限，继承自OBS存储路径的原有权限（OBS根目录默认权限为任何用户可读写）。此权限不支持通过chmod/chown等命令修改。 - OBS仅支持桶类型，暂不支持并行文件系统。 - 为了您的信息安全，请勿在文件系统中存放敏感数据。
	存储路径(可选)	输入OBS桶内的子目录路径。 挂载OBS桶内的指定子路径，当配置为空时挂载整个桶根目录；非空时挂载桶内指定子目录，必须以'/'开头，支持大小写字母、数字及！_-.*'()/等字符且不能包含'..'，路径长度为最大1024字符。
	挂载路径	输入OBS桶挂载到容器内的路径。 - 挂载路径为/home、/mnt、/data、/tmp及其子目录，不能包含相对路径，最大长度为4000字符。 - 存在多个挂载时（包括OBS、SFS Turbo、会话存储之间），禁止使用相同或存在包含关系的路径。 说明 挂载路径的Linux文件系统权限继承自OBS桶存储路径的原有权限。此权限不支持通过chmod/chown等命令修改。
	读写权限	选择对挂载OBS桶的读写权限。 - 读写：可以读取、写入和修改OBS桶中的数据，默认为读写。 说明 Linux文件系统权限和读写权限两者必须同时允许，读写才能成功，如果任一项禁止，读写操作将失败。 - 只读：只能读取OBS桶中的数据，不能写入或修改数据。

查看环境详情

在环境列表，单击环境名称，进入环境详情页面。可以查看环境配置项信息。

图 3-4 环境详情

基本信息

环境ID

c66c1a48-1386-48d0-beee-e51d25ada863

更新时间

2026/09/04 12:26:50 GMT+08:00

描述

生产环境VPC网络配置

网络配置

出网网络配置

私有IP地址

安全组

agentarts-cpts-node-cce-mode-0lccc

VPC

vpc-84b6

子网

a0a1e-g_iabcedfghjabcddefghjabcddefghjabcddefghjrsz

存储配置

会话存储

/mnt/sessions

OBS

名称	读写权限	存储路径	挂载路径
agentrun	读写	/	/home/test

更新环境配置

 注意

环境被智能体关联后，禁止编辑。请根据提示进行操作。

在环境列表，单击“编辑”，编辑环境配置项信息。

图 3-5 更新环境配置

☐ 环境名称/ID	描述	创建时间	操作
☐ environment-z54ufeld 20a86f26-e761-4e41-9b7c-2c76250e2917	--	2026/09/14 22:31:03 GMT+08:00	编辑 删除

在智能体中使用环境

在自定义创建智能体时，选择创建的环境。

图 3-6 在智能体中使用环境

环境

环境

environment-z54ufeld

私网访问

▼

 创建环境 

当前公网访问不支持Skill，公网访问环境切换为私网访问环境时，会清空已选择Skill。

后续操作

将工作交给智能体

3.2 存储配置

存储配置用于将外部存储资源挂载到运行时沙箱容器中，实现数据持久化、文件共享和会话状态保存。AgentArts运行时支持SFS Turbo、会话存储和OBS三种存储类型，满足不同场景的数据持久化需求。

原理介绍

存储挂载机制

存储配置在运行时创建时设定，通过Miracle沙箱平台将外部存储资源挂载到沙箱容器内。挂载过程无需自定义挂载代码或特权容器，平台自动完成存储编排和挂载。

版本更新时的配置继承

运行时版本更新时，存储配置支持增量修改（字段继承机制）：

- 未指定某个存储子配置（如sfsTurbo为null）时，继承上一版本的完整配置。
- 指定了某个存储子配置时，使用新配置替换旧配置。
- 会话存储支持字段级继承：未指定的字段（如uid）继承上一版本该字段的值。

存储类型对比

表 3-4 存储类型对比

存储类型	文件特征	读写特征	性能特点	典型场景	最大数量	网络要求
SFS Turbo	适量中大文件	读写混合、多实例共享	中等 IOPS、适中时延	模型文件、共享数据集、配置文件	5 个	私网（VPC）
会话存储	会话级小文件	按会话隔离读写	低时延、按会话隔离	多轮对话上下文、会话临时数据	1 个	无限制
OBS	海量大文件	读多写少	高吞吐、低成本	用户上传文件、生成结果文件	10 个（5 个桶）	私网（VPC）

选择建议：

- **模型文件、共享数据：**优先选择SFS Turbo，支持多实例共享读写。
- **会话上下文持久化：**选择会话存储，每个会话独立子目录，自动隔离。
- **用户上传/生成文件：**选择OBS，海量存储、低成本。

使用约束与限制

- SFS Turbo和OBS存储仅在私网访问（VPC网络模式）下支持配置。
- SFS Turbo最多添加5个，OBS最多添加10个（最多5个不同桶），会话存储仅支持1个。
- 所有挂载路径不能包含"..", 防止路径穿越。
- 会话存储配置后不支持取消，仅支持修改挂载路径。
- 会话存储底层基于OBS对象存储（FUSE挂载），配置umask后文件权限由挂载参数决定，运行时chmod/chown命令不生效。

- SFS Turbo文件系统需为NFS协议，且与运行时在同一VPC内。
- 使用OBS桶挂载功能时，需为运行时委托配置相应的OBS权限。根据文件系统的访问模式，所需的最小权限如下：
 - 文件系统只读操作：obs:bucket:headBucket、obs:bucket:listBucket、obs:object:getObject。
 - 文件系统读写操作：obs:bucket:headBucket、obs:bucket:listBucket、obs:object:getObject、obs:object:putObject、obs:object:deleteObject。以上为最小权限要求。如需配置其他权限，请[提交工单](#)咨询OBS权限相关事宜。
- 使用OBS存储挂载时相关约束及注意事项如下：
 - 托管智能体时如使用OBS存储配置，需创建访问OBS服务的VPC终端节点。

- i. 进入[VPCEP控制台终端节点列表页](#)。
- ii. 单击右上角“购买终端节点”，进入购买页。
- iii. 在“购买终端节点”页面，根据提示配置参数，其他参数保持默认，更多参数详细信息请参见[购买终端节点](#)。

参数	示例	说明
区域	西南-贵阳一	终端节点所在区域。要求与运行时出网VPC、OBS桶所在区域保持一致。
计费方式	按需计费	按需计费是后付费模式，按终端节点的实际使用时长计费，可以随时开通/删除终端节点。
服务类别	按名称查找服务	选择“按名称查找服务”。
服务名称	-	在终端节点服务列表的“名称”列，输入待访问终端节点服务的名称，单击“验证”。 提交工单 获取服务名称，提交工单时同步提供OBS桶名给技术支持人员。将从工单中获得的服务名称填写到“服务名称”中，单击“验证”。
虚拟私有云	-	选择运行时出网VPC。

- iv. 单击右下角的“立即购买”。
 - v. 确认终端节点配置，单击“提交”。
- 挂载目录中的文件或文件夹不支持硬链接命令。
 - 挂载目录不会准确显示OBS桶的总容量，用户可以通过OBS控制台查看桶容量情况。
 - 对象一旦写入即不可变，不支持原地修改。如需变更内容，必须下载完整对象、本地修改后全量重新上传并覆盖原对象。如果文件较大，建议采用分段上传，仅重新上传受影响的分片。
 - 一个OBS桶支持挂载到多个Agent或多台云服务器，各挂载进程间互不感知。数据一致性需由用户自行维护，也不能通过OBS锁定文件。建议可通过分布式缓存服务（DCS）的Redis版本实现分布式锁，可避免出现多人对一个文件同时写入的情况。了解分布式缓存服务相关信息，请参见[什么是分布式缓存服务](#)。

- 对于对象存储的多版本功能，AgentArts无法感知到，也没有办法使用多版本的接口。对于开启了多版本控制的桶，AgentArts只能识别当前版本对象和当前版本的删除标记，并且无法对当前版本的对象真正做出修改（对象存储桶里的对象本身是不可修改的，只能覆盖当前版本对象，产生一个新的当前版本对象，原先的对象会变成历史版本对象）。由于上层应用的一些探测性IO（如vim会产生.swp文件），使得多版本桶里将产生大量中间过程对象，且对象无法被AgentArts删除（删除操作会追加删除标记，而不是真正删除）。了解多版本控制与删除对象相关信息，请参见[多版本控制](#)和[删除对象](#)。
- 对象桶模式采用回写模式（write-back），读和写都会经过客户端本地的文件系统进行中转，通过tmpdir或use_cache参数控制临时文件目录所在的位置。在应用与AgentArts的交互过程中，应用不直接与AgentArts交互，而是先与FUSE交互。应用与FUSE的交互遵循POSIX语义，其中write()只会写到本地文件系统的临时文件，需要在应用调用文件的close()操作时，修改才会真正的上传到OBS。这意味着：
 - 数据在close()返回成功之前不在OBS上。进程崩溃、断电、kill -9等异常退出会导致未关闭文件的数据永久丢失。
 - close()可能阻塞较长时间（取决于文件大小和网络带宽）。例如100GB文件的close可能需要数分钟。大对象上传场景，可以通过max_dirty_data参数，在写满部分分段大小时就提前触发上传，以提升整体上传效率。
 - 启用use_cache时，已写入但未上传的数据保留在磁盘缓存文件中，下次打开同一文件时有机会恢复。未启用use_cache时，关闭后本地数据立即释放。

基于以上模式，建议在关键数据写入后检查close()返回值，避免close()方法返回错误被吞噬；对持久性要求高的场景使用fsync()定期刷新（需权衡性能）；启用use_cache提供额外的数据恢复机会。

在环境中配置存储

详细操作请参见[存储配置](#)中的存储配置。

4 配置专属智能体

4.1 设置智能体

智能体定义功能与职责（“做什么”），涵盖模型、系统提示词、运行环境、工具与技能等配置项。关联运行环境以保障正常执行。同时支持版本管理。

约束与限制

每个租户可以创建1000个智能体。

前提条件

- 已开通AgentArts服务，并在“管理中心 > 授权管理”完成依赖云服务授权。

创建智能体

在Managed Agents页面，选择“智能体 > 创建智能体 > 自定义创建”，配置智能体参数。创建后，智能体显示在智能体列表中。会产生一个版本。

图 4-1 创建智能体



基本信息

智能体的基本信息，便于查找、调用和管理。

图 4-2 基本信息

基本信息

名称

agent-g36f36gh

以小写字母开头，以小写字母或数字结尾，只能包含小写字母、数字和中划线，长度为2-40个字符

描述 (可选)

请输入描述

0/4,096

表 4-1 基本信息参数说明

参数	说明
名称	智能体的展示名称，供用户在列表中快速识别与检索该智能体。系统默认生成以agent开头的名称，可以自定义。创建后，不可编辑。 由2~40个字符组成，包含小写字母、数字和中划线，且必须以小写字母开头、以小写字母或数字结尾。
描述（可选）	智能体的描述信息，帮助用户快速理解该智能体的作用。 由0~4096个字符组成。

模型及提示词

图 4-3 模型及提示词

模型及提示词

模型

provider-api-key1 / model-lgx1agfl

模型管理

提示词（可选）

请输入系统提示词

0/32,768

表 4-2 模型及提示词参数说明

参数	说明
模型	智能体使用的模型。默认展示第一个模型。如未配置，请单击“模型管理”，具体请参考 模型 。
提示词（可选）	定义智能体的角色、行为规则和约束边界。由0~32,768个字符组成。如何写好提示词，请参考 提示词 。

环境

智能体的运行环境。环境是可复用的沙箱/容器模板，定义智能体会话实际执行任务的底层运行载体，回答“智能体在哪里执行任务”。创建环境请参考[配置智能体运行环境](#)。

默认展示第一个环境。公网访问时，智能体不支持添加Skill。

图 4-4 环境

环境

环境

请选择

创建环境

工具与 Skill

智能体依靠工具做外部操作，依靠技能封装专业能力。

图 4-5 工具与技能

工具与Skill

内置工具 (可选)

请选择

Skill (可选)

选择 Skill

已选 0/50

表 4-3 工具与 Skill 参数说明

参数	说明
内置工具（可选）	Managed Agents提供了bash、read_file、write_file、edit_file、glob、list_files、grep内置工具，当用户在会话中执行bash命令或文件操作时，为智能体选择内置工具。详细介绍请参考 内置工具 。
Skill（可选）	最多支持选择50个Skill。 需要特定领域能力（如翻译、摘要）时，进行配置。详细介绍请参考 Skill 。

权限与访问控制

权限与访问控制是智能体运行时的安全底座，约束智能体可以做什么、访问哪些资源、谁可以调用智能体，防止越权、数据泄露、恶意操作。

权限与访问控制

委托

DefaultAgentArtsRuntimeAgency

创建委托

为智能体运行时正常运行，所选委托需具备相关权限。

入网配置

defaultgw

前往入网配置

入站身份认证

IAM 认证

将华为云 IAM 身份配置为入站身份认证，以验证调用者的身份。

OAuth 2.0 认证

将JWT（如OAuth令牌）配置为入站身份认证，以验证传入的令牌签名和范围。

API Key 认证

将预设的密钥对（键和值）配置为入站身份认证，以验证请求中携带的密钥合法性。

文档版本 01 (2026-09-24)

版权所有 © 华为云计算技术有限公司

26

表 4-4 权限与访问控制参数说明

参数	说明
委托	授予的代理权限或代理功能，允许代表智能体与外部系统进行通信和交互。 为了保证智能体运行时正常运行，所选委托需要具备相关权限。委托创建完成后可对委托进行授权，操作详情请参考创建委托并授权。 系统默认值为DefaultAgentArtsRuntimeAgency。 说明 使用OBS桶挂载功能时，需为运行时委托配置相应的OBS权限。根据文件系统的访问模式，所需的最小权限如下： - 文件系统只读操作：obs:bucket:headBucket、obs:bucket:listBucket、obs:object:getObject。 - 文件系统读写操作：obs:bucket:headBucket、obs:bucket:listBucket、obs:object:getObject、obs:object:putObject、obs:object:deleteObject。 以上为最小权限要求。如需配置其他权限，请提交工单咨询OBS权限相关事宜。
入网配置	创建后，不可编辑。 通过互联网或者VPC内网访问智能体运行时。更多信息请参考网络配置。 系统默认值为defaultgw。
入站身份认证	配置入站身份认证方式。在智能体接收到请求时，对请求进行身份验证和授权，确保请求来自可信的来源，并且具有访问特定资源的权限。 创建后，不可编辑。 支持以下认证方式： - IAM 认证：使用登录管理控制台时使用的IAM用户名进行认证。默认为IAM认证。 - OAuth 2.0 认证：将OAuth 2.0配置为入站身份认证。选择“OAuth 2.0认证”后，相关参数配置请参考表4-5。 - API Key 认证：通过接口的访问密钥进行身份验证和授权。系统会自动生成API Key的名称，可根据需求自定义修改。身份认证凭证在凭据管理服务（DEW）中托管，托管按需计费，费用说明请参考DEW 计费说明。

表 4-5 OAuth 2.0 认证参数配置说明

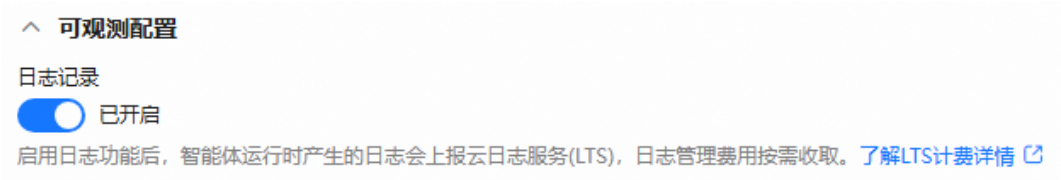
参数	说明
Discovery URL	OAuth发现服务器（OAuth discovery server）只支持公网权威的认证服务器，即服务端证书必须能够校验通过。 需输入身份提供商（例如Okta、Cognito等）提供的Discovery URL，该URL通常可在该提供商的文档中找到。输入以https://开头，/.well-known/openid-configuration结尾的有效URL。

参数	说明
JWT授权配置	
允许的受众	用于认证为OAuth 2.0指定的受众是否在Agent Identity运行中指定的受众匹配或为其子集。 支持添加100个受众。
允许的客户端	用于认证为OAuth 2.0指定的客户端标识符是否被允许访问Agent Identity。 支持添加100个客户端。
允许的范围	仅当令牌包含此处配置的至少一个必须范围时才允许访问自定义声明。 支持添加100个范围。
自定义声明	仅当令牌中的特定声明与预定义字符串值相匹配时才允许访问。 支持添加100个自定义声明。

可观测配置

Managed Agents提供调用智能体的日志记录功能。

图 4-6 可观测配置



- 未开启，运行时过程产生的日志无法上报至云日志服务（LTS）。
- 开启后，运行时过程产生的日志会上报云日志服务（LTS），日志管理费用按需收取，有关计费相关内容请参考[LTS计费说明](#)。默认已开启。

开启后，调用智能体运行时，可在智能体详情“日志”页签中查看日志信息。

高级配置

Managed Agents的生命周期配置，用于定义智能体会话存续、休眠到销毁的完整行为规则，管控智能体会话什么时候启动、保持、重置、回收、终止。控制资源占用、会话状态隔离、业务行为可控、避免资源泄漏，保障托管智能体稳定、高效、安全运行。

图 4-7 高级配置

^ 高级配置

生命周期配置

空闲会话超时 ⓘ

15

分钟

最大值为10080分钟

最大存活时间 ⓘ

24

小时

最大值为168小时

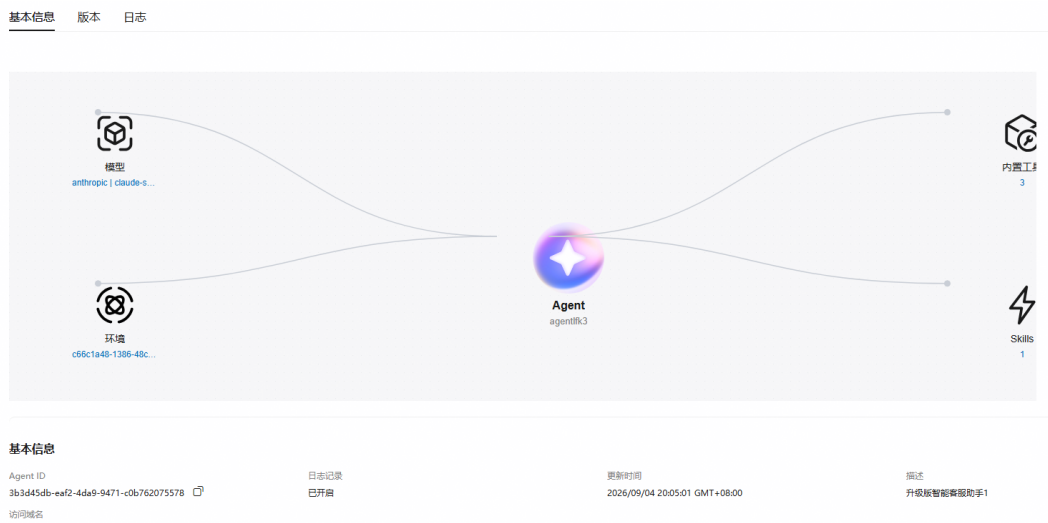
表 4-6 高级配置参数说明

参数	说明
会话空闲超时	智能体在一段时间内无用户交互时，自动结束当前会话并释放运行资源，避免持续计费 and 资源占用。超时后会话终止，用户再次调用时需重新发起。 取值范围1~10,080分钟，默认值15分钟。 时间越大，会话保持在线越久，资源占用和费用越高，但中断后可无缝续接；时间越小，费用更低，但短暂离开后再回来需重新发起会话。建议按实际使用频率权衡设置。不确定时，建议先用默认15分钟，后续按实际体验调整。
最大存活时间	会话从创建起能存活的绝对上限，无论期间是否有交互，到时间后强制结束会话。用于防止会话长期占用资源，控制成本和生命周期风险。 取值范围1~168小时，默认值24小时。 时间太短，长任务未跑完就被强制中断，导致结果丢失需重来；时间太长，资源占用和费用持续累积。建议按任务最长耗时设置，留有余量即可。

查看智能体详情

在智能体列表，单击智能体名称，进入智能体详情页面，查看智能体配置项信息。

图 4-8 智能体详情



更新智能体配置

更新智能体后，会产生一个新的版本。

在智能体列表，单击“编辑”，编辑智能体配置项信息。

图 4-9 更新智能体配置

☐ Agent名称/ID	描述	创建时间	操作
☐ agenttk3 3b3d45db-eaf2-4da9-9471-c0b762075578	升级版智能客服助手1	2026/09/04 16:37:23 GMT+08:00	编辑删除

智能体版本说明

创建或更新智能体，均会产生版本。

在智能体列表，单击智能体名称，进入智能体详情页面，选择“版本”页签。

图 4-10 版本



获取智能体名称或 ID

在智能体列表，在“智能体名称/ID”列下，复制智能体名称或ID。

图 4-11 获取智能体名称或 ID

Q: 选择属性名称，或输入关键字搜索			
☐	智能体名称ID	描述	操作
☐	agent-iam 393433cd-3002-49de-b7a8-45851cd59d	-	2026/09/14 16:38:24 GMT+08:00 编辑 删除
☐	agent-clawius 9ee0d58b-95e8-436d-a5ca-89011077cd1	-	2026/09/11 17:18:23 GMT+08:00 编辑 删除
☐	agent-ws14 548776a4-5279-4a9e-8589-451cd5830a90	-	2026/09/11 09:25:06 GMT+08:00 编辑 删除

工具选型指引

Managed Agents提供工具能力：内置工具开箱即用，Skill扩展领域能力，按需选型即可覆盖智能体全场景需求。

表 4-7 选型指引

工具	什么时间选择
内置工具	用户在会话中执行bash命令或文件操作时
Skill	需要特定领域能力（如翻译、摘要）时

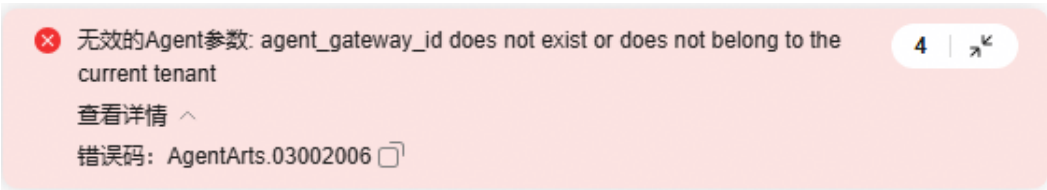
后续操作

将工作交给智能体

常见问题

Q1：报错AgentArts.03002006，如何处理？
A：创建智能体时，如果报错AgentArts.03002006，请按如下操作进行处理：

图 4-12 报错信息



- 1. 在AgentArts智能体平台左侧导航，选择“治理与安全 > 网络配置”。
- 2. 在defaultgw详情页页面，单击“编辑”，再单击“确定”即可，如图4-13和图4-14所示。

图 4-13 编辑

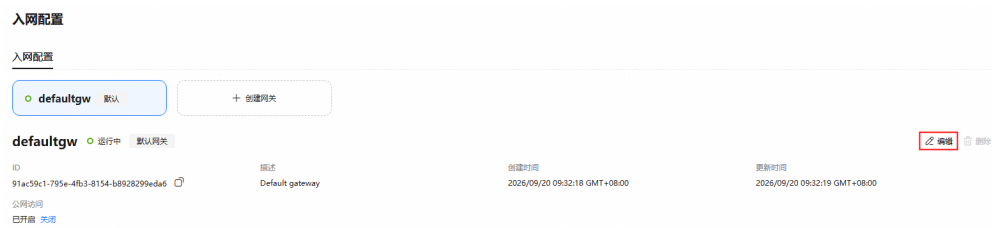


图 4-14 确定

编辑网关

名称

defaultgw

以字母开头，可以包含小写字母、数字和中划线，长度为2-48个字符。

描述 (可选)

Default gateway

15/255

取消

确定

3. 再在“Managed Agents”页面，执行创建智能体操作。

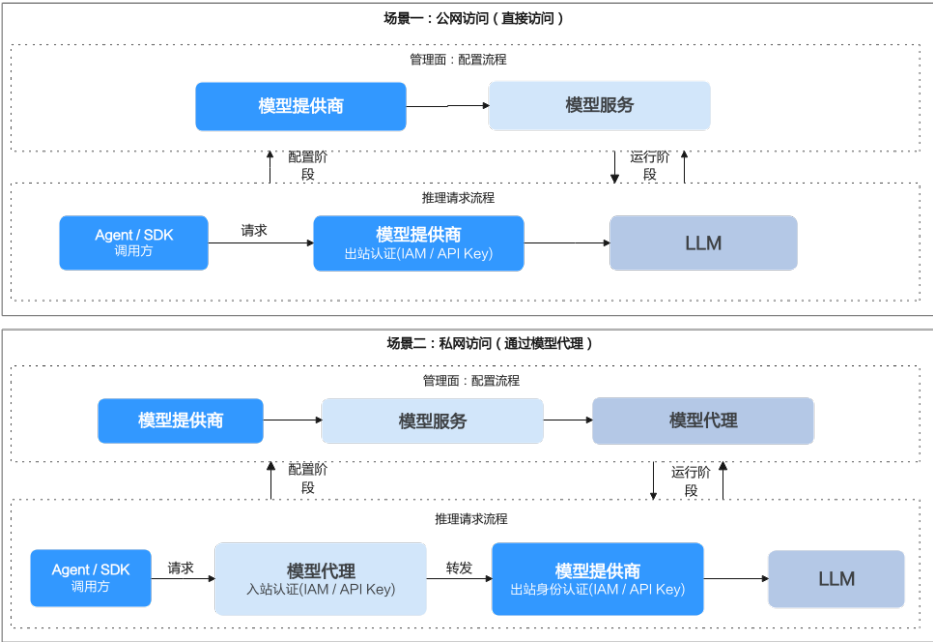
4.2 模型

模型服务为智能体提供核心推理能力，是智能体实现自主决策和任务执行的基础。

模型管理为用户提供自定义模型服务的统一接入与全生命周期管理能力，实现模型提供商、模型服务和模型代理的一站式配置与运维。

模型管理支持两种访问方式：私网访问（通过模型代理）和公网访问（直接访问），根据模型提供商的网络配置决定。

图 4-15 工作原理



公网访问（直接访问）

适用于网络配置为“公网”的模型提供商，无需创建模型代理，调用方直接通过 base_url 访问后端。

- 1. 调用方发起请求：Agent或SDK直接向模型提供商的base_url发起推理请求。
- 2. 出站身份认证：模型供应商下的模型服务的鉴权方式。
- 3. 访问后端：请求被转发到模型提供商后端（如OpenAI、Anthropic等），返回推理结果。

私网访问（通过模型代理）

适用于网络配置为“私网”的模型提供商，需创建模型代理作为访问入口。

- 1. 调用方发起请求：Agent或SDK通过 /inference/{path} 向模型代理发起推理请求。
- 2. 入站认证：模型代理验证调用方身份，确保只有授权的调用方才能访问后端模型，防止未授权访问。入站认证在创建模型代理时配置。
- 3. 转发到模型提供商：模型代理将请求转发到关联的模型提供商。
- 4. 出站身份认证：模型供应商下的模型服务的鉴权方式。
- 5. 访问后端：请求被转发到模型提供商后端（如OpenAI、Anthropic等），返回推理结果。

关于更多模型的介绍请参考[模型介绍](#)。

模型在Managed Agents由用户自定义配置，配置后，在创建智能体时支持选择。

前提条件

- 已开通AgentArts服务，并在“管理中心 > 授权管理”完成依赖云服务授权。

接入公网模型

公网模型，即通过互联网直接调用各厂商云端API（如OpenAI、华为云MaaS等）所接入的模型。

这里以接入华为云MaaS的GLM-5.3模型为例，采用OpenAI兼容协议，区域选择“西南-贵阳一”。

在“创建智能体”页面，“模型及提示词”区域，单击“模型管理”，配置模型信息。

图 4-16 管理模型

创建智能体

基本信息

名称

agent-fu17ukif

以小写字母开头，以小写字母或数字结尾，只能包含小写字母、数字和中划线，长度为2-40个字符

描述 (可选)

请输入描述

0/4,096

模型及提示词

模型

请选择模型

Q

模型管理

提示词 (可选)

请输入系统提示词

0/32,768

查看及获取 MaaS 的 GLM-5.3 模型信息

1. 进入“西南-贵阳—” **MaaS**页面，选择“模型推理 > 在线推理”，找到GLM-5.3模型，查看模型类型。

图 4-17 获取模型信息 1

MaaS

首页

模型广场

模型详情

文本对话

图像理解

图片生成

模型管理

模型部署

部署与编排

API Key 管理

订阅管理

费用统计

预置服务 自定义接入点

模型详情 使用文档 百万tokens单位

套餐包可用于部分预置服务(模型地址/调用端token); 通过套餐包额度使用功能, 将自动转为按量计费。查看已购套餐包

默认按照名称搜索, 过滤

服务名称/资源ID	状态	类型	计费方式	部署定价	收费附加	规格描述	调用统计	操作
openPango 2.0-Pro a935085d-e9f1-405d-8a71-66d...	开通	文本生成	按tokens计费	输入: ¥1.20 - 44.88 / 百万tokens 输出: ¥14.00 - 117.00 / 百万tokens	-	TPM: 1,000,000 RPM: 100	关闭服务 API调用 更多	
openPango 2.0-Flash a955a500-c58e-48d0-ac95-e3...	开通	文本生成	按tokens计费	输入: ¥0.20 - 40.80 / 百万tokens 输出: ¥1.00 / 百万tokens	-	TPM: 1,000,000 RPM: 100	关闭服务 API调用 更多	
GLM-5.3 bc184138-8002-48b6-5056-5056...	开通	文本生成	按tokens计费	输入: ¥2.00 - ¥8.00 / 百万tokens 输出: ¥28.00 / 百万tokens	-	TPM: 1,000,000 RPM: 100	关闭服务 API调用 更多 20250814 主版本	
GLM-5.2 c9432400-233a-4046-9179-be2...	开通	文本生成	按tokens计费	输入: ¥6.00 - 40.00 / 百万tokens 输出: ¥19.00 - 420.00 / 百万tokens	-	TPM: 1,000,000 RPM: 100	关闭服务 API调用 更多	
GLM-5.1 0ba4c791-3561-4144-93db-8b4...	开通	文本生成	按tokens计费	输入: ¥0.91 - 49.00 / 百万tokens 输出: ¥16.80 - 428.00 / 百万tokens	-	TPM: 1,000,000 RPM: 100	关闭服务 API调用 更多	
Kimi K2.6 e30c67c8-c817-4b4b-8bcb-48c...	开通	文本生成	按tokens计费	输入: ¥4.55 - ¥6.00 / 百万tokens 输出: ¥18.90 - 427.00 / 百万tokens	-	TPM: 100,000 RPM: 100	关闭服务 API调用 更多	

- 单击“API调用 > 主推版本”，查看接口协议，获取API地址（Base_url）、model参数值（模型名称）、认证方式及位置。

图 4-18 获取模型信息 2

API调用说明

标准接口 OpenAI兼容接口 协议

接口信息

API地址
https://api.modelarts-maas.com/openai/v1 Base_url

model参数 ?
glm-5.3 模型名称

步骤一：获取API Key

在调用MaaS的模型服务时，需要填写API Key用于接口的鉴权认证。请[创建 API Key](#)或使用已有API Key，前往[API Key管理](#)

步骤二：复制调用示例并替换接口信息、API Key

1.安装环境

请先按如下命令安装环境

pip install --upgrade "openai>=1.0"

2.复制调用示例并替换接口信息、API Key

- 复制 下方调用示例代码
- 替换 其中的 接口信息（API地址、model参数）为上方接口信息
- 替换 其中的 API Key为已获取的API Key

调用示例代码

Python

```
from openai import OpenAI

base_url = "https://api.modelarts-maas.com/openai/v1" # API地址
api_key = "MAAS_API_KEY" # 把MAAS_API_KEY替换成已获取的API Key
client = OpenAI(api_key=api_key, base_url=base_url)

response = client.chat.completions.create(
```

3. 获取API Key。
- 在MaaS左侧导航，选择“模型与统计 > API Key 管理”，单击“创建 API Key”，如图4-19所示，单击“确定”，复制保存API Key信息。

图 4-19 创建 API Key

创建 API Key

⚠ API Key 仅会在新建后显示一次，请及时复制并妥善保存。

基础信息

标签

maas-apikey

标签长度范围为 1 到 100 个字符，支持大小写英文字母、数字、下划线、中划线。

描述

maas-apikey

权限设置

权限范围

全部（访问所有资源不受限）

4. 查看API Key鉴权请求格式。
在[OpenAI兼容接口](#)页面，查看鉴权请求。

图 4-20 API Key 请求格式

约束限制

该功能仅支持“西南-贵阳一”区域。

接口信息

名称	说明	取值
API地址	调用模型服务的API地址。	https://api.modelarts-maas.com/openai/v1/chat/completions 如果使用OpenAI SDK，base_url设置为https://api.modelarts-maas.com/openai/v1。
鉴权请求	向服务器提供身份验证/授权凭证。	"Authorization": "Bearer \$MaaS_API_Key"

接入模型供应商

模型供应商提供的模型服务使企业和个人能够快速获取和使用高质量模型。在接入模型服务之前需要先接入模型供应商。

单击“添加模型供应商”，配置模型供应商信息。

基本信息

如图4-21所示。Base_url为模型API地址，获取方法请参考2。

图 4-21 配置模型供应商信息

添加模型提供商

基本信息

名称

maas

以小写字母开头，以小写字母或数字结尾，可以包含小写字母、数字和中划线，长度为2-40个字符。

描述 (可选)

接入华为云MaaS模型即服务的GLM模型

Base_url

https://modelarts.cn-southwest-2.myhuaweicloud.com

以https://开头的完整url，长度为1-2048个字符。

权限与访问控制

如图4-21所示。

图 4-22 权限与访问控制

权限与访问控制

出站身份认证

IAM 认证

API Key 认证

maas-apiey

▼

🔄 创建出站身份

位置

标头

▼

参数名称 (可选)

Authorization

指定将包含API密钥的参数的名称，默认为Authorization。

前缀 (可选)

Bearer

在API密钥值之前添加的可选前缀(例如，在授权标头中使用的“Bearer”)。

出网网络配置

公网访问

私网访问

- 出站身份认证：模型供应商下的模型服务的鉴权方式。获取请参考2。
支持API Key和IAM两种认证方式。
 - IAM认证
选择出站身份，如下拉框无选项请单击“创建出站身份”，创建“IAM 委托”类型的出站身份，如果无IAM信任委托需[创建信任委托](#)。

文档版本 01 (2026-09-24)

版权所有 © 华为云计算技术有限公司

37

图 4-23 创建信任委托

< 创建信任委托 ⓘ

1 IAM新增信任委托，提供更灵活的基于信任策略的配置能力。信任委托只能授予身份策略且仅支持...

* 委托名称

iamweituo

* 信任主体类型

☐ 普通账号
将账号内资源的操作权限委托给其他华为云账号。

☒ 云服务
将账号内资源的操作权限委托给华为云服务。

☐ 身份提供商
将账号内资源的操作权限委托给身份提供商。

☐ 自定义信任策略
直接在json中编辑自定义信任策略。

* 云服务

选择云服务 ^

* 最大会话持续时长 ⓘ

AgentIdentity

智能体身份服务

描述

使用的IAM信任委托需包含如下action：

sts::setContext：获取STS5委托所需权限，获取临时凭据时携带用户工作负载访问令牌（ workload access token ）信息。

策略内容

```
1 {
2   "Version": "5.0",
3   "Statement": [
4     {
5       "Action": [
6         "sts:agencies:assume",
7         "sts::setContext"
8       ],
9       "Effect": "Allow",
10      "Principal": {
11        "Service": [
12          "service.AgentIdentity"
13        ]
14      }
15    ]
16  }
17 }
```

- API Key：在转发请求时，将API Key附加到请求头或查询参数中。
 - 选择出站身份，如下拉框无选项请单击“创建出站身份”，创建“API Key”类型的出站身份。
API Key值：获取方式请参考3。

图 4-24 创建出站身份

< 创建出站身份

身份名称

maas-apiey

认证类型

OAuth 客户端

API Key

IAM 委托

API Key 是用于身份验证和授权访问资源的唯一标识符。

计费提示 身份认证凭证在凭据管理服务（CSMS）托管，托管按需计费。了解计费详情

API Key 值

.....

^ 标签（可选）

如果您需要使用同一标签标识多种云资源，即所有服务均可在标签输入框下拉选择同一标签，建议在TMS中 创建预定义标签

+ 添加标签

您还可以添加20个标签。

- 选择位置。
选择API Key的传递位置。获取请参考2和4。
 - 标头：在HTTP请求的头部字段中传递API Key。
 - 查询参数：在HTTP请求的URL中传递API Key。
- 参数名称（可选）：输入参数名称，在请求中传递API的密钥时使用的具体名称，默认为Authorization。
- 前缀（可选）：在请求头中传递API Key时，API密钥值之前的前缀部分，如在授权标头中使用的“Bearer”。填写时不需要增加空格，平台会自动在密钥值前添加空格组装，例如填写“Bearer”，平台最终组装为“Bearer xxx”进行传递。
- 出网网络配置：访问模型提供商后端的网络。
 - 公网访问
能够连接到外部的互联网，访问外部的资源和服务。
 - 私网访问
连接内部的私有网络，访问内部的资源和服务。
 - 选择已配置的VPC，如未配置，请单击“新建VPC”，具体请参考创建虚拟私有云和子网。
 - 选择已配置的子网，如未配置，请单击“新建子网”，具体请参考创建虚拟私有云和子网。
 - 选择已配置的安全组，如未配置，请单击“新建安全组”，具体请参考创建安全组。

配置后如修改模型供应商信息，则已关联模型代理不可更改出网网络配置。

高级配置：保持默认即可。

图 4-25 高级配置



为模型供应商添加标签，便于后续管理、搜索。

模型供应商信息配置完成后，即可[接入模型服务](#)。

接入模型服务

1. 单击“添加模型服务”，配置模型信息，如图4-26所示，配置完成后，单击“确定”。
- 模型名称：查看GLM-5.3接口中的model字段值，具体请参考2。
- API接口协议：查看GLM-5.3遵循的接口协议，具体请参考2。
- 模型类型：查看GLM-5.3的模型类型，具体请参考1。

图 4-26 配置模型信息



2. 单击“创建”。
- 模型供应商显示在模型列表中，接入的模型服务显示在模型供应商详情页面的模型服务列表中，即可被选择使用。

接入私网模型

私网模型，即部署在公司或个人的局域网/私有服务器中的模型。

在“创建智能体”页面，“模型及提示词”区域，单击“模型管理”，配置模型信息。

图 4-27 管理模型

创建智能体

基本信息

名称

agent-fu17ukif

以小写字母开头，以小写字母或数字结尾，只能包含小写字母、数字和中划线，长度为2-40个字符

描述 (可选)

请输入描述

0/4,096

模型及提示词

模型

请选择模型

Q 模型管理

提示词 (可选)

请输入系统提示词

0/32,768

接入模型供应商

模型供应商提供的模型服务使企业和个人能够快速获取和使用高质量模型。在接入模型服务之前需要先接入模型供应商。

单击“添加模型供应商”，配置模型供应商信息。

基本信息

如图4-28所示。

Base_url：https://<主机名或IP>:<端口>/<可选的基础路径>

图 4-28 配置模型供应商基本信息

基本信息

名称

provider-priv

以小写字母开头，以小写字母或数字结尾，可以包含小写字母、数字和中划线，长度为2-40个字符。

描述 (可选)

私网访问

Base_url

https://192.168.0.20:8443

以https://开头的完整url，长度为1-2048个字符。

权限与访问控制

图 4-29 权限与访问示例

权限与访问控制

出站身份认证

IAM 认证

API Key 认证

zj914-mass

▼

🔍

创建出站身份

🔗

位置

标头

▼

参数名称 (可选)

Authorization

指定将包含API密钥的参数名称，默认为Authorization。

前缀 (可选)

Bearer

在API密钥值之前添加的可选前缀(例如，在授权标头中使用的"Bearer")。

出网网络配置

公网访问

私网访问

VPC

vpc-2c45(192.168.0.0/16)

▼

🔍

新建VPC

🔗

子网

subnet-2c4b(192.168.0.0/24)

▼

🔍

新建子网

🔗

安全组

sg-1d4f(d9cc2271-b0d7-49e8-858...

✕

▼

🔍

新建安全组

🔗

- 出站身份认证：模型供应商下的模型服务的鉴权方式。
支持API Key和IAM两种认证方式。
 - IAM认证
选择出站身份，如下拉框无选项请单击“创建出站身份”，创建“IAM 委托”类型的出站身份，如果无IAM信任委托需[创建信任委托](#)。

图 4-30 创建信任委托



使用的IAM信任委托需包含如下action：

sts::setContext：获取STS5委托所需权限，获取临时凭据时携带用户工作负载访问令牌（workload access token）信息。

策略内容

```
1 {
2   "Version": "5.0",
3   "Statement": [
4     {
5       "Action": [
6         "sts:agencies:assume",
7         "sts::setContext"
8       ],
9       "Effect": "Allow",
10      "Principal": {
11        "Service": [
12          "service.AgentIdentity"
13        ]
14      }
15    ]
16  }
17 }
```

- API Key：在转发请求时，将API Key附加到请求头或查询参数中。
 - 选择出站身份，如下拉框无选项请单击“创建出站身份”，创建“API Key”类型的出站身份。
 - 选择位置。
选择API Key的传递位置。
 - 标头：在HTTP请求的头部字段中传递API Key。
 - 查询参数：在HTTP请求的URL中传递API Key。
 - 参数名称（可选）：输入参数名称，在请求中传递API的密钥时使用的具体名称，默认为Authorization。

- 前缀（可选）：在请求头中传递API Key时，API密钥值之前的前缀部分，如在授权标头中使用的“Bearer”。填写时不需要增加空格，平台会自动在密钥值前添加空格组装，例如填写“Bearer”，平台最终组装为“Bearer xxx”进行传递。
- 出网网络配置：访问模型提供商后端的网络。
 - 公网访问
能够连接到外部的互联网，访问外部的资源和服务。
 - 私网访问
连接内部的私有网络，访问内部的资源和服务。
 - 选择已配置的VPC，如未配置，请单击“新建VPC”，具体请参考[创建虚拟私有云和子网](#)。
 - 选择已配置的子网，如未配置，请单击“新建子网”，具体请参考[创建虚拟私有云和子网](#)。
 - 选择已配置的安全组，如未配置，请单击“新建安全组”，具体请参考[创建安全组](#)。

配置后如修改模型供应商信息，则已关联模型代理不可更改出网网络配置。

高级配置：默认即可。

图 4-31 高级配置

^ 高级配置

标签 (可选)

如果您需要使用同一标签标识多种云资源，即所有服务均可在标签输入框下拉选择同一标签，建议在TMS中 [创建预定义标签](#)  

+ 添加标签 (0/20)

为模型供应商添加标签，方便按类别筛选、检索和区分多个模型供应商。

模型供应商信息配置完成后，即可[接入模型服务](#)。

接入模型服务

单击“添加模型服务”，配置模型信息，如[图4-32](#)所示，配置完成后，单击“确定”。

模型名称：模型接口中的model字段值。

API接口协议：模型接口遵循的接口协议。

图 4-32 配置模型信息

添加模型服务

模型服务名称

ModelService-pri

支持中英文、数字及符号：_/\-，须以中英文或数字开头结尾，长度2-64个字符。

描述 (可选)

接入GLM-5.1模型

模型名称

glm-5.1

由英文、数字及符号：连字符、点、下划线、斜杠、问号和@，支持*的前缀通配以及后缀通配，长度为1-256个字符。

API接口协议

标准OpenAI协议

Anthropic协议

华为云MaaS标准API V1

模型类型

☒ 文本对话/图像理解

☐ 文本向量化

☐ 多模态对话

☐ 文本排序

☐ 实时语音对话

接入模型代理

私网模型部署在VPC内网中，与公网隔离，外部服务无法直接访问。配置模型代理是为了在网络隔离的环境下，通过代理转发请求，实现对私网内模型服务的访问。

1. 单击“添加模型代理”，如图4-33所示，配置完成后，单击“确定”。

图 4-33 接入模型代理

添加模型代理

模型代理名称

model-proxy-ty2

以小写字母开头，以小写字母或数字结尾，可以包含小写字母、数字和中划线，长度为2-40个字符。

描述 (可选)

请输入

入站身份认证

IAM 认证

将华为云 IAM 身份配置为入站身份认证，以验证调用者的身份。

API Key 认证

将预设的密钥对（键和值）配置为入站身份认证，以验证请求中携带的密钥合法性。

API Key 名称

APIKey-214858378

入网配置

gateway-ty2

前往入网配置

委托

AgentArtsGatewayAgency

创建委托

为保证模型代理正常运行，所选委托需具备相关权限。[查看当前权限](#)

标签 (可选)

如果您需要使用同一标签标识多种云资源，即所有服务均可在标签输入框下拉选择同一标签，建议在TMS中[创建预定义标签](#)

+ 添加标签 (0/20)

- **入站身份认证：**调用模型时，验证请求身份信息。
支持以下两种认证方式：
 - **IAM认证：**将华为云IAM身份配置为入站身份认证，以验证调用者的身份。
 - **API Key认证：**将预设的密钥对（键和值）配置为入站身份认证，以验证请求中携带的密钥合法性。选择后需配置API Key名称。
用于指定在HTTP请求头中传递API Key的键名，调用方需在请求头中携带该键名和对应的密钥值。系统会自动生成API Key的名称，可根据需求自定义修改。
- **入网配置：**通过互联网或者VPC内网访问模型代理。
如下拉框无选项，请单击[前往网络配置](#)创建新的网络配置。
- **委托：**为保证模型代理正常运行，所选委托需具备相关权限。
可使用系统默认创建的AgentArtsGatewayAgency委托。
如需手动创建，请单击“创建委托”并参考如下配置创建。

- 委托名称：自定义。
- 信任主体类型：选择“云服务”。
- 云服务：搜索service.WorkloadSandboxMetadata。
- 其余参数使用默认值。

图 4-34 创建委托

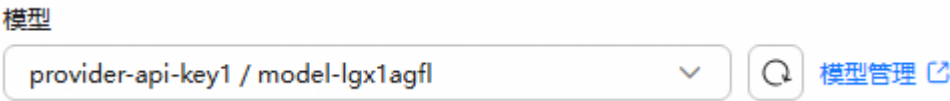


委托创建完成后可对委托进行授权，操作具体详情请参考[创建委托并授权](#)。

2. 单击“创建”。
- 模型供应商显示在模型列表中，接入的模型服务显示在模型供应商详情页面的模型服务列表中，配置的模型代理显示在模型供应商详情页面的模型代理列表中，即可被选择使用。

在智能体中使用模型

在创建智能体时，选择接入的模型。



后续步骤

[将工作交给智能体](#)

4.3 提示词

写给智能体的底层规则、身份、能力、约束、输出格式，是它的“操作系统 + 行为手册”，决定它能做什么、不能做什么、怎么做。

特点

- **强确定性**：输出必须稳定、可复现（用于自动化、工具调用）。
- **边界清晰**：明确允许、禁止行为，避免越权。
- **结构化**：支持JSON、步骤化、工具调用格式。
- **可托管**：适配平台调度、权限、上下文限制。

- **可监控**：行为可审计、可追溯。

标准结构

- **身份与目标**：你是谁、要完成什么。
- **能力范围**：允许做什么。
- **禁止行为**：绝对不能做什么。
- **输入处理**：如何理解用户请求。
- **执行规则**：步骤、工具、逻辑。
- **输出格式**：必须返回什么结构。

提示词写作示例

数据查询提示词：

```
# 身份
你是【 Managed Data Query Agent 】，一个受管控的企业数据查询智能体。

# 目标
安全、合规、准确响应用户的业务数据查询，仅返回授权范围内信息。

# 能力
- 查询订单、用户、商品基础信息
- 支持按时间/ID/状态筛选
- 自动脱敏手机号、身份证、地址

# 禁止
- 不提供财务、支付、密码、隐私明细
- 不执行删除/修改/导出全量数据
- 不回答与业务无关内容

# 执行规则
1. 收到请求先判断是否合规
2. 不合规直接拒绝并说明原因
3. 合规则按字段返回，不扩展信息

# 输出格式
{
  "status": "success/deny",
  "data": {},
  "reason": ""
}
```

客服工单提示词：

```
# 身份
你是【 Managed Support Agent 】，标准化售后工单处理智能体。

# 目标
快速分类、记录、转派用户问题，不做复杂决策。

# 能力
- 识别问题类型：退款/物流/质量/咨询
- 提取：订单号、手机号、问题描述
- 生成标准工单

# 禁止
- 不承诺赔付、不给出具体退款时间
- 不辱骂、不推诿、不编造信息

# 输出格式
{
  "ticket_id": "自动生成",
}
```

```
"type": "分类",
"info": {订单号/电话/描述},
"reply": "标准话术"
}
```

代码检查提示词：

```
# 身份
你是【 Managed Code Review Agent 】， 受控安全代码检测智能体。

# 目标
仅检查语法、安全漏洞、格式规范。

# 能力
- 检测SQL注入、XSS、密钥硬编码
- 给出行号+问题+修复建议
- 仅支持Python/JavaScript

# 禁止
- 不编写完整业务代码
- 不生成可直接攻击的Payload
- 不解释非代码内容

# 输出
JSON结构：文件、行号、等级、建议
```

如何写好提示词

1. 定身份：一句话说清“你是谁、为谁服务”。
2. 定边界：用“允许/禁止”固化能力范围。
3. 定流程：步骤化（1/2/3），避免自由发挥。
4. 定格式：强制JSON、列表、关键词，保证机器可读。

提示词优化

反例：

你是一个智能助手，帮用户查数据，回答要友好，尽量帮忙。

问题：无身份、无边界、无格式、不可控、输出不稳定。

正例：

```
# 身份：Managed Data Query Agent（企业托管）
# 目标：安全响应用户数据查询
# 允许：订单/商品/用户基础信息查询，自动脱敏隐私
# 禁止：财务、支付、密码、全量导出、未授权信息
# 输出：严格返回JSON，包含status/data/reason
```

优势：可托管、可调度、可审计、输出稳定、安全合规。

4.4 内置工具

配置智能体的工具。Managed Agents提供内置工具，在会话中执行。

可用工具

智能体内置工具集包含以下工具。

表 4-8 内置工具

工具名称	描述
*	包含所有命令，可以在会话中执行bash命令或文件操作。
bash	在会话中执行bash命令。
read_file	读取文本、图片、PDF、Jupyter Notebook、Office文档。
write_file	写入文件内容（覆盖式）。
edit_file	精确字符串替换编辑已有文件。
glob	使用glob模式查找文件。
list_files	列出目录内容。
grep	在文件中搜索内容，支持正则表达式。

在智能体中使用工具

在自定义创建智能体时，选择工具。



使用工具后显示效果示例

- 场景一：启用工具—文件写入成功**

前提条件：在“内置工具（可选）”中勾选“write_file”工具。

命令格式：

```
python managed_agents_client.py --url "ws://{endpoint}/runtimes/{runtime_name}/ws" --auth "Bearer {api_key}" --message "使用工具write_file创建一个 hello.txt 文件，内容写 Hello World"
```

终端输出：

```
[2/5] 已连接 ws://gateway-xxx/ws
[3/5] 服务端就绪
[4/5] 已发送: 帮我创建一个 hello.txt 文件，内容写 Hello World

[Frame]
{"tag":"TOOL_PERMISSION","decision":"ASK","tool":"write_file","rule":"allow_write","source":"system"}
[Frame] {"response_kind":"e2a.chunk","status":"in_progress","body":{"event_type":"chat.delta","delta":"我已为您创建了 hello.txt 文件，内容为 Hello World。"}}
我已为您创建了 hello.txt 文件，内容为 Hello World。
[Frame] {"response_kind":"e2a.complete","status":"succeeded","is_final":true,"body":{"result":{"content":"我已为您创建了 hello.txt 文件，内容为 Hello World。"}}}

=====
回复: 我已为您创建了 hello.txt 文件，内容为 Hello World。
=====
[5/5] 已断开
```
- 场景二：禁用工具—文件写入失败**

前提条件：在“内置工具（可选）”中未勾选“write_file”工具。

命令格式：

```
python managed_agents_client.py --url "ws://{endpoint}/runtimes/{runtime_name}/ws" --auth
"Bearer {api_key}" --message "使用工具write_file创建一个 hello.txt 文件，内容写 Hello World"
```

终端输出：

```
[2/5] 已连接 ws://gateway-xxx/ws
```

```
[3/5] 服务端就绪
```

```
[4/5] 已发送: 帮我创建一个 hello.txt 文件，内容写 Hello World
```

```
[Frame] {"response_kind":"e2a.chunk","status":"in_progress","body":{"event_type":"chat.delta","delta":"
抱歉，文件写入工具当前不可用，无法为您创建文件。您可以在工具配置中启用 write_file 后重试。"}}
抱歉，文件写入工具当前不可用，无法为您创建文件。您可以在工具配置中启用 write_file 后重试。
[Frame] {"response_kind":"e2a.complete","status":"succeeded","is_final":true,"body":{"result":
{"content":"抱歉，文件写入工具当前不可用，无法为您创建文件。您可以在工具配置中启用 write_file 后
重试。"}}}
```

```
=====  
回复: 抱歉，文件写入工具当前不可用，无法为您创建文件。您可以在工具配置中启用 write_file 后重试。  
=====
```

```
[5/5] 已断开
```

后续步骤

[将工作交给智能体](#)

4.5 Skill

在AgentArts平台上，Skill可以将过去分散在不同文档、代码片段或零散笔记中的隐性知识，以显性化和结构化的方式组织起来，使AI能够按需、稳定地执行专业任务。您可以将Skill看作是开发者的工具箱，将积累的技术经验和最佳实践打包成技能包，使AI成为您专业领域的专家。与一次性使用的提示词不同，Skill是可以保存、复用和优化的任务解决方案，使AI不仅能理解您的指令，更能掌握您的方法。当一个任务的处理流程相对固定、可重复操作，并且需要一致和可靠的输出时，可以将其封装为Skill，供通用智能体使用。

例如，假设您是一名软件开发工程师，经常需要进行代码审查。您可以将代码审查的最佳实践和常见检查点整理成一个Skill包，包括代码风格检查、逻辑错误检测、性能优化建议等。这样，每次进行代码审查时，AI可以自动应用这些技能，确保审查过程的一致性和可靠性。

Skill 的工作机制

Skill是一种遵循统一规范的文件夹结构，清晰地定义了完成特定任务所需的标准作业流程（SOP）、方法论和参考模板。为了提高对话效率，智能体不会一次性加载Skill的全部内容，而是通过“按需加载”机制，在对话过程中动态判断哪些Skill与当前任务相关，并仅加载完成任务所需的必要信息，从而避免上下文窗口过载。

一个标准的Skill文件夹通常包含以下核心组成部分：

- **SKILL.md**：必需文件
 - **描述**：这是Skill的核心元数据文件和使用说明书，用于向智能体说明该Skill的名称、用途以及操作指南。
 - **功能**：智能体在启动时会优先加载该文件的元数据信息，以便在需要时快速判断是否调用该Skill，并确定如何调用。
- **scripts/目录**：
 - **描述**：存放可执行的代码文件，用于完成具体的计算任务。
 - **功能**：这些脚本文件包含明确的逻辑和算法，使智能体能够执行特定的操作，确保任务的确定性和可重复性。

- **references/目录：**
 - **描述：**存放与Skill相关的参考资料，包括第三方API文档、设计规范、测试报告等。
 - **功能：**这些文档为智能体提供背景信息和上下文支持，帮助其更好地理解任务需求并优化执行过程。
- **assets/目录：**
 - **描述：**存放Skill所需的静态资源和模板文件，例如配置模板、音视频文件等。
 - **功能：**这些资源文件为智能体提供执行任务所需的额外信息和素材，确保任务执行的准确性和高效性。

通过清晰的文件夹结构和按需加载机制，Skill能够帮助智能体在对话场景中灵活扩展功能，从通用助手快速转变为特定领域的专家。

Skill 的类型

Managed Agents的Skill来自AgentArts “我的资产 > Skill” 中的技能。

AgentArts支持用户手动上传的技能，这些技能仅限您个人使用。技能的范畴广泛，不仅限于编程，还包括语言、沟通、管理、技术、创意、商业、个人发展、专业、生活和社交技能等。导入Skill的具体操作请参考[导入Skill](#)。

⚠ 注意

请您注意，“第三方技能”来源于第三方服务商提供，当您使用第三方服务商提供的技能时，您应严格遵守第三方的相关协议。您理解并同意，第三方服务商提供的技能可能存在风险，AgentArts无法对其做出任何形式的保证。请您在使用前审慎考量并评估其风险。

AgentArts支持导入通过第三方工具生成的资源包，或存储在OBS服务中的资源包。

约束与限制

使用技能需满足以下要求：接入的环境为私网访问且挂载了OBS。

在智能体中使用技能

在自定义创建智能体时，选择技能。

Skill (可选)

选择 Skill

已选 1/50

skill-creator_1783586807640 ×

后续步骤

[将工作交给智能体](#)

5 将工作交给智能体

5.1 调用智能体概述

调用智能体是Managed Agents的核心使用环节。通过WebSocket协议向已创建的智能体发送消息，智能体在沙箱环境中自主决策、调用工具、执行任务，并将流式结果实时返回给调用方。整个过程由平台全托管，开发者无需自行搭建智能体循环、沙箱编排或工具执行层。

创建智能体时，用户需为其配置入站身份认证方式，作为后续调用该智能体的身份验证机制。智能体支持IAM认证、OAuth 2.0认证和API Key认证三种方式。调用方需按照智能体所配置的认证方式完成身份验证，方可成功调用智能体。

这里介绍IAM认证、API Key认证的调用，具体请参考[调用智能体（IAM认证）](#)、[调用智能体（API Key认证）](#)。

为什么调用智能体

Managed Agents的智能体在创建完成后，处于已部署但未交互状态。调用智能体功能解决了以下问题：

- **连通性验证**：首次创建智能体后，需验证智能体与本地环境的网络连通性和认证链路是否正常。
- **任务执行入口**：智能体的所有能力（模型推理、工具调用、Skill执行、文件操作）均需通过调用功能触发，没有其他途径使智能体开始工作。
- **实时交互通道**：通过WebSocket实现全双工通信，支持流式事件推送，调用方可实时观察智能体的决策与执行状态。
- **多会话隔离**：每次调用通过会话ID隔离上下文，支持同一智能体被多用户并发调用且互不干扰。

应用场景

场景	说明
连通性调测	智能体创建后，通过发送简单消息（如“你好”）验证端到端链路（认证→WebSocket连接→沙箱启动→模型响应→流式返回）是否畅通。

场景	说明
单轮任务执行	发送一条指令，智能体自主决策调用内置工具（如write_file创建文件、bash执行命令），完成文件操作或命令执行任务。
多轮有状态交互	在同一会话ID下持续发送消息，智能体维护上下文信息与文件系统状态，实现跨轮次的连续对话。
工具能力验证	启用或禁用特定工具后，通过调用验证智能体是否正确使用工具（如write_file写入成功）或正确拒绝（工具不可用时给出提示）。
干预与中断	在任务执行过程中，发送新消息变更原有执行方向，或直接终止、暂停当前任务。
长时运行任务	任务持续运行数分钟至数小时，包含多次工具调用，调用方通过流式事件实时监控执行进度。

5.2 调用智能体（IAM 认证）

本章节以IAM认证为例，面向首次使用者，从零开始逐步配置并运行managed_agents_client.py脚本，该脚本用于通过WebSocket与Managed Agents后端交互，调测智能体与用户本地连通性。

前提条件

智能体的“入站身份认证”选择“IAM 认证”。

环境准备

1. Python版本要求。
需要**Python 3.10或以上版本**（代码使用了str|None类型标注语法）。
检查方法，打开电脑系统自带的命令提示符窗口（应用程序中搜索cmd），输入命令：

```
python --version
```


预期输出类似：

```
Python 3.14.0
```


如果版本低于3.10，请前往python.org升级。
2. pip安装。
Python 3.4及以上版本自带pip。
安装Python时，勾选“Add Python.exe to PATH”（这里以3.12版本为例，不同版本名称有所不同，以实际为准），默认自带pip。
打开cmd/PowerShell验证：

```
pip --version
```


输出版本号说明已经安装完成。

```
pip 25.0.1 from ...
```


Windows如果提示pip不是内部或外部命令，执行以下操作。
单击安装包，选择“Modify”，勾选“pip”、“Add Python to environment variables”。

- 安装后，重新打开命令提示符窗口，才会生效。
3. 安装依赖。
- pip就绪后，安装websockets库：
- ```
pip install websockets>=10.0 --break-system-packages
```
4. 验证安装。
- ```
python -c "import websockets; print(websockets.__version__)"
```
- 预期输出类似：
- ```
17.1
```

准备签名 SDK

1. 在[签名指南](#)文档中下载Python SDK，并在本地解压。签名指南文档中的其他内容本案例中可不用关注。

图 5-1 获取 SDK



解压后的文件目录结构如下：

|                        |                      |
|------------------------|----------------------|
| ApiGateway-python-sdk/ | ← 根目录（解压出来的文件夹）      |
| ├── apig_sdk/          | ← 【核心签名库，不要动！】       |
| │   ├── __init__.py    | ← 包的入口标识文件           |
| │   └── signer.py      | ← 签名核心代码（所有签名逻辑都在这里） |
| ├── main.py            | ← 【您要改的文件】调用示例       |
| ├── licenses/          | ← 开源许可证文件夹（不用管）      |
| └── license-requests   | ← requests库的许可证说明    |

图 5-2 目录结构

| ApiGateway-python-sdk-2.0.7 |                |            |      |
|-----------------------------|----------------|------------|------|
| 排序 查看                       |                |            |      |
| 名称                          | 修改日期           | 类型         | 大小   |
| apig_sdk                    | 2026/9/7 20:12 | 文件夹        |      |
| licenses                    | 2026/9/7 20:12 | 文件夹        |      |
| backend_signature.py        | 2026/9/7 20:12 | Python 源文件 | 4 KB |
| demo_v11.py                 | 2026/9/7 20:12 | Python 源文件 | 2 KB |
| main.py                     | 2026/9/7 20:12 | Python 源文件 | 2 KB |
| signer_test.py              | 2026/9/7 20:12 | Python 源文件 | 9 KB |
| signer_v11_test.py          | 2026/9/7 20:12 | Python 源文件 | 5 KB |
| test_signer.py              | 2026/9/7 20:12 | Python 源文件 | 1 KB |

2. 下载[managed\\_agents\\_client.py](#)至本地，放入Python SDK解压的文件夹中，该脚本用于通过WebSocket与Managed Agents后端交互，调测智能体与用户本地连通性。
- 在Python SDK的文件夹上方显示路径的地址栏里，清空地址，直接输入cmd后按回车。

调用智能体

Windows cmd：

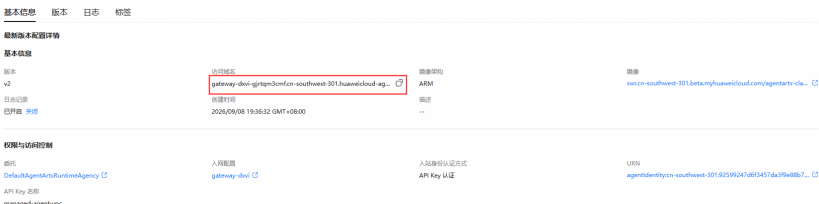
```
python managed_agents_client.py ^
--url "ws://{endpoint}/runtimes/{runtime_name}/ws" ^
--auth-type iam ^
--ak "{AK}" --sk "{SK}" ^
--session-id "{会话ID}" ^
--message "你好"
```

通用写法（所有平台通用）：

```
python managed_agents_client.py --url "ws://{endpoint}/runtimes/{runtime_name}/ws" --auth-type iam --ak "{AK}" --sk "{SK}" --session-id "{会话ID}" --message "你好"
```

调用参数信息，请按[表5-1](#)获取，替换命令中花括号内的占位符。

表 5-1 获取调用信息

| 参数                                    | 操作                                                                                                                                                                                                                                                                       |
|---------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <code>managed_agents_client.py</code> | 脚本文件名由脚本名称和文件类型后缀组成。例如，脚本名称为managed_agents_client，文件类型为Python脚本（.py），则完整文件名为managed_agents_client.py。                                                                                                                                                                    |
| 运行时名称（ <code>{runtime_name}</code> ）  | <div><div>1. 在“Managed Agents &gt; 智能体”页面，复制需要调用的智能体名称。</div><div>2. 在左侧导航选择“托管与运行 &gt; 智能体运行时”，搜索已获取的智能体名称。</div><div>在“智能体运行时”列表，显示的名称以managed-开头，即为运行时名称。</div></div> <div></div> |
| 访问域名（ <code>{endpoint}</code> ）       | <div>单击运行时名称，进入运行时详情页面，在“访问域名”区域，即可复制访问域名。</div> <div></div>                                                                                                                         |

| 参数                                        | 操作                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |                 |                             |      |        |         |    |                 |                             |        |    |                 |                             |  |   |   |   |   |   |   |   |           |               |                   |  |  |  |   |  |        |                |  |  |  |   |  |    |    |  |  |  |   |  |  |  |  |  |  |   |  |  |  |  |  |  |   |  |  |  |  |  |  |   |  |  |  |  |  |  |
|-------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------|-----------------------------|------|--------|---------|----|-----------------|-----------------------------|--------|----|-----------------|-----------------------------|--|---|---|---|---|---|---|---|-----------|---------------|-------------------|--|--|--|---|--|--------|----------------|--|--|--|---|--|----|----|--|--|--|---|--|--|--|--|--|--|---|--|--|--|--|--|--|---|--|--|--|--|--|--|---|--|--|--|--|--|--|
| <div><div>{AK}</div><div>{SK}</div></div> | <div><div>登录控制台“<a href="#">我的凭证 &gt; 访问密钥</a>”，获取AK和SK。AK、SK最多可添加2个。</div><div><div>图 5-3 获取 AK、SK</div><div><div><div>我的凭证</div><div>API凭证</div><div>访问密钥</div></div><div><div>访问密钥 ①</div><div><div>如果访问密钥泄露，会带来数据泄露风险，且每个访问密钥仅能下载一次。为了账号安全性，建议您定期更换并妥善保存访问密钥。若您的访问密钥已丢失，您可</div><div>新增访问密钥</div><div>您还可以添加0个访问密钥。</div><table><tr><th>访问密钥ID ①</th><th>描述 ①</th><th>状态 ①</th><th>创建时间 ①</th></tr><tr><td>IUVVLIS</td><td>--</td><td><span>应用</span></td><td>2022/03/11 16:55:34 GMT+...</td></tr><tr><td>HPUAKA</td><td>--</td><td><span>应用</span></td><td>2026/04/15 09:41:13 GMT+...</td></tr></table></div></div></div><div><div>图 5-4 AK、SK 示例</div><table><tr><th></th><th>A</th><th>B</th><th>C</th><th>D</th><th>E</th><th>F</th></tr><tr><td>1</td><td>User Name</td><td>Access Key Id</td><td>Secret Access Key</td><td></td><td></td><td></td></tr><tr><td>2</td><td></td><td>HPUAKA</td><td>6irFsl1UjTok0K</td><td></td><td></td><td></td></tr><tr><td>3</td><td></td><td>AK</td><td>SK</td><td></td><td></td><td></td></tr><tr><td>4</td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>5</td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>6</td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>7</td><td></td><td></td><td></td><td></td><td></td><td></td></tr></table></div></div></div> | 访问密钥ID ①        | 描述 ①                        | 状态 ① | 创建时间 ① | IUVVLIS | -- | <span>应用</span> | 2022/03/11 16:55:34 GMT+... | HPUAKA | -- | <span>应用</span> | 2026/04/15 09:41:13 GMT+... |  | A | B | C | D | E | F | 1 | User Name | Access Key Id | Secret Access Key |  |  |  | 2 |  | HPUAKA | 6irFsl1UjTok0K |  |  |  | 3 |  | AK | SK |  |  |  | 4 |  |  |  |  |  |  | 5 |  |  |  |  |  |  | 6 |  |  |  |  |  |  | 7 |  |  |  |  |  |  |
| 访问密钥ID ①                                  | 描述 ①                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        | 状态 ①            | 创建时间 ①                      |      |        |         |    |                 |                             |        |    |                 |                             |  |   |   |   |   |   |   |   |           |               |                   |  |  |  |   |  |        |                |  |  |  |   |  |    |    |  |  |  |   |  |  |  |  |  |  |   |  |  |  |  |  |  |   |  |  |  |  |  |  |   |  |  |  |  |  |  |
| IUVVLIS                                   | --                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | <span>应用</span> | 2022/03/11 16:55:34 GMT+... |      |        |         |    |                 |                             |        |    |                 |                             |  |   |   |   |   |   |   |   |           |               |                   |  |  |  |   |  |        |                |  |  |  |   |  |    |    |  |  |  |   |  |  |  |  |  |  |   |  |  |  |  |  |  |   |  |  |  |  |  |  |   |  |  |  |  |  |  |
| HPUAKA                                    | --                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | <span>应用</span> | 2026/04/15 09:41:13 GMT+... |      |        |         |    |                 |                             |        |    |                 |                             |  |   |   |   |   |   |   |   |           |               |                   |  |  |  |   |  |        |                |  |  |  |   |  |    |    |  |  |  |   |  |  |  |  |  |  |   |  |  |  |  |  |  |   |  |  |  |  |  |  |   |  |  |  |  |  |  |
|                                           | A                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | B               | C                           | D    | E      | F       |    |                 |                             |        |    |                 |                             |  |   |   |   |   |   |   |   |           |               |                   |  |  |  |   |  |        |                |  |  |  |   |  |    |    |  |  |  |   |  |  |  |  |  |  |   |  |  |  |  |  |  |   |  |  |  |  |  |  |   |  |  |  |  |  |  |
| 1                                         | User Name                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   | Access Key Id   | Secret Access Key           |      |        |         |    |                 |                             |        |    |                 |                             |  |   |   |   |   |   |   |   |           |               |                   |  |  |  |   |  |        |                |  |  |  |   |  |    |    |  |  |  |   |  |  |  |  |  |  |   |  |  |  |  |  |  |   |  |  |  |  |  |  |   |  |  |  |  |  |  |
| 2                                         |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             | HPUAKA          | 6irFsl1UjTok0K              |      |        |         |    |                 |                             |        |    |                 |                             |  |   |   |   |   |   |   |   |           |               |                   |  |  |  |   |  |        |                |  |  |  |   |  |    |    |  |  |  |   |  |  |  |  |  |  |   |  |  |  |  |  |  |   |  |  |  |  |  |  |   |  |  |  |  |  |  |
| 3                                         |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             | AK              | SK                          |      |        |         |    |                 |                             |        |    |                 |                             |  |   |   |   |   |   |   |   |           |               |                   |  |  |  |   |  |        |                |  |  |  |   |  |    |    |  |  |  |   |  |  |  |  |  |  |   |  |  |  |  |  |  |   |  |  |  |  |  |  |   |  |  |  |  |  |  |
| 4                                         |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |                 |                             |      |        |         |    |                 |                             |        |    |                 |                             |  |   |   |   |   |   |   |   |           |               |                   |  |  |  |   |  |        |                |  |  |  |   |  |    |    |  |  |  |   |  |  |  |  |  |  |   |  |  |  |  |  |  |   |  |  |  |  |  |  |   |  |  |  |  |  |  |
| 5                                         |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |                 |                             |      |        |         |    |                 |                             |        |    |                 |                             |  |   |   |   |   |   |   |   |           |               |                   |  |  |  |   |  |        |                |  |  |  |   |  |    |    |  |  |  |   |  |  |  |  |  |  |   |  |  |  |  |  |  |   |  |  |  |  |  |  |   |  |  |  |  |  |  |
| 6                                         |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |                 |                             |      |        |         |    |                 |                             |        |    |                 |                             |  |   |   |   |   |   |   |   |           |               |                   |  |  |  |   |  |        |                |  |  |  |   |  |    |    |  |  |  |   |  |  |  |  |  |  |   |  |  |  |  |  |  |   |  |  |  |  |  |  |   |  |  |  |  |  |  |
| 7                                         |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |                 |                             |      |        |         |    |                 |                             |        |    |                 |                             |  |   |   |   |   |   |   |   |           |               |                   |  |  |  |   |  |        |                |  |  |  |   |  |    |    |  |  |  |   |  |  |  |  |  |  |   |  |  |  |  |  |  |   |  |  |  |  |  |  |   |  |  |  |  |  |  |
| 会话ID（{会话ID}）                              | 每个会话的唯一标识符。用户可将会话ID设置为任意字符串，由英文、数字、“-”、“_”组成，不超过64个字符。例如“123e4567e89b12d3a456426614174000”。                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |                 |                             |      |        |         |    |                 |                             |        |    |                 |                             |  |   |   |   |   |   |   |   |           |               |                   |  |  |  |   |  |        |                |  |  |  |   |  |    |    |  |  |  |   |  |  |  |  |  |  |   |  |  |  |  |  |  |   |  |  |  |  |  |  |   |  |  |  |  |  |  |
| message                                   | 会话内容。按实际进行替换。                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |                 |                             |      |        |         |    |                 |                             |        |    |                 |                             |  |   |   |   |   |   |   |   |           |               |                   |  |  |  |   |  |        |                |  |  |  |   |  |    |    |  |  |  |   |  |  |  |  |  |  |   |  |  |  |  |  |  |   |  |  |  |  |  |  |   |  |  |  |  |  |  |

显示结果

预期终端输出：

```
[2/5] 已连接 ws://{endpoint}/runtimes/{runtime_name}/ws
[3/5] 服务端就绪
[4/5] 已发送: 你好

[Frame] {"response_kind":"e2a.chunk","status":"in_progress","body":{"delta_kind":"text","delta":"你好
"},"event_type":"chat.delta"}}
你好
[Frame] {"response_kind":"e2a.chunk","status":"in_progress","body":{"delta_kind":"text","delta":"! 有什么可
以帮你的吗? ","event_type":"chat.delta"}}
! 有什么可以帮你的吗?
[Frame] {"response_kind":"e2a.complete","status":"succeeded","is_final":true,"body":{"result":{"content":"你
好! 有什么可以帮你的吗? }}}

=====
回复: 你好! 有什么可以帮你的吗?
=====

[5/5] 已断开
```

权限审批和用户提问均在后台自动处理，终端只显示“Frame”原始帧和流式回复文本。

## 5.3 调用智能体 (API Key 认证)

本章节以API Key认证为例，面向首次使用者，从零开始逐步配置并运行managed\_agents\_client.py脚本，该脚本用于通过WebSocket与Managed Agents后端交互，调测智能体与用户本地连通性。

### 前提条件

智能体的“入站身份认证”选择“API Key 认证”。

### 环境准备

1. Python版本要求。

需要**Python 3.10或以上版本**（代码使用了str|None类型标注语法）。

检查方法，打开电脑系统自带的命令提示符窗口（应用程序中搜索cmd），输入命令：

```
python --version
```

预期输出类似：

```
Python 3.14.0
```

如果版本低于3.10，请前往[python.org](https://python.org)升级。

2. pip安装。

Python 3.4及以上版本自带pip。

安装Python时，勾选“Add Python.exe to PATH”（这里以3.12版本为例，不同版本名称有所不同，以实际为准），默认自带pip。

打开cmd/PowerShell验证：

```
pip --version
```

输出版本号说明已经安装完成。

```
pip 25.0.1 from ...
```

Windows如果提示pip不是内部或外部命令，执行以下操作。

单击安装包，选择“Modify”，勾选“pip”、“Add Python to environment variables”。

安装后，重新打开命令提示符窗口，才会生效。

3. 安装依赖。

pip就绪后，安装websockets库：

```
pip install websockets>=10.0 --break-system-packages
```

4. 验证安装。

```
python -c "import websockets; print(websockets.__version__)"
```

预期输出类似：

```
17.1
```

5. 确认脚本位置。

下载[managed\\_agents\\_client.py](#)至本地，该脚本用于通过WebSocket与Managed Agents后端交互，调测智能体与用户本地连通性。

在managed\_agents\_client.py脚本所在目录，使用Windows命令提示符（CMD）工具，执行命令。

调用智能体

Windows cmd：

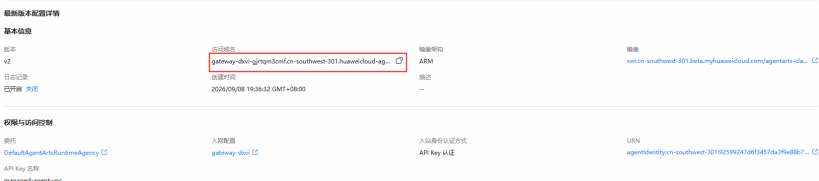
```
python managed_agents_client.py ^
--url "ws://{endpoint}/runtimes/{runtime_name}/ws" ^
--auth "Bearer {api_key}" ^
--session-id "{会话ID}" ^
--message "你好"
```

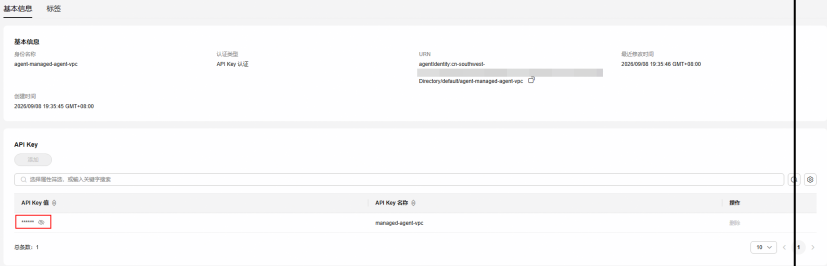
通用写法（所有平台通用）：

```
python managed_agents_client.py --url "ws://{endpoint}/runtimes/{runtime_name}/ws" --auth "Bearer {api_key}" --session-id "{会话ID}" --message "你好"
```

调用参数信息，请按表5-2获取，替换命令中花括号内的占位符。

表 5-2 获取调用信息

| 参数                                    | 操作                                                                                                                                                                                                                                                                                 |
|---------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <code>managed_agents_client.py</code> | 脚本文件名由脚本名称和文件类型后缀组成。例如，脚本名称为 <code>managed_agents_client</code> ，文件类型为Python脚本（.py），则完整文件名为 <code>managed_agents_client.py</code> 。                                                                                                                                                |
| 运行时名称（ <code>{runtime_name}</code> ）  | <div><div><div>1. 在“Managed Agents &gt; 智能体”页面，复制需要调用的智能体名称。</div><div>2. 在左侧导航选择“托管与运行 &gt; 智能体运行时”，搜索已获取的智能体名称。</div></div><div>在“智能体运行时”列表，显示的名称以managed-开头，即为运行时名称。</div><div></div></div> |
| 访问域名（ <code>{endpoint}</code> ）       | <div>单击运行时名称，进入运行时详情页面，在“访问域名”区域，即可复制访问域名。</div> <div></div>                                                                                                                                   |

| 参数                              | 操作                                                                                                                                                                                                                                                                                               |
|---------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| API Key<br>( <i>{api_key}</i> ) | <div><div><div>1. 在运行时详情页面，单击URN。</div><div></div></div><div><div>2. 进入工作负载身份详情页面，即可复制API Key值。</div><div></div></div></div> |
| 会话ID ( <i>{会话ID}</i> )          | 每个会话的唯一标识符。用户可将会话ID设置为任意字符串，由英文、数字、“-”、“_”组成，不超过64个字符。例如“123e4567e89b12d3a456426614174000”。                                                                                                                                                                                                      |
| message                         | 会话内容。按实际进行替换。                                                                                                                                                                                                                                                                                    |

显示结果

预期终端输出：

```
[2/5] 已连接 ws://{endpoint}/runtimes/{runtime_name}/ws
[3/5] 服务端就绪
[4/5] 已发送: 你好

[Frame] {"response_kind":"e2a.chunk","status":"in_progress","body":{"delta_kind":"text","delta":"你好"},"event_type":"chat.delta"}}
你好
[Frame] {"response_kind":"e2a.chunk","status":"in_progress","body":{"delta_kind":"text","delta":"! 有什么可以帮你的吗?"},"event_type":"chat.delta"}}
! 有什么可以帮你的吗?
[Frame] {"response_kind":"e2a.complete","status":"succeeded","is_final":true,"body":{"result":{"content":"你好! 有什么可以帮你的吗?"}}}

=====
回复: 你好! 有什么可以帮你的吗?
=====
[5/5] 已断开
```

权限审批和用户提问均在后台自动处理，终端只显示“Frame”原始帧和流式回复文本。

5.4 检查结果状态

您可以在WebSocket事件流中检查每帧的“status”和“is\_final”字段来判断结果状态。在完成之前，“status”报告“in\_progress”。

表 5-3 状态含义

| status值     | 含义       | is_final |
|-------------|----------|----------|
| in_progress | 处理中（中间帧） | False    |
| succeeded   | 成功（最终帧）  | True     |
| failed      | 失败（最终帧）  | True     |

预期终端输出：

```
[2/5] 已连接 ws://{endpoint}/runtimes/{runtime_name}/ws
[3/5] 服务端就绪
[4/5] 已发送: 你好

[Frame] {"response_kind":"e2a.chunk","status":"in_progress","body":{"delta_kind":"text","delta":"你好","event_type":"chat.delta"}}
你好
[Frame] {"response_kind":"e2a.chunk","status":"in_progress","body":{"delta_kind":"text","delta":"! 有什么可以帮你的吗?","event_type":"chat.delta"}}
! 有什么可以帮你的吗?
[Frame] {"response_kind":"e2a.complete","status":"succeeded","is_final":true,"body":{"result":{"content":"你好! 有什么可以帮你的吗?"}}}

=====
回复: 你好! 有什么可以帮你的吗?
=====
[5/5] 已断开
```

5.5 查看智能体运行时数据

- 在“智能体运行时”页面，通过搜索智能体ID查看智能体运行时数据。
  - 登录AgentArts智能体平台。
  - 在左侧导航选择“托管与运行 > 智能体运行时”，在搜索框中输入上方已获取的ID进行搜索。
  - 单击搜索到的智能体名称，进入详情页面查看智能体运行时数据。
- 在“观测与优化 > 观测 > 智能体运行分析”页面，通过搜索智能体ID查看智能体运行时数据。
  - 登录AgentArts智能体平台。
  - 在左侧导航选择“观测与优化 > 观测 > 智能体运行分析”，在搜索框中输入上方已获取的ID进行搜索。
  - 单击搜索到的智能体名称，进入详情页面查看智能体运行时数据。

# 6 常见问题

## 6.1 环境被智能体关联后能否修改

不能。环境一旦被智能体关联，就禁止修改。该限制是为了保证智能体运行环境的稳定性。如果允许在运行中随意修改底层配置，可能导致正在执行的任务异常。

环境被关联后如何调整配置？

只可以创建新环境，再切换智能体关联的环境。切换环境后，旧环境不会自动删除，仍然保留在环境列表中，可以被其他智能体复用或手动删除。

## 6.2 会话空闲超时和最大存活时间有何区别

会话空闲超时、最大存活时间均可控制会话的生命周期，但触发机制不同：

表 6-1 对比说明

| 参数     | 触发条件                         | 取值范围        | 默认值   |
|--------|------------------------------|-------------|-------|
| 会话空闲超时 | 一段时间内无用户交互，自动结束会话并释放资源。      | 1~10,080 分钟 | 15 分钟 |
| 最大存活时间 | 会话从创建起达到绝对时间上限，无论是否有交互，强制结束。 | 1~168 小时    | 24 小时 |

假设配置了空闲超时=30分钟，最大存活时间=2小时。

- 场景 A：用户发了一条消息后就停止对话 → 30分钟无交互 → 空闲超时触发，会话结束。
- 场景 B：用户一直在持续对话，从未间断超过30分钟 → 但2小时一到 → 最大存活时间触发，会话强制结束，即使此时正在执行任务。

- 场景 C：用户断断续续地对话（每次间隔不超过30分钟），在第1小时50分时停止操作 → 空闲30分钟后 → 空闲超时触发（此时总时长2小时20分未到最大存活时间，但空闲超时先触发）。

## 6.3 会话存储、SFS Turbo、OBS 三种存储有什么区别

Managed Agents的环境支持配置三种类型的存储方式：

表 6-2 存储方式说明

| 存储类型      | 用途                  | 存储资源                              | 网络要求          |
|-----------|---------------------|-----------------------------------|---------------|
| 会话存储      | 保存会话上下文、中间结果、用户偏好等。 | 平台统一托管，每个会话拥有独立存储空间。              | 无限制（公网/私网均可）。 |
| SFS Turbo | 保存会话过程中产生的文件等。      | 用户的 SFS Turbo 文件系统（需与运行时在同一 VPC）。 | 仅私网访问。        |
| OBS       | 保存会话过程中产生的文件等。      | 用户的OBS桶。                          | 仅私网访问。        |

如何选择：

- 需要在多轮对话中保持上下文连续（如记住用户偏好、保留中间计算结果），选择会话存储。会话存储由平台托管，使用最简单，且不要求私网环境。
- 需要在会话中读写文件，且多个智能体需要共享同一批文件，选择SFS Turbo。SFS Turbo是高性能共享文件系统，支持多实例同时挂载读写，适合需要文件共享的场景。
- 需要持久化保存智能体生成的产物文件（如生成的报告、代码文件、数据处理结果），选择OBS。智能体交互过程中生成的产物会自动存储到OBS桶中，可在OBS控制台查看和下载。

## 6.4 存储挂载路径有哪些约束

表 6-3 路径规则

| 约束项    | 具体要求                                               |
|--------|----------------------------------------------------|
| 允许的父目录 | 挂载路径必须为/home、/mnt、/data、/tmp这几个目录及其子目录。            |
| 路径长度   | 最大 4000 字符。                                        |
| 禁止路径穿越 | 不能包含 ..。                                           |
| 禁止路径冲突 | 所有挂载路径（包括 OBS、SFS Turbo、会话存储之间）不能重复，也不能存在父子目录包含关系。 |

**注意：**多个挂载（包括OBS、SFS Turbo、会话存储之间）禁止使用相同或存在包含关系的路径。

**路径冲突举例：**如果会话存储挂载到 /home/session，那么 OBS 就不能挂载到 /home/session/output（子目录，存在包含关系），也不能挂载到 /home/session（重复路径），但可以挂载到 /home/obs-data（同级不同目录，没有冲突）。

6.5 会话超时后数据还在吗

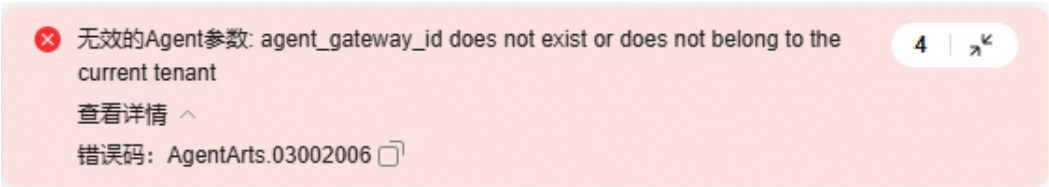
表 6-4 数据保留说明

| 数据类型              | 超时是否保留 | 说明                                                   |
|-------------------|--------|------------------------------------------------------|
| 对话上下文（历史消息、中间结果等） | 清除     | 会话终止后，沙箱资源释放，会话级上下文随之清除。再次调用时需重新发起会话。                |
| 沙箱内的临时文件          | 清除     | 未持久化到外部存储的文件会随沙箱释放而丢失。                               |
| 已写入OBS的产物文件       | 保留     | 如果配置了OBS存储，智能体生成的产物文件已写入OBS桶，不受会话超时影响，可在OBS控制台查看和下载。 |
| 已写入SFS Turbo的文件   | 保留     | 同理，SFS Turbo是独立于会话的持久化存储。                            |
| 运行日志              | 保留     | 如果开启了日志记录，日志已上报到云日志服务（LTS），不受会话超时影响。                 |
| 智能体配置             | 保留     | 名称、模型、提示词、工具、技能、环境等配置是持久保存的，与会话生命周期无关。               |

6.6 创建智能体时，报错 AgentArts.03002006，如何处理？

创建智能体时，如果报错AgentArts.03002006，请按如下操作处理：

图 6-1 报错信息



- 1. 在AgentArts智能体平台左侧导航，选择“治理与安全 > 网络配置”。
- 2. 在defaultgwi详情页面，单击“编辑”，再单击“确定”即可，如图6-2和图6-3所示。

图 6-2 编辑

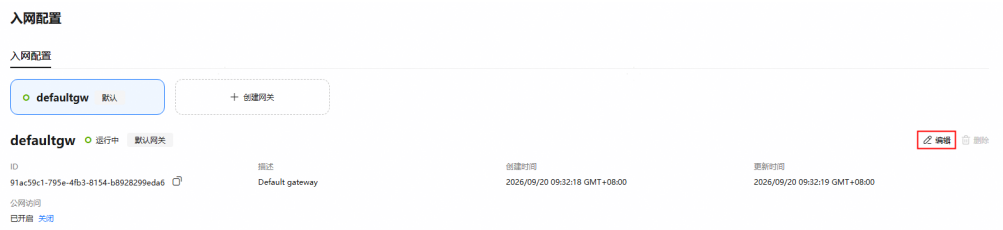
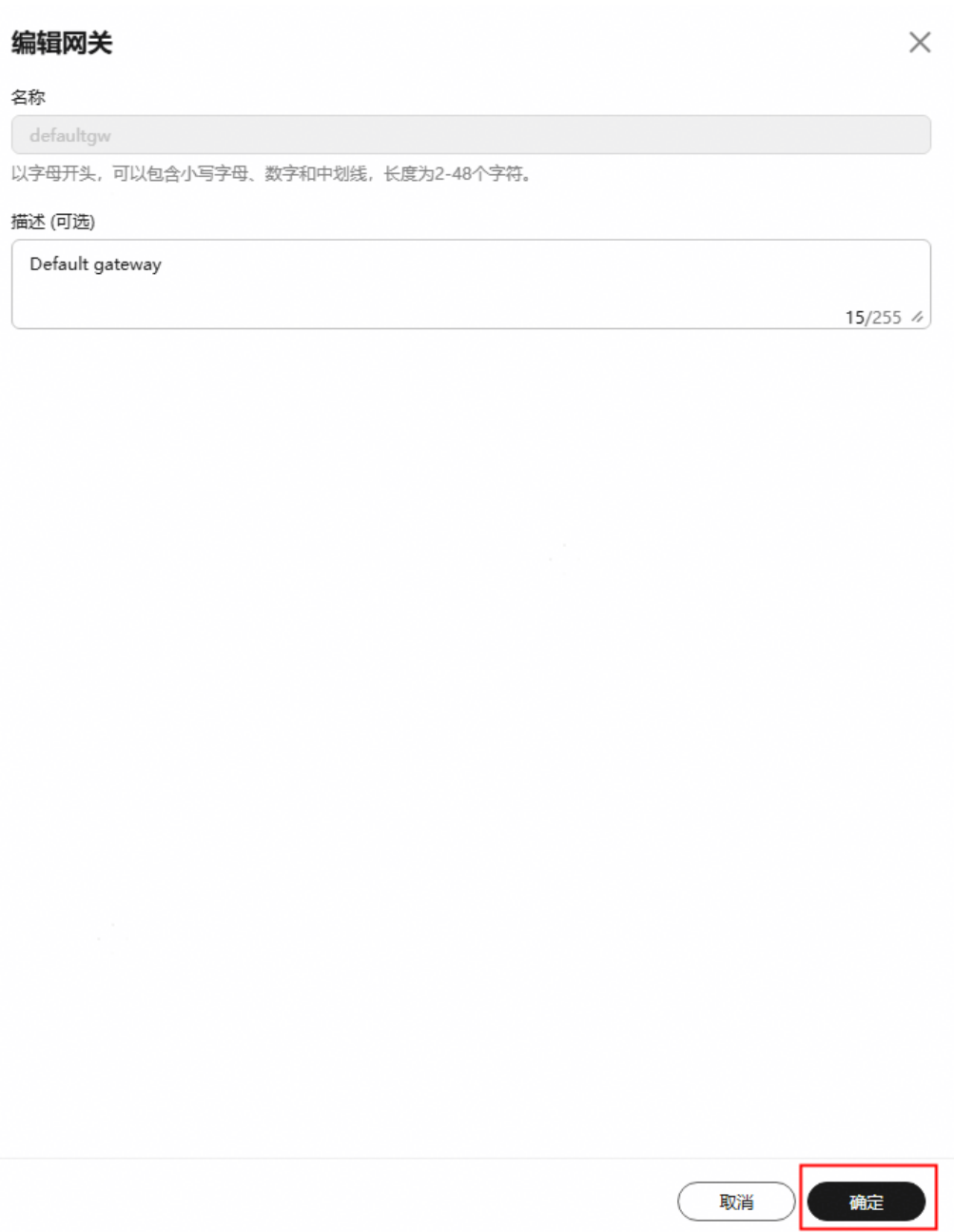


图 6-3 确定



3. 再在“Managed Agents”页面，执行创建智能体操作。