

MaaS 模型即服务

接入 AI 工具

文档版本 01
发布日期 2026-10-12



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 华为云码道（CodeArts）代码智能体.....	1
2 果办 OfficeAce.....	5
3 OpenClaw.....	9
4 OpenCode.....	16
5 Claude Code.....	20
6 Dify.....	24
7 Cherry Studio.....	33
8 Cursor.....	38
9 Cline.....	42
10 RAGFlow.....	48
11 Deep Research.....	54
12 TRAE.....	60

1 华为云码道（CodeArts）代码智能体

华为云码道代码智能体支持企业接入第三方大语言模型，可根据业务场景灵活定制，并满足企业私有化部署、数据安全与合规管控要求，实现模型能力与业务深度适配。

华为云码道代码智能体专业版套餐支持通过控制台和客户端配置模型，基础版套餐和体验版套餐仅支持在客户端配置模型。

- 控制台：仅华为云码道企业管理员可配置模型，并可管控华为云码道企业成员在客户端配置模型的权限。
- 客户端：所有华为云码道企业成员均可配置模型。

前提条件

- 已注册华为云账号，并进行实名认证。具体操作，请参见[注册华为账号并开通华为云和实名认证](#)。
- 已开通MaaS预置服务，具体操作请参见[开通预置服务](#)。
- 已购买华为云码道代码智能体**企业版套餐**、**个人版套餐（新版套餐）**或**专业版套餐**、**基础版套餐**、**体验版套餐（旧版套餐）**。
 - **企业版套餐或专业版套餐**支持通过控制台和客户端配置模型。
 - **个人版套餐或基础版套餐、体验版套餐**仅支持在客户端配置模型。
- 已获取华为云码道企业成员或企业管理员权限。
- 已添加华为云码道团队成员，具体操作请参见[管理成员及团队](#)。
- 安装华为云码道代码智能体客户端，详细操作请参见[华为云码道代码智能体 快速入门](#)。

计费说明

在调用MaaS模型服务时，将根据实际使用的Tokens数量进行计费。计费详情请参见[MaaS模型服务计费项](#)。

约束限制

- 华为云码道接入第三方模型仅支持OpenAI规范的Chat Completions API格式和Anthropic规范的Messages API格式。
- 华为云码道自定义的模型仅可由系统内置智能体、AgentTeam和自定义的智能体调用，智能问答无法调用。

控制台操作步骤

1. 使用已授权**企业管理员**角色的账号登录**码道代码智能体**控制台。
2. 在左侧导航栏选择“企业设置 > 模型配置”，进入“模型配置”页面。
3. 在“企业内置模型”区域，单击“新建模型”，弹出“新建模型”对话框。
4. 参考**表1-1**，接入MaaS模型。

图 1-1 接入 MaaS 模型

新建模型

模型提供商

自定义配置

模型提供商

华为云MaaS

模型ID

自定义

自定义模型ID

openpangu-2.0-flash

查询模型ID

模型展示名称

MaaS-openpangu-2.0-flash

24/64

API Key

获取API Key

可用范围

☒ 企业 ☐ 团队

高级配置（可选）

自定义模型URL

自定义编辑

https://api.modelarts-maas.com/v2

上下文窗口

输入 184000 输出 16000

描述

openPangu-2.0-Pro 是基于昇腾 NPU 训练的大规模混合专家（MoE）语言模型

表 1-1 接入 MaaS 模型参数说明

参数	说明
模型提供商	选择“华为云MaaS”。
模型ID	在下拉列表中选择模型名称，例如：“Kimi-K2.6”。 如果下拉列表中没有需要的模型名称，请选择“自定义”，并在“自定义模型ID”中填写模型名称，例如：“openpangu-2.0-flash”当前支持模型参见模型列表。
模型展示名称	模型展示名称可与系统内置模型名称相同，但不可与其他已创建的自定义模型名称相同。 设置模型在华为云码道代码智能体中显示的名称，长度不超过64个字符。 输入“MaaS-openPangu-2.0-Flash”。

参数	说明
可用范围	设置模型的可用范围，创建后不可修改。 - 企业：企业全员可用。 - 团队：仅指定团队可用。 企业管理员可使用所有企业级和团队级配置的模型。
API Key	获取API Key 。
自定义模型URL	选择模型供应商后，保持默认。 如果接入Token Plan需要开启“自定义编辑”，并填写兼容OpenAI接口协议工具： https://api.modelarts-maas.com/plan/v2
上下文窗口	可选，保持默认。
描述	可选，保持默认。

- 5. 单击“确定”，完成MaaS模型的接入。
在“企业内置模型”区域，查看新接入的MaaS模型。
- 6. 打开华为云码道代码智能体客户端的聊天界面，在输入框下方选择“内置 > 智能体”，在右侧模型列表中选择MaaS模型。在输入框中输入“什么是二叉树”，若模型返回正确内容，说明MaaS模型可正常使用。

图 1-2 使用 MaaS 模型



2 果办 OfficeAce

果办OfficeAce是办公场景专属的桌面AI助手，致力于以AI重构桌面办公体验。不同于传统的对话AI，果办OfficeAce是能直接操控软件、处理文件、执行流程的数字办公专家。它通过自动化处理重复事务、智能整合信息及协作辅助，将用户从繁琐操作中解放出来，大幅提升办公效率，让智能办公触手可及。

前提条件

- 接入前请确保已开通预置模型服务，如未开通，请前往[预置服务](#)开通。
- 已安装并登录果办OfficeAce。具体操作请参见[安装果办OfficeAce](#)。

操作步骤

1. [登录果办OfficeAce](#)客户端，在左侧导航栏，选择“模型”，单击右上角的“接入模型”，选择“接入华为云MaaS模型”。
2. 选择“接入类型”为“华为云MaaS 按需计费”，输入API Key。

API Key是华为云MaaS的接口访问凭证，用于接口调用的身份鉴权。如果没有API Key，请前往[API Key 管理](#)获取。

图 2-1 接入华为云 MaaS 按需计费模型

接入MaaS模型

接入类型

华为云MaaS Token Plan

华为云MaaS 按需计费

API Key

.....

创建API Key

选择模型

模型服务

☐ 全选

☐ DeepSeek-V3

☐ Wan2.2-T2V-A14B

☐ qwen-image-edit-2509

☐ vidu/viduq3-pro_start-end2video

☐ Wan2.2-I2V-A14B

☐ vidu/viduq3-pro_text2video

☐ qwen2.5-vl-72b

☐ bge-m3

没有开通?

华为云MaaS 模型推理服务

为您提供模型在线推理服务，开通后体验或调用以实际用量进行扣费

去开通

取消

连通性测试

确认

3. 在“选择模型”区域，选择需要接入的模型（支持多选），不同模型的说明请参见[模型列表](#)。
4. 选择模型后，单击“连通性测试”，测试通过表示接入成功。
以接入模型“qwen2.5-vl-72b”为例，勾选qwen2.5-vl-72b模型，单击“连通性测试”，显示连通性测试通过，如图2-2所示。

图 2-2 接入模型示例

接入MaaS模型

×

接入类型

华为云MaaS Token Plan

华为云MaaS 按需计费

API Key

.....

创建API Key

选择模型

模型服务

☐ 全选

☐ DeepSeek-V3

☐ Wan2.2-T2V-A14B

☐ qwen-image-edit-2509

☐ vidu/viduq3-pro_start-end2video

☐ Wan2.2-I2V-A14B

☐ vidu/viduq3-pro_text2video

☒ qwen2.5-vl-72b

☐ bge-m3

1个模型连通性测试通过

没有开通?

华为云MaaS 模型推理服务

为您提供模型在线推理服务，开通后体验或调用以实际用量进行扣费

去开通

取消

连通性测试

确认

说明

连通性测试的状态说明：

- ：连通性测试通过，表示该模型已开通，且能正常调用。
- ：连通性测试不通过，可能原因包括：该模型未开通或API Key配置有误。
 - 如果模型未开通，则前往[预置服务](#)开通。
 - 如果API Key配置有误，请前往[API Key 管理](#)重新创建。
- ：不支持连通性测试，表示该模型暂不具备在线检测能力，请通过实际调用确认其可用性。

5. 连通性测试通过后，单击“确认”。

在模型界面可查看已接入的按需计费模型，如下图所示。

文档版本 01 (2026-10-12)

版权所有 © 华为云计算技术有限公司

7

图 2-3 已接入的按需计费模型信息



3 OpenClaw

在软件开发过程中，开发者经常面临代码编写效率低、调试困难等问题。OpenClaw是一款开源的个人AI助手平台，通过接入大模型可帮助开发者高效完成代码开发、调测。然而，如何将这些先进的AI能力快速集成到现有的开发环境中，成为了一个新的挑战。您可通过简单配置，把MaaS模型API接入OpenClaw，从而轻松实现AI能力的集成，提高开发效率。本文档旨在帮助开发者快速配置并高效、安全地在OpenClaw环境中集成MaaS模型的API。

前提条件

- 已注册华为云账号，并进行实名认证。具体操作，请参见[注册华为账号并开通华为云和实名认证](#)。
- 已开通预置服务，或已创建接入点。具体操作，请参见[开通预置服务](#)或[创建自定义接入点](#)。

计费说明

在调用MaaS模型服务时，将根据实际使用的Tokens数量进行计费。计费详情请参见[MaaS文本生成模型](#)。

约束限制

该功能仅支持“西南-贵阳一”区域。

步骤一：安装 OpenClaw

安装OpenClaw包括Windows以及MacOS/Linux两种方式。

Windows

1. 安装[Node.js 22或更高版本](#)。
`node --version` # 运行该命令可验证Node.js版本
2. 在命令行界面中执行以下命令，安装OpenClaw。
`iwr -useb https://openclaw.ai/install.ps1 | iex`
或通过npm全局安装：
`npm install -g openclaw@latest`

MacOS/Linux

1. 安装Node.js 22或更高版本。
node --version #运行该命令可验证Node.js版本
2. 执行以下命令，安装OpenClaw。
curl -fsSL https://openclaw.ai/install.sh | bash
或通过npm全局安装：
npm install -g openclaw@latest

步骤二：完成初始配置（首次安装）

表 3-1 OpenClaw 配置

配置类型	配置项	配置示例
基本配置	I understand this is personal-by-default and shared/multi-user use requires lock-down. Continue?	选择 "Yes"
	Setup mode	选择 “QuickStart”
模型配置	Model/auth provider	选择 "Skip for now", 后续可以配置。
	Filter models by provider	选择 "All providers"
	Default model	使用默认配置
功能配置	Select channel (QuickStart)	选择 “Skip for now” ，后续可以配置。
	Search provider	选择 “Skip for now” ，后续可以配置。
	Configure skills now? (recommended)	选择 “No” ，后续可以配置。
	Enable hooks?	按空格键选中 “Skip for now” 选项，按回车键进入下一步。
	How do you want to hatch your bot?	选择 "Hatch in TUI"。

步骤三：在 OpenClaw 中接入 MaaS 模型 API

请按需选择以下操作。

Web UI 方式

1. 在终端执行以下命令，打开OpenClaw Web UI。
openclaw dashboard

2. 在Web UI的左侧菜单栏中选择“配置 > RAW（或Config > RAW）”，查看并修改配置信息。

表 3-2 配置参数说明

参数	说明
baseUrl	https://api.modelarts-maas.com/v2
apiKey	获取API Key 。
models	支持的模型及model参数值，请参见 支持模型 。

- 首次配置：复制以下内容到Raw JSON5输入框，替换已有内容。
- ```
{
 "models": {
 "providers": {
 "hw-maas-pan": {
 "baseUrl": "https://api.modelarts-maas.com/v2",
 "apiKey": "<API_KEY>",
 "api": "openai-completions",
 "models": [
 {
 "id": "glm-5.2",
 "name": "glm-5.2",
 "contextWindow": 64000,
 "maxTokens": 64000,
 "input": [
 "text"
]
 },
 {
 "id": "deepseek-v4-flash",
 "name": "deepseek-v4-flash",
 "contextWindow": 96000,
 "maxTokens": 96000,
 "input": [
 "text"
]
 }
]
 }
 }
 },
 "agents": {
 "defaults": {
 "model": {
 "primary": "hw-maas-pan/glm-5.2"
 },
 "models": {
 "hw-maas-pan/glm-5.2": {},
 "hw-maas-pan/deepseek-v4-flash": {}
 }
 }
 },
 "gateway": {
 "mode": "local"
 }
}
```
- 已有配置：如果您已经配置过OpenClaw，请勿直接覆盖原有配置，建议根据提供的配置更新models、agents和gateway节点信息。其中models节点包含支持的模型列表。

3. 配置完成后，单击右上角“Save”保存，然后单击“Update”使配置生效。  
保存成功后，apiKey将显示为“\_\_OPENCLAW\_REDACTED\_\_”。仅用于前端界面隐藏，不影响实际调用。

终端方式

1. 在终端编辑openclaw.json配置文件。  
cat ~/.openclaw/openclaw.json
2. 按下表修改配置信息。

表 3-3 配置参数说明

| 参数      | 说明                                        |
|---------|-------------------------------------------|
| baseUrl | https://api.modelarts-maas.com/v2         |
| apiKey  | <a href="#">获取API Key</a> 。               |
| models  | 支持的模型及model参数值，请参见 <a href="#">支持模型</a> 。 |

- 首次配置：复制以下内容到配置文件。

```
{
 "models": {
 "providers": {
 "hw-maas-pan": {
 "baseUrl": "https://api.modelarts-maas.com/v2",
 "apiKey": "<API_KEY>",
 "api": "openai-completions",
 "models": [
 {
 "id": "glm-5.2",
 "name": "glm-5.2",
 "contextWindow": 64000,
 "maxTokens": 64000,
 "input": [
 "text"
]
 },
 {
 "id": "deepseek-v4-flash",
 "name": "deepseek-v4-flash",
 "contextWindow": 96000,
 "maxTokens": 96000,
 "input": [
 "text"
]
 }
]
 }
 }
 },
 "agents": {
 "defaults": {
 "model": {
 "primary": "hw-maas-pan/glm-5.2"
 },
 "models": {
 "hw-maas-pan/glm-5.2": {},
 "hw-maas-pan/deepseek-v4-flash": {}
 }
 }
 }
}
```

```
 },
 "gateway": {
 "mode": "local"
 }
 }
}
```

- 已有配置：如果您已经配置过OpenClaw，请勿直接覆盖原有配置，建议根据提供的配置更新models、agents和gateway节点信息。其中models节点包含支持的模型列表。
3. 保存文件并退出，执行以下命令使配置生效。
- ```
openclaw gateway restart
```

步骤四：使用 OpenClaw

1. 您可以任选以下方式使用OpenClaw。

- **Web UI:**

新开一个终端，执行以下命令，浏览器将自动打开OpenClaw的操作界面（Web UI），然后在Web UI中开始对话。

```
openclaw dashboard
```

- **TUI:**

新开一个终端，执行以下命令，进入TUI界面，然后开始对话。

```
openclaw tui
```

2. 切换模型

- **在当前会话切换模型（临时有效）**

在终端输入openclaw tui，进入TUI界面，使用/model <模型名称>在当前会话中切换模型。修改示例如下：

```
/model hw-maas-pan/glm-5.2
```

界面返回提示“model set to <模型名称>”，表示生效。

- **切换默认模型（永久有效）**

如需在每次新会话中使用指定模型，修改agents.defaults.model.primary字段为目标模型，请参考[步骤二：完成初始配置](#)。修改示例如下：

```
{  
  "agents": {  
    "defaults": {  
      "model": {  
        "primary": "hw-maas-pan/glm-5.2"  
      }  
    }  
  }  
}
```

3. OpenClaw会话效果如图所示：

```
When reading HEARTBEAT.md, use workspace file /root/.openclaw/workspace/HEARTBEAT.md (exact case). Do not read docs/heartbeat.md.
Current time: Monday, April 27th, 2026 - 8:18 PM (Asia/Shanghai) / 2026-04-27 12:18 UTC

HEARTBEAT_OK

请输出冒泡排序算法的python代码

def bubble_sort(arr):
    """
    Bubble Sort algorithm - O(n^2)
    Repeatedly swap adjacent elements if they're in wrong order.
    """
    n = len(arr)
    for i in range(n):
        # Flag to optimize - stop if no swaps occurred
        swapped = False
        for j in range(0, n - i - 1):
            if arr[j] > arr[j + 1]:
                arr[j], arr[j + 1] = arr[j + 1], arr[j]
                swapped = True
        # If no swaps, array is already sorted
        if not swapped:
            break
    return arr

# Example usage
if __name__ == "__main__":
    numbers = [64, 34, 25, 12, 22, 11, 90]
    print(f"Original: {numbers}")
    print(f"Sorted: {bubble_sort(numbers.copy())}")

时间复杂度:
- 最好: O(n) - 已排序时
- 平均/最坏: O(n^2)

空间复杂度: O(1) - 原地排序

特点: 简单易懂, 适合小数据集或教学演示。大数据集建议用快排或归并排序。
connected | idle
```

常见问题

● OpenClaw如何开启深度思考模式？

📖 说明

OpenClaw支持设置的深度思考level: off、minimal、low、medium、high、xhigh。具体参见 [Thinking Levels](#)。

有三种方式：

方式一：在消息前面添加内联指令：/think:<level> 或/t <level>，示例如下：

```
/think:high 你的问题
# 或者使用简写模式
/t high 你的问题
```

方式二：在会话级别开启深度思考模式，单独发送一条指令：/think:<level> 或/t <level>，示例如下：

```
/think:high
# 或者使用简写模式
/t high
```

方式三：全局设置开启深度思考模式。使用 config 命令修改全局配置：openclaw config set agents.defaults.thinkingDefault <level>，示例如下：

```
openclaw config set agents.defaults.thinkingDefault high
# 重启 gateway 才能生效
openclaw gateway restart
# 查看配置是否生效
openclaw config get agents.defaults.thinkingDefault
```

● 执行openclaw gateway restart命令后，显示“Gateway service disabled”，如何处理？

执行如下命令，手动启动gateway服务。启动成功之后会回显如下图所示信息：

```
openclaw gateway
```

```
[root@ecs-08087514-test ~]# openclaw gateway
^COpenClaw 2026-4-24 (c6c6d6) - I keep secrets like a vault... unless you print them in debug logs again.

19:57:20 [gateway] loading configuration.
19:57:20 [gateway] resolving authentication.
19:57:20 [gateway] starting...
19:57:20 [gateway] starting HTTP server...
19:57:20 [canvas] host mounted at http://127.0.0.1:18789/ _openclaw_/canvas/ (root /root/.openclaw/canvas)
19:57:20 [health-monitor] started (interval: 300s, startup-grace: 60s, channel-connect-grace: 120s)
19:57:20 [gateway] agent model: ba-mas-pai-gle-5
19:57:20 [gateway] ready (0 plugins: acps, bonjour, browser, device-pair, phone-control, talk-voice, 1.8s)
19:57:20 [gateway] log file: /tmp/openclaw/openclaw-2026-04-27.log
19:57:20 [gateway] starting channels and sidecars...
19:57:20 [plugins] bonjour: nothing detected (non-announced service; attempting re-advertise (gateway fdm-ecs-08087514-test (OpenClaw) (2) _openclaw_gw_ftp.local. host-openclaw (2).local. port-18789 state-probing)
19:57:20 [model-pricing] OpenRouter pricing fetch failed: TypeError: fetch failed
19:57:20 [plugins] embedded acps runtime backend registered (cad: /root/.openclaw/workspace)
19:57:20 [browser/server] browser control listening on http://127.0.0.1:18791/ (auth-token)
19:57:20 [heartrate] started
19:57:20 [plugins] bonjour: gateway name conflict resolved; newName="ecs-08087514-test (OpenClaw) (2)"
19:57:20 [plugins] bonjour: gateway hostname conflict resolved; newHostName="OpenClaw (2)"
19:57:20 [plugins] bonjour: advertised gateway fdm-ecs-08087514-test (OpenClaw) (2) _openclaw_gw_ftp.local. host-openclaw (2).local. port-18789 state-announcing
19:57:20 [plugins] embedded acps runtime backend probe failed: embedded ACP runtime probe failed (agent-codes; command-ops @zed-industries/codes.acp@9.11.1; cad-/root/.openclaw/workspace; ACP agent exited before initialize completed (exit=127, signal=null); /root/.npm/_npx/4280x170818c35/node_modules/@zed-industries/codes-acp-linux-x64/bin/codes-acp; error while loading shared libraries: libhol.so.3: cannot open shared object file: No such file or directory)
19:57:20 [gateway] device pairing auto-approved device-F958ba28f4f55ffff070b4dbb0b64dc31a87f0d9556c0379085a5c6b8a33 role=operator
19:02:42 [ui] res < chat.history 574100 conn=1770c414-0ea2-1ba4f02010-4c31
```

4 OpenCode

在软件开发过程中，开发者经常面临代码编写效率低、调试困难等问题。OpenCode是一款AI编程工具，通过接入大模型可帮助开发者高效完成代码开发、调测。然而，如何将这些先进的AI能力快速集成到现有的开发环境中，成为了一个新的挑战。您可以通过简单配置，把MaaS模型API接入OpenCode，从而轻松实现AI能力的集成，提高开发效率。本文档旨在帮助开发者快速配置并高效、安全地在OpenCode环境中集成MaaS模型的API。

前提条件

- 已注册华为云账号，并进行实名认证。具体操作，请参见[注册华为账号并开通华为云和实名认证](#)。
- 已开通预置服务，或已创建接入点。具体操作，请参见[开通预置服务](#)或[创建自定义接入点](#)。

计费说明

在调用MaaS模型服务时，将根据实际使用的Tokens数量进行计费。计费详情请参见[MaaS文本生成模型](#)。

约束限制

该功能仅支持“西南-贵阳一”区域。

步骤一：安装 OpenCode

1. 安装[Node.js 18或更高版本](#)。
`node -v` #运行该命令可验证Node.js版本
2. 打开命令行界面，执行以下命令，安装OpenCode。
`npm install -g opencode-ai`
3. 运行以下命令验证安装。若有版本号输出，则表示安装成功。若安装失败，请检查Node.js版本是否为18及以上。可运行`node -v`查看当前版本。
`opencode -v`

步骤二：在 OpenCode 中接入 MaaS 模型 API

1. 前往OpenCode的配置文件路径（Windows路径：C:\Users\您的用户名\.config\opencode\）下编辑opencode.json文件。本文以模型GLM-5.2为例，参考[表4-1](#)进行配置。

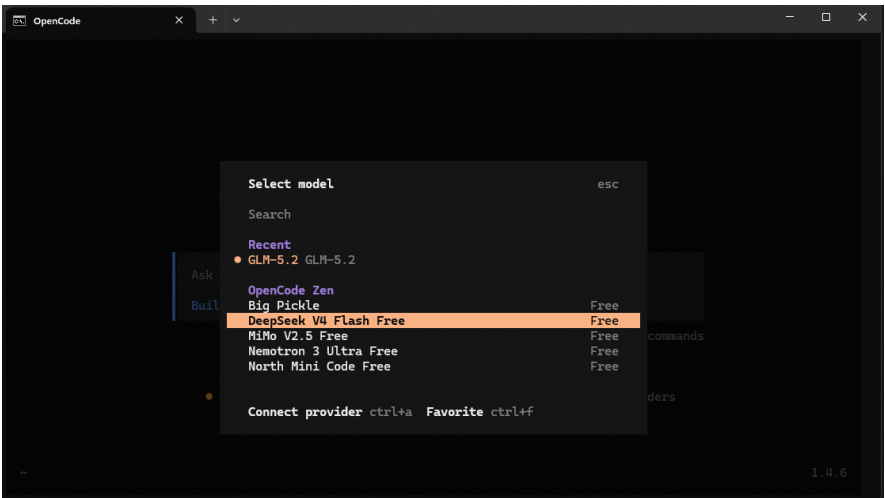
表 4-1 配置参数说明

参数	说明	示例
name	自定义模型名称	MaaS
baseURL	https://api.modelarts-maas.com/openai/v1或https://api.modelarts-maas.com/v2	https://api.modelarts-maas.com/openai/v1
apiKey	获取API Key。	*****SOYrkryjO-Jcd-X0-YQIA_oAllv*****-xVhElZgwtV3-3aXiuiac6bW*****
models	支持的模型及model参数值，请参见支持模型。	glm-5.2

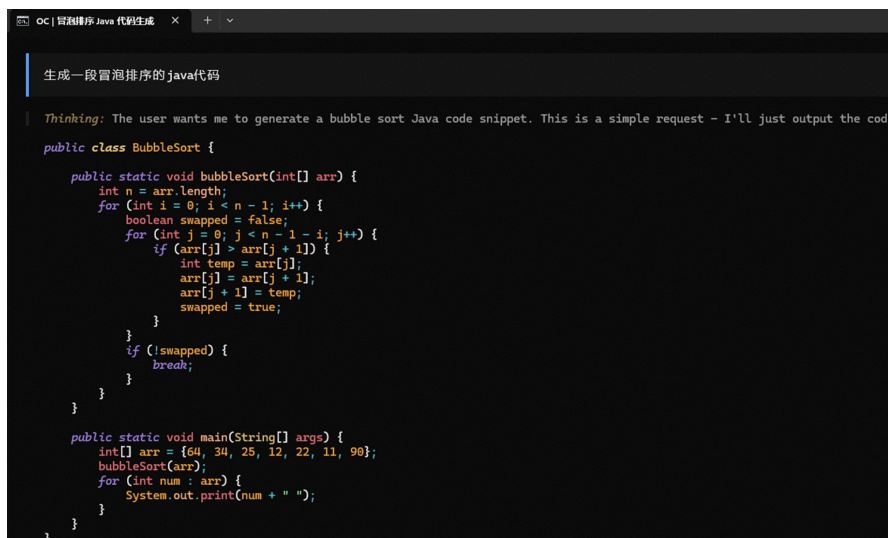
- <API_KEY>: 替换为您自己的获取API Key。

```
{
  "$schema": "https://opencode.ai/config.json",
  "provider": {
    "myprovider": {
      "npm": "@ai-sdk/openai-compatible",
      "name": "MaaS",
      "options": {
        "baseURL": "https://api.modelarts-maas.com/openai/v1",
        "apiKey": "<API_KEY>"
      },
    },
    "models": {
      "glm-5.2": {
        "name": "glm-5.2"
      }
    }
  }
}
```

2. 在终端执行如下命令开启OpenCode。
- opencode
3. 输入/models，选择配置的GLM-5.2模型并在OpenCode中使用。



4. 对话，如下图所示。



```
生成一段冒泡排序的java代码

Thinking: The user wants me to generate a bubble sort Java code snippet. This is a simple request - I'll just output the code.

public class BubbleSort {

    public static void bubbleSort(int[] arr) {
        int n = arr.length;
        for (int i = 0; i < n - 1; i++) {
            boolean swapped = false;
            for (int j = 0; j < n - 1 - i; j++) {
                if (arr[j] > arr[j + 1]) {
                    int temp = arr[j];
                    arr[j] = arr[j + 1];
                    arr[j + 1] = temp;
                    swapped = true;
                }
            }
            if (!swapped) {
                break;
            }
        }
    }

    public static void main(String[] args) {
        int[] arr = {64, 34, 25, 12, 22, 11, 90};
        bubbleSort(arr);
        for (int num : arr) {
            System.out.print(num + " ");
        }
    }
}
```

常见问题

● OpenCode如何开启深度思考模式？

- a. 编辑OpenCode的配置文件`~/config/opencode/opencode.json`，在参数`models`模型信息配置中新增`options`配置，内容为`{"thinking": {"type": "enabled"}}`。参考示例如下：

```
{
  "$schema": "https://opencode.ai/config.json",
  "provider": {
    "myprovider": {
      "npm": "@ai-sdk/openai-compatible",
      "name": "MaaS",
      "options": {
        "baseUrl": "https://api.modelarts-maas.com/openai/v1",
        "apiKey": "<API_KEY>"
      },
      "models": {
        "glm-5.2": {
          "name": "glm-5.2",
          "options": {
            "thinking": {
              "type": "enabled"
            }
          }
        }
      }
    }
  }
}
```

- b. 配置完成后执行命令“`opencode`”重新启动。如果在对话时仍看不到Thinking信息，可以在输入框按`ctrl+p`，搜索`think`，选中`Show thinking`，则可以查看思考的信息。



5 Claude Code

在软件开发过程中，开发者经常面临代码编写效率低、调试困难等问题。Claude Code 是一款AI编程工具，通过接入大模型可帮助开发者高效完成代码开发、调测。然而，如何将这些先进的AI能力快速集成到现有的开发环境中，成为了一个新的挑战。您可以通过简单配置，把MaaS模型API接入Claude Code，从而轻松实现AI能力的集成，提高开发效率。

前提条件

已开通预置服务，或已创建接入点。具体操作，请参见[开通预置服务](#)或[创建自定义接入点](#)。

计费说明

在调用MaaS模型服务时，将根据实际使用的Tokens数量进行计费。计费详情请参见[MaaS文本生成模型](#)。

约束限制

该功能仅支持“西南-贵阳一”区域。

步骤一：安装 Claude Code

1. 安装[Node.js 18或更高版本](#)。
2. 安装[Git for Windows](#)（仅限Windows用户）。
3. 打开命令行界面，执行以下命令，安装Claude Code。

```
npm install -g @anthropic-ai/claude-code
```
4. 安装结束后，执行以下命令验证安装是否成功。

```
claude --version
```

如果显示版本号，则表示安装成功。

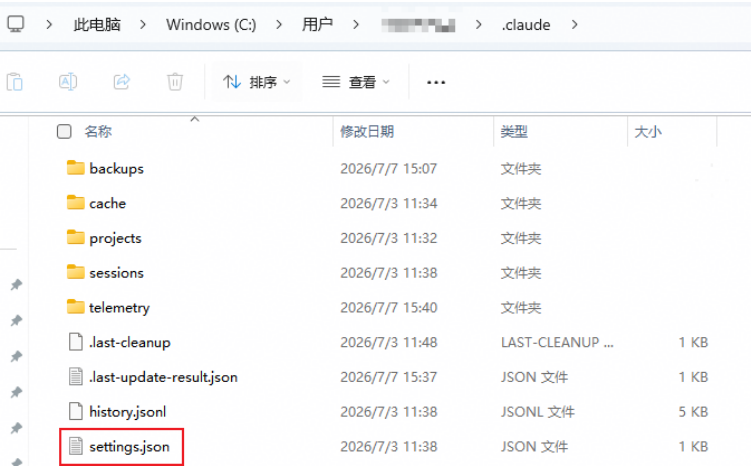
步骤二：在 Claude Code 配置 MaaS 模型 API

详细配置如下：

Windows

1. Claude Code安装完成后，前往Claude Code的配置文件路径（C:\Users\您的用户名\.claude\）下找到settings.json。如果没有settings.json文件，请在对应路径下新增一个settings.json空文件。

图 5-1 Windows 路径



2. 编辑settings.json文件。如果是新建的settings.json空文件，将下面的配置拷贝至settings.json文件。
- <API_KEY>：替换为您自己的API Key，如果没有请前往[API Key管理页面](#)创建。
 - <Model_Name>：替换为需要使用的模型信息，如glm-5.2。支持的模型信息参见[支持的模型](#)。

```
{
  "env": {
    "ANTHROPIC_AUTH_TOKEN": "<API_KEY>",
    "ANTHROPIC_BASE_URL": "https://api.modelarts-maas.com/anthropic",
    "ANTHROPIC_MODEL": "<Model_Name>",
    "ANTHROPIC_DEFAULT_HAIKU_MODEL": "<Model_Name>",
    "ANTHROPIC_DEFAULT_SONNET_MODEL": "<Model_Name>",
    "ANTHROPIC_DEFAULT_OPUS_MODEL": "<Model_Name>"
  }
}
```

示例如下：

图 5-2 配置示例

```
{
  "env": {
    "ANTHROPIC_AUTH_TOKEN": "gA4ocsWhHiQMOGjiM...",
    "ANTHROPIC_BASE_URL": "https://api.modelarts-maas.com/anthropic",
    "ANTHROPIC_MODEL": "glm-5.2",
    "ANTHROPIC_DEFAULT_HAIKU_MODEL": "glm-5.2",
    "ANTHROPIC_DEFAULT_SONNET_MODEL": "glm-5.2",
    "ANTHROPIC_DEFAULT_OPUS_MODEL": "glm-5.2"
  },
  "theme": "dark"
}
```

3. 保存配置文件，重新打开一个终端即可生效。

macOS/Linux

1. Claude Code安装完成后，前往Claude Code的配置文件路径（`~/.claude/`）下找到`settings.json`。如果没有`settings.json`文件，请在对应路径下新增一个`settings.json`空文件。
2. 编辑`settings.json`文件。如果是新建的`settings.json`空文件，将下面的配置拷贝至`settings.json`文件后再编辑。

执行如下命令编辑`settings.json`文件。

```
nano ~/.claude/settings.json
```

- `<API_KEY>`：替换为您自己的API Key，如果没有请前往[API Key管理页面](#)创建。
- `<Model_Name>`：替换为需要使用的模型信息，如`glm-5.2`。支持的模型信息参见[支持的模型](#)。

```
{
  "env": {
    "ANTHROPIC_AUTH_TOKEN": "<API_KEY>",
    "ANTHROPIC_BASE_URL": "https://api.modelarts-maas.com/anthropic",
    "ANTHROPIC_MODEL": "<Model_Name>",
    "ANTHROPIC_DEFAULT_HAIKU_MODEL": "<Model_Name>",
    "ANTHROPIC_DEFAULT_SONNET_MODEL": "<Model_Name>",
    "ANTHROPIC_DEFAULT_OPUS_MODEL": "<Model_Name>"
  }
}
```

示例如下：

图 5-3 配置示例

```
{
  "env": {
    "ANTHROPIC_AUTH_TOKEN": "gA4ocsWhHiQMOGJiM...",
    "ANTHROPIC_BASE_URL": "https://api.modelarts-maas.com/anthropic",
    "ANTHROPIC_MODEL": "glm-5.2",
    "ANTHROPIC_DEFAULT_HAIKU_MODEL": "glm-5.2",
    "ANTHROPIC_DEFAULT_SONNET_MODEL": "glm-5.2",
    "ANTHROPIC_DEFAULT_OPUS_MODEL": "glm-5.2"
  },
  "theme": "dark"
}
```

3. 保存配置文件，重新打开一个终端即可生效。

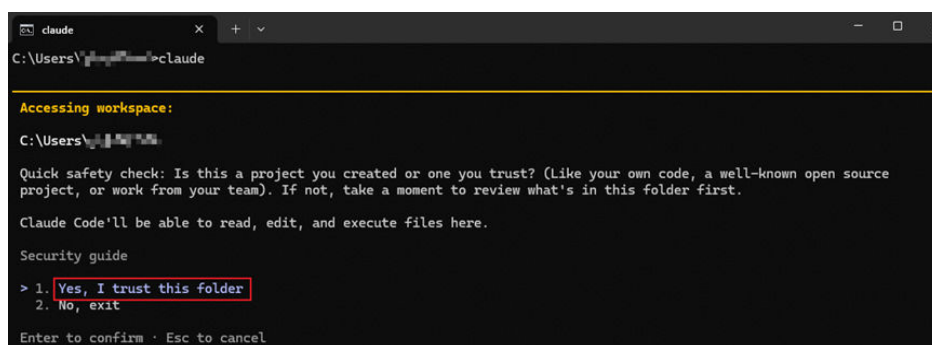
步骤三：使用 Claude Code

1. 启动Claude Code。

打开终端，进入项目目录，执行`claude`命令，即可开始使用Claude Code。

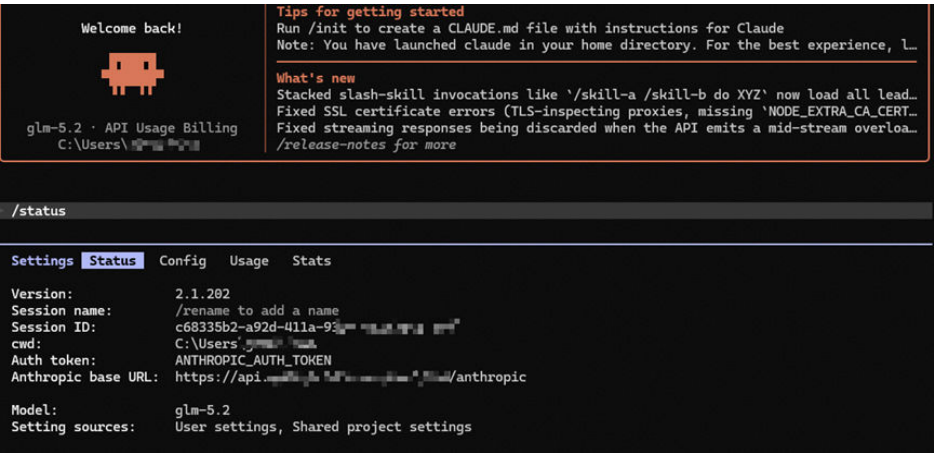
```
claude
```

启动后，选择信任此文件夹，允许Claude Code访问该文件夹中的文件。如下图所示：



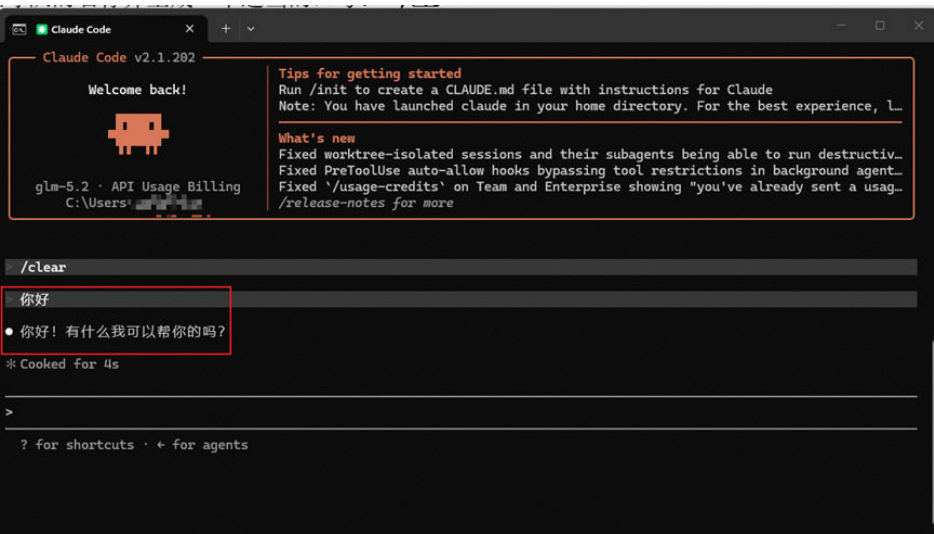
2. 模型状态验证：输入/status确认模型状态。

图 5-4 查看模型状态



3. 在Claude Code中对话。

图 5-5 对话示例



相关文档

- 在[MaaS查看在线推理的调用数据和监控指标](#)：MaaS提供调用统计功能，支持查看模型服务在指定时间段内的调用数据和监控指标详情，帮助您了解服务的使用情况和性能变化，从而更有效地进行模型评估、问题定位、故障排除和性能优化。
- 关于使用MaaS模型服务对接通讯或办公平台的最佳实践，请参见[OpenClaw最佳实践汇总](#)。

6 Dify

客服系统是企业与消费者沟通的重要桥梁，然而传统的人工客服存在着人工成本高昂、数据收集和分析困难、难以24小时提供服务等痛点。

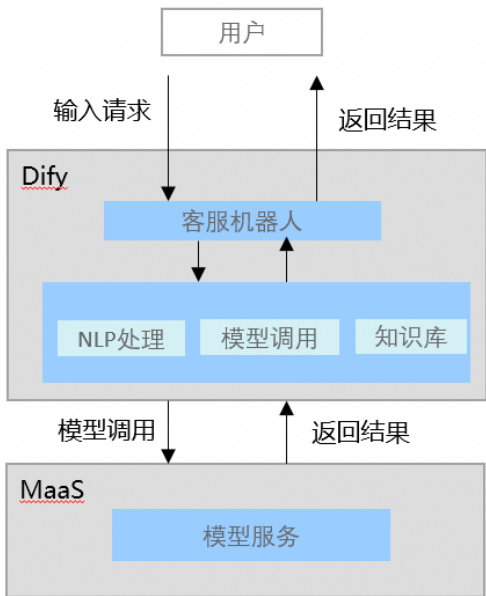
随着AI大模型的崛起，如何利用先进的大语言模型构建AI智能客服系统，提供更加自然、流畅的对话体验，提高用户使用效率，实现24*7的无间断服务，降低企业在客服场景下的投入成本，提升用户满意度成为企业和用户关注的热门领域。

基于开源框架Dify构建智能对话助手，系统支持用户直接调用大模型进行开放式问答，适用于智能客服、知识库问答、个性化推荐等多种业务场景。MaaS集成了DeepSeek、GLM等系列大模型，基于昇腾AI云服务进行全面适配和优化，使得精度和性能显著提升。

本文以GLM-5.1模型为例，介绍如何使用MaaS的模型API接入Dify（开源Agent平台），快速构建AI对话机器人并将其嵌入在网页中。

方案架构

图 6-1 方案架构



1. 用户通过Dify搭建的客服机器人界面提交问题请求。
2. Dify将请求发送给MaaS 模型服务。
3. MaaS 大语言模型处理请求后将结果返回给Dify。
4. Dify将结果优化处理后通过客服机器人界面返回给用户。

计费说明

本文支持使用预置服务进行调用。计费如下：

预置服务：提供商用级别的推理API服务，无需等待部署，开通后您可直接体验或调用模型服务，按Tokens数量计费。详细信息，请参见[推理服务计费项](#)。

约束限制

该功能仅支持“西南-贵阳一”区域。

前提条件

- 已注册华为云账号，并进行实名认证。具体操作，请参见[注册华为账号并开通华为云和实名认证](#)。
- 请参考[配置MaaS访问授权](#)进行委托授权。

步骤一：基于云服务器单机部署 Dify 平台

Dify是一个能力丰富的开源AI应用开发平台，为大型语言模型（LLM）应用的开发而设计。它巧妙地结合了后端即服务（Backend as Service）和LLMOps的理念，提供了一套易用的界面和API，加速了开发者构建可扩展的生成式AI应用的过程。

1. 基于Flexus云服务器X实例快速部署Dify平台。具体操作，请参见[快速搭建Dify-LLM应用开发平台](#)云服务器单机部署方式。

部署Dify平台的区域要与MaaS体验模型的区域一致，本文选择在“西南-贵阳一”区域部署Dify平台。

2. 部署完成后，登录Dify。

首次登录需注册管理员账号，依次填写邮箱、账号、密码后，单击“设置”进行保存后，即可登录使用。

图 6-2 注册管理员账户

步骤二：获取 MaaS 平台的模型信息

在MaaS平台，您可以调用预置服务模型。在模型调用时，需要获取模型服务的API地址和模型名称，并创建API Key。

使用预置服务

1. 在**MaaS控制台**左侧导航栏，选择“模型推理 > 在线推理”。
2. 在“预置服务”页签，在目标服务右侧的“操作”列，单击“开通服务”。
3. 在开通预置模型服务对话框，按需勾选预置服务，勾选“我已阅读并同意上述说明，及《MaaS 模型即服务声明》”，单击“一键开通”。
勾选目标预置服务后，会自动开通该服务下的所有模型版本。
4. 在已开通的预置服务右侧，单击“操作”列下的“API调用”，选择版本。
5. 在“API调用说明”页面，单击“创建 API Key”，配置相关参数，单击“确定”。

API Key用于接口的鉴权认证，最多可创建30个密钥。API Key的标签在创建完成后，不支持修改。每个密钥仅在创建时显示一次，请确保妥善保存。如果密钥丢失，无法找回，需要重新创建API Key以获取新的访问密钥。关于API Key参数说明，请参见**创建API Key**。

图 6-3 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

请输入标签

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

请输入API Key描述

权限设置

权限范围

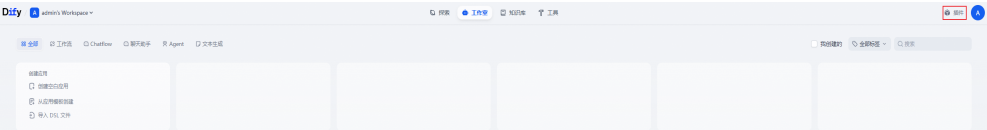
全部（访问所有资源不受限）

6. 在“您的API Key”对话框，复制密钥并保存至安全位置。保存完毕后，单击“我已保存，确认关闭”。
- 单击“我已保存，确认关闭”后将无法再次查看密钥。
7. 在“API调用说明”页面，查看调用该服务需要的API地址和模型名称信息，在后续Dify配置中使用。

步骤三：在 Dify 平台对接 MaaS 模型服务

1. 打开您的Dify平台，单击右上角“插件”按钮。

图 6-4 Dify 平台



2. 单击“探索Marketplace”按钮，在搜索框输入“MaaS”，选择“华为云MaaS平台”，单击“安装”按钮，弹出“安装插件”页面。

图 6-5 搜索插件



3. 在“安装插件”页面，单击“安装”，安装MaaS插件。

图 6-6 安装插件



说明

如已安装MaaS插件，需要更新MaaS插件至最新版本。

4. MaaS插件安装成功之后，单击Dify页面右上角用户名称，下拉并单击“设置”，进入设置界面。

图 6-7 设置界面



5. 在左侧导航栏单击“模型供应商”，单击MaaS插件右侧的“设置”按钮，配置API 密钥，如下图所示：

图 6-8 模型供应商



6. 在API 密钥授权配置界面，填写**步骤二：获取MaaS平台的模型信息**获取的API Key，单击“保存”，密钥授权成功。

图 6-9 API 密钥授权设置



7. 在插件下方单击“显示N个模型”，弹出模型列表，如下图所示。

图 6-10 添加模型

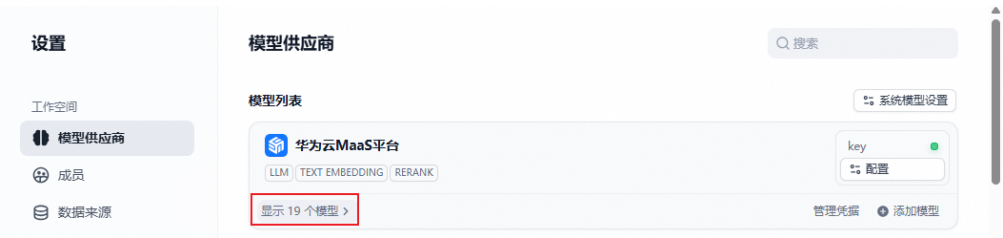
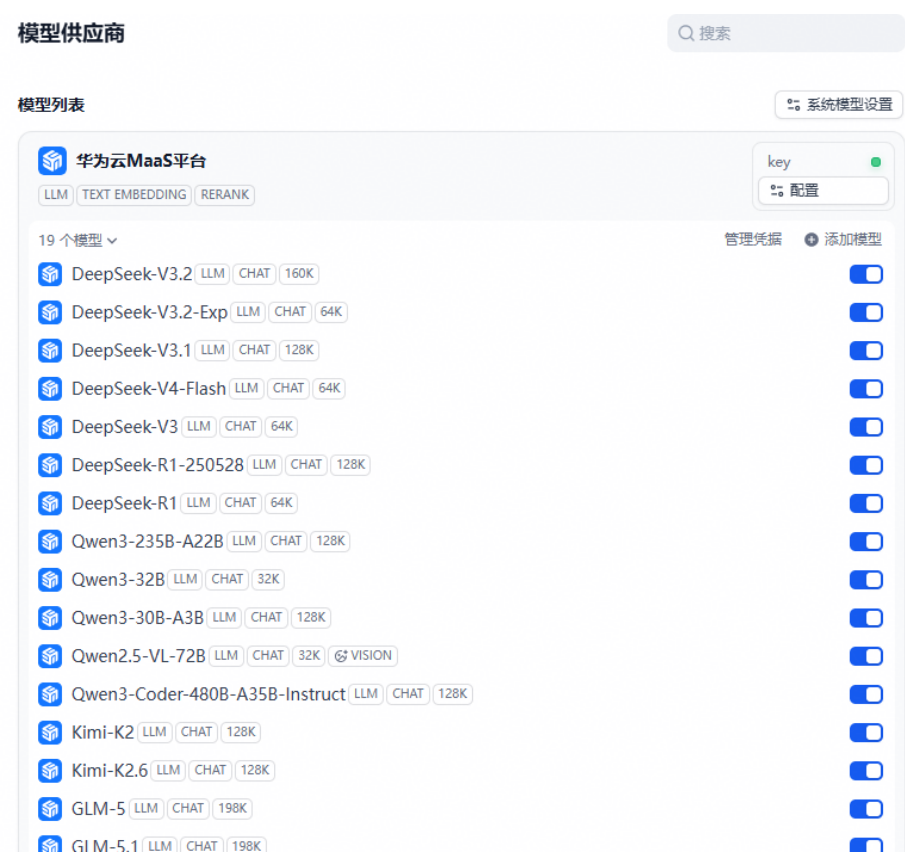


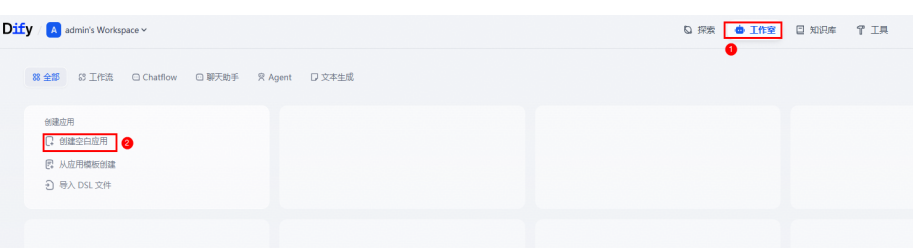
图 6-11 模型列表



步骤四：在 Dify 中创建客服机器人

- 1. 进入Dify平台“工作室”，单击“创建空白应用”。

图 6-12 创建空白应用



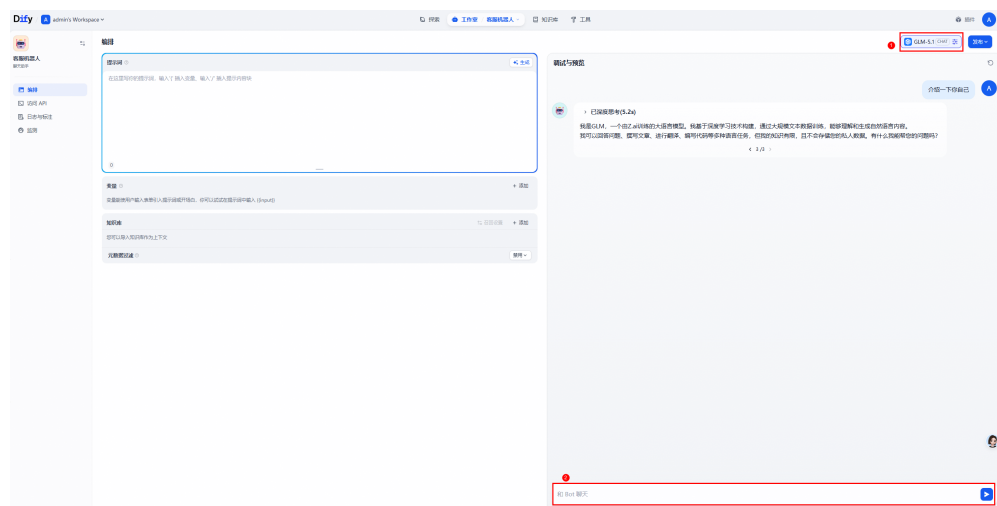
- 2. 在“创建空白应用”页面，选择“聊天助手”，填写“应用名称 & 图标”，单击“创建”，客服机器人创建成功。

图 6-13 创建应用



3. 客服机器人创建成功后，自动进入应用调试界面。右上角选择已开通的模型，在下方聊天框输入信息，模型正常答复后，说明客服机器人调试成功。

图 6-14 应用调试页面



4. 页面右上角单击“发布”，应用发布成功。可根据实际应用场景对该应用进行“发布更新”、“运行”、“嵌入网站”等操作，如下图所示：

图 6-15 应用发布页面



相关文档

Dify平台的更多应用，详情请参考[使用Dify](#)。

常见问题

在MaaS创建API Key后需要等待多久才能生效？

MaaS API Key在创建后不会立即生效，通常需要等待几分钟才能生效。

7 Cherry Studio

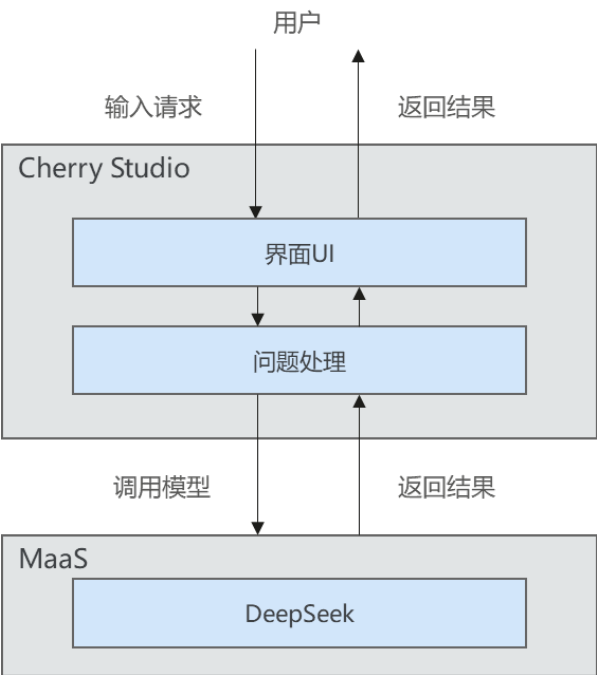
在人工智能飞速发展的当下，大语言模型（LLMs）已成为推动自然语言处理应用革新的关键力量。

与此同时，为了让用户更便捷地使用这些大语言模型，各种模型调用工具和平台应运而生。Cherry Studio便是其中一款功能强大的开源多模型桌面客户端，支持Windows、macOS和Linux系统。它集成了多种主流大语言模型（例如OpenAI、DeepSeek、Gemini等），并支持本地模型运行。此外，它还具备丰富的功能，例如对话知识库、AI绘画、翻译、多模型切换等。

本文介绍如何使用MaaS的模型API接入Cherry Studio，快速构建个人AI智能助手。

方案架构

图 7-1 方案架构



- 用户通过Cherry Studio提交问题请求。

- Cherry Studio将请求发送给MaaS 模型。
- MaaS模型处理请求后将结果返回给Cherry Studio。
- Cherry Studio将结果优化处理后通过界面返回给用户。

计费说明

本文支持使用预置服务进行调用。计费如下：

预置服务：提供商用级别的推理API服务，无需等待部署，开通后您可直接体验或调用模型服务，按Tokens数量计费。详细信息，请参见[推理服务计费项](#)。

约束限制

该功能仅支持“西南-贵阳一”区域。

前提条件

- 已注册华为云账号，并进行实名认证。具体操作，请参见[注册华为账号并开通华为云和实名认证](#)。
- 请参考[配置MaaS访问授权](#)进行委托授权。

步骤一：下载并安装 Cherry Studio

您可以通过[官方网站](#)或[开源地址](#)下载并安装Cherry Studio。

步骤二：获取 MaaS 模型 API 对接信息

获取MaaS模型服务的关键信息（模型 ID、API Key和API地址），用于后续Cherry Studio对接MaaS模型API。

使用预置服务

1. 在[MaaS控制台](#)左侧导航栏，选择“模型推理 > 在线推理”。
2. 在“预置服务”页签，在目标服务右侧的“操作”列，单击“开通服务”。
3. 在开通预置模型服务对话框，按需勾选预置服务，勾选“我已阅读并同意上述说明，及《MaaS 模型即服务声明》”，单击“一键开通”。
勾选目标预置服务后，会自动开通该服务下的所有模型版本。
4. 在已开通的预置服务右侧，单击“操作”列下的“API调用”，选择版本。
5. 在“API调用说明”页面，单击“创建 API Key”，配置相关参数，单击“确定”。
API Key用于接口的鉴权认证，最多可创建30个密钥。API Key的标签在创建完成后，不支持修改。每个密钥仅在创建时显示一次，请确保妥善保存。如果密钥丢失，无法找回，需要重新创建API Key以获取新的访问密钥。关于API Key参数说明，请参见[创建API Key](#)。

图 7-2 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

请输入标签

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

请输入API Key描述

权限设置

权限范围

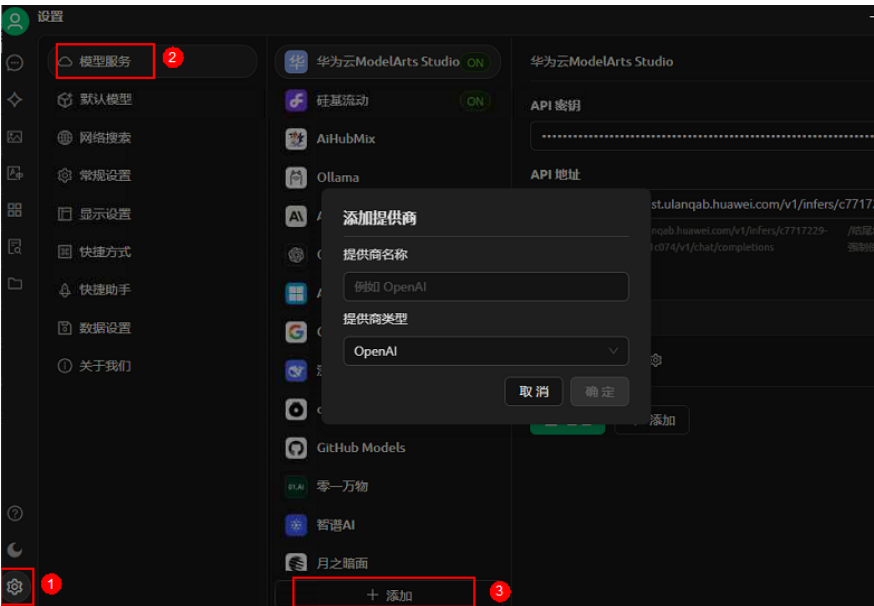
全部（访问所有资源不受限）

6. 在“您的API Key”对话框，复制密钥并保存至安全位置。保存完毕后，单击“我已保存，确认关闭”。
- 单击“我已保存，确认关闭”后将无法再次查看密钥。
7. 在“API调用说明”页面，查看调用该服务需要的API地址和模型名称信息，在后续Cherry Studio配置中使用。

步骤三：在 Cherry Studio 中配置 MaaS 模型信息

1. 在Cherry Studio添加MaaS提供商。
- a. 在Cherry Studio客户端左下角，单击设置图标，在“模型服务”中单击“添加”。

图 7-3 添加提供商



- b. 在“添加提供商”对话框，配置提供商名称和提供商类型，然后单击“确定”。

表 7-1 添加提供商参数说明

参数	说明
提供商名称	配置为“华为云MaaS”，您可以按需修改。
提供商类型	配置为“OpenAI”。

2. 在Cherry Studio添加MaaS 模型API密钥和API地址。
- a. 在Cherry Studio客户端左下角，单击设置图标。

b. 在“设置”页面，找到“华为云MaaS”选项，配置API密钥和API地址。

参数	说明
API密钥	步骤二中创建的API Key。
API地址	步骤二获取的MaaS服务的基础API地址，需要去掉地址尾部的“/v1/chat/completions”后填入。

3. 在Cherry Studio添加模型。
- a. 在“模型”区域，单击“添加”。

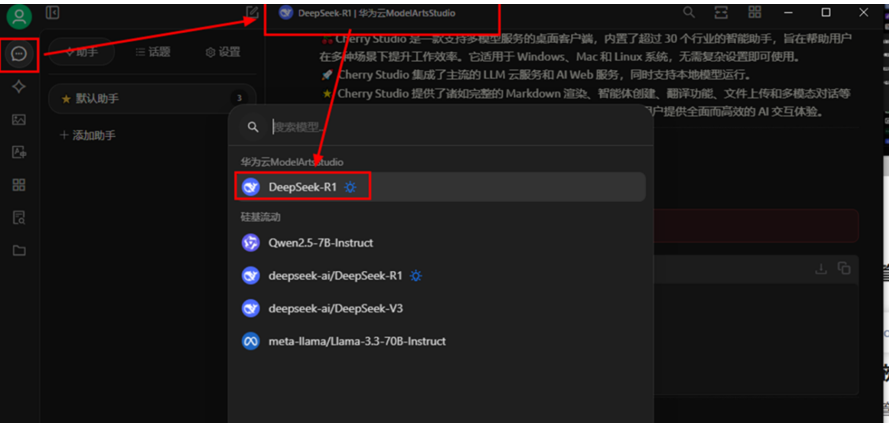
b. 在“添加模型”对话框，配置模型ID、模型名称和分组名称，单击“添加模型”。

参数	说明
模型 ID	步骤二获取的模型名称。
模型名称	自定义模型名称。
分组名称	自定义分组名称。

步骤四：在 Cherry Studio 中使用 MaaS 模型

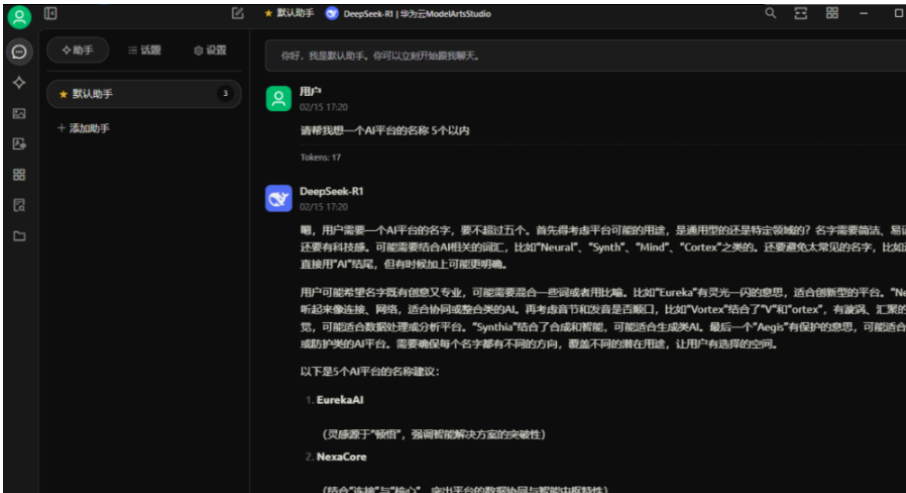
1. 在Cherry Studio左侧导航栏，单击图标，选择已配置好的模型。

图 7-4 选择模型



2. 在文本框中输入文字，开始对话。
- 您也可以选择顶部菜单中的模型名字切换模型。

图 7-5 模型问答示例



常见问题

- 在MaaS创建API Key后需要等待多久才能生效？
- MaaS API Key在创建后不会立即生效，通常需要等待几分钟才能生效。

8 Cursor

Cursor是一款基于人工智能技术的现代化代码编辑器，专为开发者设计。它结合了传统编辑器（例如VS Code）的强大功能与AI驱动的智能编程能力，能够提供智能代码补全、自然语言编程、代码库理解等核心功能，极大地提升了开发效率。此外，Cursor支持多种主流AI模型（例如OpenAI的GPT-4、DeepSeek、GLM等），并提供灵活的自定义配置选项，适合从初学者到专业开发者的广泛用户群体。

MaaS服务将DeepSeek、GLM等系列模型部署到平台，支持广大开发者进行在线体验或API调用。

本文介绍如何使用Cursor调用部署在MaaS上的DeepSeek模型，构建代码编辑器。

计费说明

本文支持使用预置服务进行调用。计费如下：

预置服务：提供商用级别的推理API服务，无需等待部署，开通后您可直接体验或调用模型服务，按Tokens数量计费。详细信息，请参见[推理服务计费项](#)。

约束限制

该功能仅支持“西南-贵阳一”区域。

前提条件

- 已注册华为云账号，并进行实名认证。具体操作，请参见[注册华为账号并开通华为云和实名认证](#)。
- 请参考[配置MaaS访问授权](#)进行委托授权。
- 已开通Cursor会员。

步骤一：下载并安装 Cursor

通过[Cursor官网](#)下载并安装Cursor。

步骤二：获取 MaaS 模型 API 对接信息

获取MaaS模型服务的关键信息（模型 ID、API Key和API地址），用于后续Cursor对接MaaS模型API。

使用预置服务

- 1. 在MaaS控制台左侧导航栏，选择“模型推理 > 在线推理”。
- 2. 在“预置服务”页签，在目标服务右侧的“操作”列，单击“开通服务”。
- 3. 在开通预置模型服务对话框，按需勾选预置服务，勾选“我已阅读并同意上述说明，及《MaaS 模型即服务声明》”，单击“一键开通”。
- 勾选目标预置服务后，会自动开通该服务下的所有模型版本。
- 4. 在已开通的预置服务右侧，单击“操作”列下的“API调用”，选择版本。
- 5. 在“API调用说明”页面，单击“创建 API Key”，配置相关参数，单击“确定”。

API Key用于接口的鉴权认证，最多可创建30个密钥。API Key的标签在创建完成后，不支持修改。每个密钥仅在创建时显示一次，请确保妥善保存。如果密钥丢失，无法找回，需要重新创建API Key以获取新的访问密钥。关于API Key参数说明，请参见[创建API Key](#)。

图 8-1 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

请输入标签

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

请输入API Key描述

权限设置

权限范围

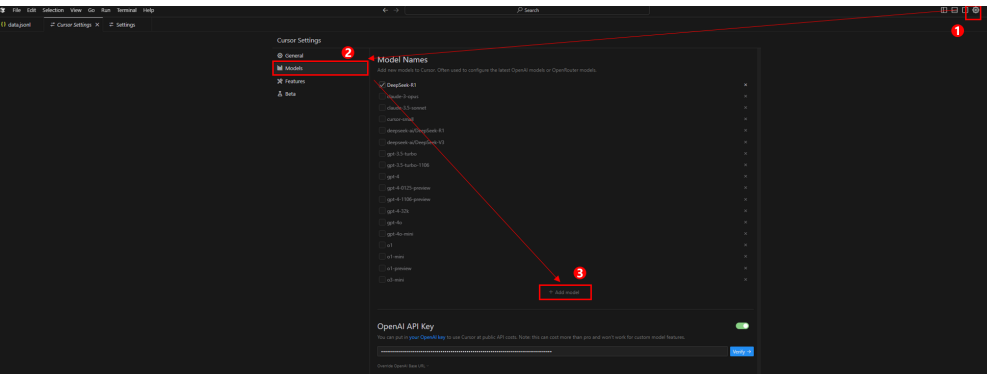
全部（访问所有资源不受限）

- 6. 在“您的API Key”对话框，复制密钥并保存至安全位置。保存完毕后，单击“我已保存，确认关闭”。
- 单击“我已保存，确认关闭”后将无法再次查看密钥。
- 7. 在“API调用说明”页面，查看调用该服务需要的API地址和模型名称信息，在后续配置中使用。

步骤三：在 Cursor 中配置 MaaS API

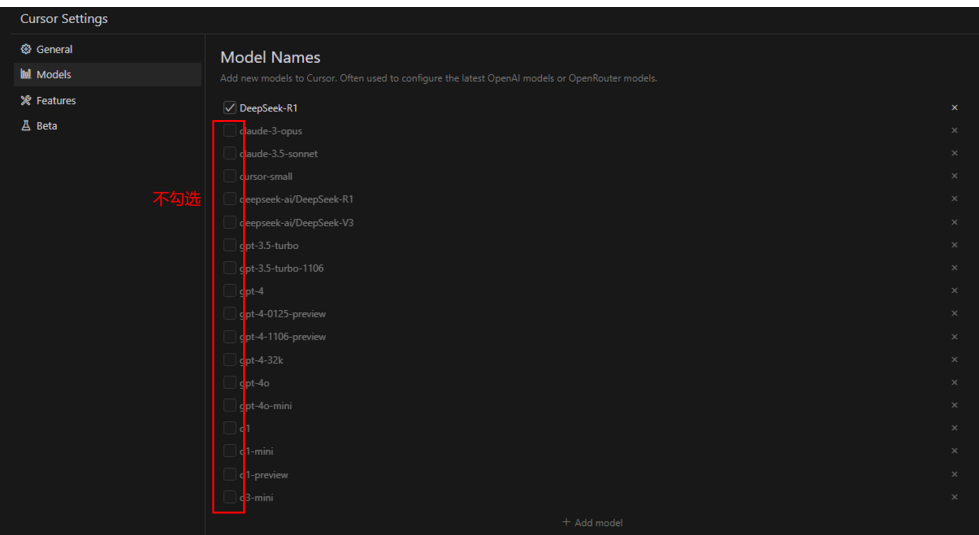
- 1. 在Cursor平台右上角单击设置图标。
- 2. 在“Cursor Settings”页面左侧导航栏，单击“Models”，然后单击“Add model”。

图 8-2 添加模型



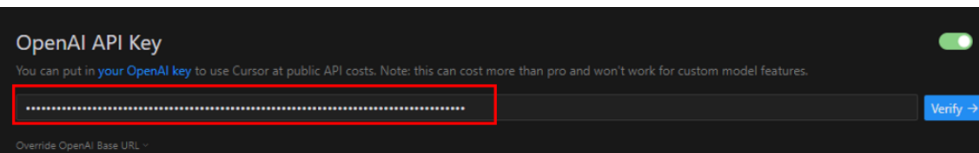
3. 在文本框中输入**步骤二**获取的模型名称，然后单击右侧的“Add model”。
4. 仅勾选刚添加的MaaS模型，其余模型去勾选（否则验证时可能会出现调不通的问题）。

图 8-3 勾选 MaaS 模型



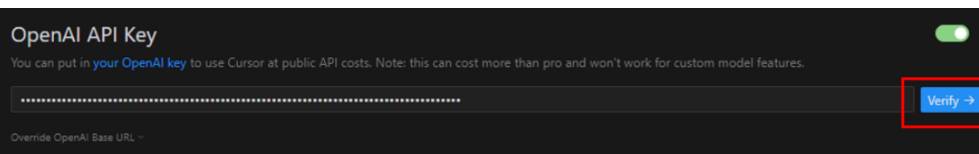
5. 在“OpenAI Key”区域填写**步骤二**创建的API Key。

图 8-4 填写 API Key



6. 单击“Override Openai Base URL”，修改基础接口地址，填入**步骤二**获取的接口地址（需去掉尾部的/chat/completions），单击“Save”。
7. 单击“Verify”验证接口连通性。如果无报错信息则配置成功，可以开始使用。

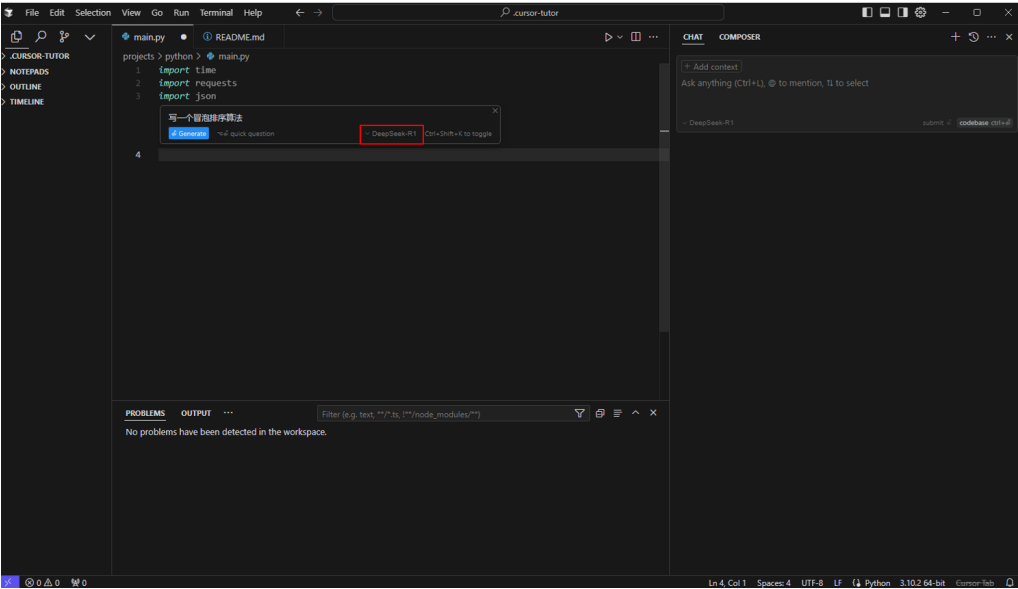
图 8-5 验证接口连通性



步骤四：在 Cursor 中使用 MaaS API 进行代码生成

在代码编辑页面，红框处选择刚配置好的模型即可进行对话、代码生成、代码解析等操作。

图 8-6 使用 MaaS API



9 Cline

Cline是一款基于大语言模型的VS Code插件，能够有效处理复杂的软件开发任务。借助VS Code开发平台，Cline为开发者带来了便捷高效的编程体验。Cline的优势如下：

- 深度融合MaaS：Cline支持接入MaaS平台的多个系列模型服务。
- 文件管理与代码纠错：支持便捷地创建和编辑文件，实时监控Linter和编译器的错误信息。一旦发现代码中存在缺少导入、语法错误等问题，能迅速进行分析诊断，并给出对应的修复建议，极大地提升代码编写的流畅性和准确性，保障开发工作进行顺利。
- 终端交互与即时响应：集成便捷的终端交互界面，允许开发者在其中执行各类命令，并实时查看命令输出。当文件编辑完成后，Cline可帮助开发者快速定位并解决服务器出现的问题，使开发流程紧密衔接，有效提高开发效率。
- Web一站式解决方案：对于Web开发任务，Cline可以在无头浏览器中启动网站，自动模拟用户的单击、输入、滚动等操作，并实时捕获截图和控制台日志。通过对这些数据的深入分析，精准定位并修复运行时的错误和视觉错误，确保Web应用的高质量交付。

本文介绍如何使用Cline调用部署在MaaS上的模型，构建AI编程助手。

约束限制

该功能仅支持“西南-贵阳一”区域。

计费说明

本文支持使用预置服务进行调用。计费如下：

预置服务：提供商用级别的推理API服务，无需等待部署，开通后您可直接体验或调用模型服务，按Tokens数量计费。详细信息，请参见[推理服务计费项](#)。

前提条件

- 已注册华为云账号，并进行实名认证。具体操作，请参见[注册华为账号并开通华为云和实名认证](#)。
- 请参考[配置MaaS访问授权](#)进行委托授权。

支持的模型

支持文本生成类的模型，且模型的上下文长度（序列长度）≥32K。

您可以登录[MaaS控制台](#)，在“模型广场”页面右上角，单击“全部模型”，在“筛选”区域，“模型类型”选择“文本生成”，“上下文长度”选择“32K”和“64K”，查看支持的模型。

图 9-1 在模型广场查看支持的模型

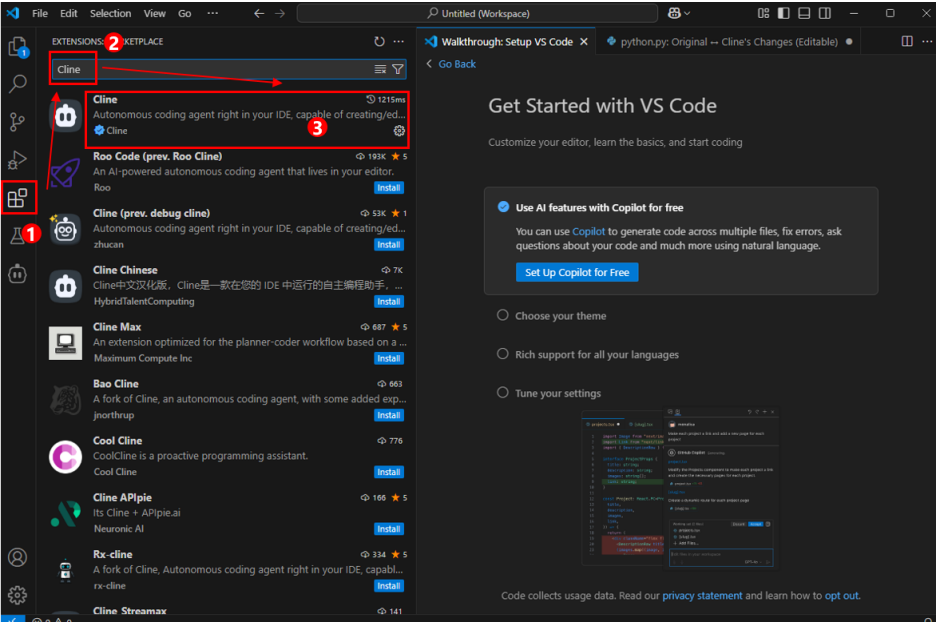


步骤一：在 VS Code 中安装 Cline

1. 打开VS Code，在左侧导航栏单击图标，在搜索框中输入“Cline”，然后单击“Install”。

在左侧导航栏出现小机器人图标，表示Cline安装完成。

图 9-2 安装 Cline



步骤二：MaaS 模型 API 接入准备

获取MaaS模型服务的关键信息（模型 ID、API Key和API地址），用于后续Cline对接MaaS模型API。

使用预置服务

1.

在MaaS控制台左侧导航栏，选择“模型推理 > 在线推理”。
2.

在“预置服务”页签，在目标服务右侧的“操作”列，单击“开通服务”。
3.

在开通预置模型服务对话框，按需勾选预置服务，勾选“我已阅读并同意上述说明，及《MaaS 模型即服务声明》”，单击“一键开通”。
勾选目标预置服务后，会自动开通该服务下的所有模型版本。
4.

在已开通的预置服务右侧，单击“操作”列下的“API调用”，选择版本。
5.

在“API调用说明”页面，单击“创建 API Key”，配置相关参数，单击“确定”。
- API Key用于接口的鉴权认证，最多可创建30个密钥。API Key的标签在创建完成后，不支持修改。每个密钥仅在创建时显示一次，请确保妥善保存。如果密钥丢失，无法找回，需要重新创建API Key以获取新的访问密钥。关于API Key参数说明，请参见[创建API Key](#)。
- 文档版本 01 (2026-10-12)

版权所有 © 华为云计算技术有限公司

44

图 9-3 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

请输入标签

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

请输入API Key描述

权限设置

权限范围

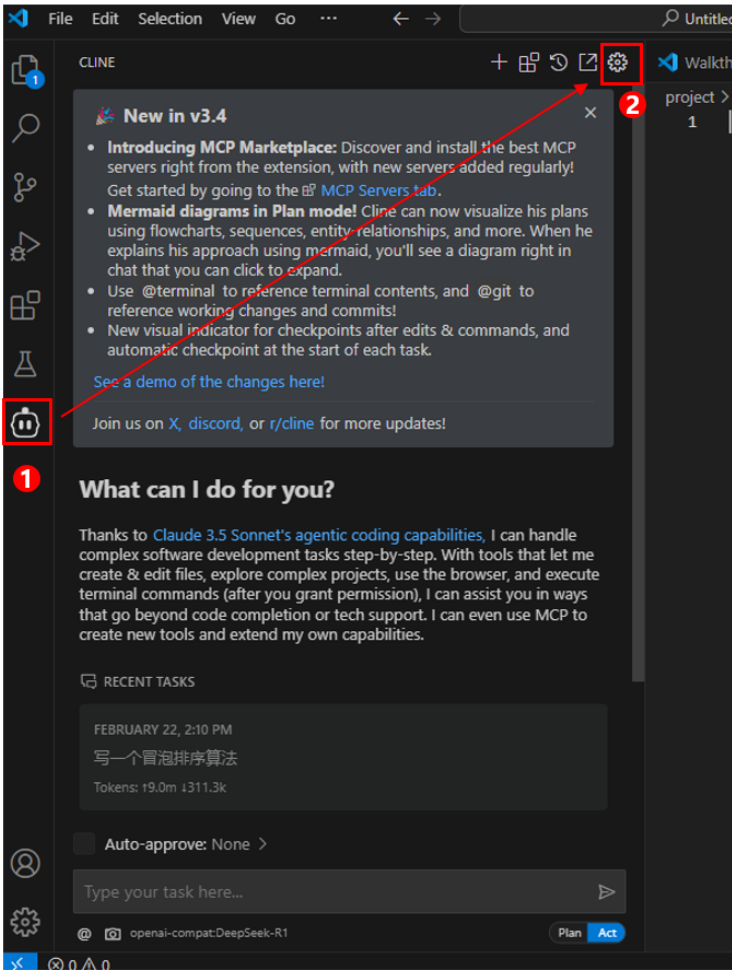
全部（访问所有资源不受限）

6. 在“您的API Key”对话框，复制密钥并保存至安全位置。保存完毕后，单击“我已保存，确认关闭”。
- 单击“我已保存，确认关闭”后将无法再次查看密钥。
7. 在“API调用说明”页面，查看调用该服务需要的API地址和模型名称信息，在后续配置中使用。

步骤三：在 Cline 中配置 MaaS API

1. 配置MaaS模型服务。
- a. 打开VS Code，在左侧导航栏单击图标，打开Cline插件，在右上角单击图标。

图 9-4 打开 Cline 插件



b. 在“Settings”页面，配置相关信息，然后单击“Done”。

表 9-1 Cline 配置说明

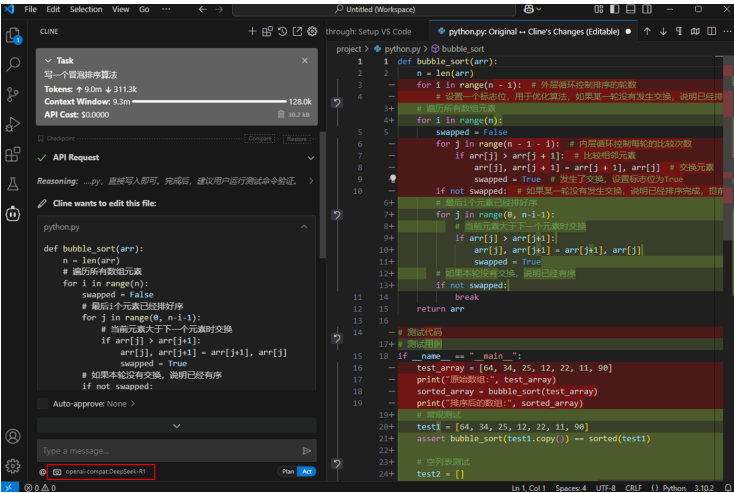
参数	说明
API Provider	选择“OpenAI Compatible”。
Base URL	步骤二获取的API地址，需要去掉尾部的“/chat/completions”后填入。
API Key	步骤二创建的API Key。
Model ID	步骤二获取的模型名称。

2. 通过VS Code的Cline插件调用MaaS API实现自动化代码生成。

a. 在VS Code左侧导航栏，单击图标。

- b. 在左下角红框处选择已配置的MaaS服务，进行对话和代码辅助生成。
Cline功能强大，可以进行代码生成，对写好的代码进行纠错、优化等操作。

图 9-5 代码生成示例



10 RAGFlow

RAGFlow是一款基于深度文档理解构建的开源RAG引擎，专为处理复杂数据结构与生成高质量问答而设计。该引擎结合了先进的检索技术和大型语言模型（LLM），支持多种数据格式（例如PDF、Word、Excel、图片等）的深度解析和知识提取。通过多路召回、融合重排序和可视化文本切片等功能，RAGFlow提供了精准、可解释且高效的信息处理能力，适用于从个人知识库到企业级智能客服的多元化应用场景。

MaaS服务将DeepSeek系列模型部署到平台，基于华为云昇腾云服务的全栈优化适配，可获得持平全球高端GPU部署模型的效果，提供稳定的生产级服务能力，满足业务商用部署需求，支持广大开发者在线体验或端外调用。

本文介绍如何使用RAGFlow调用部署在MaaS上的DeepSeek模型，快速构建AI助理。

计费说明

本文支持使用预置服务进行调用。计费如下：

预置服务：提供商用级别的推理API服务，无需等待部署，开通后您可直接体验或调用模型服务，按Tokens数量计费。详细信息，请参见[推理服务计费项](#)。

约束限制

该功能仅支持“西南-贵阳一”区域。

前提条件

- 已注册华为云账号，并进行实名认证。具体操作，请参见[注册华为账号并开通华为云和实名认证](#)。
- 请参考[配置MaaS访问授权](#)进行委托授权。

步骤一：下载并安装 RAGFlow

您可以通过[官方网站](#)或[Github](#)下载并安装RAGFlow。

步骤二：MaaS 模型 API 接入准备

获取MaaS DeepSeek模型服务的关键信息，用于后续RAGFlow对接MaaS DeepSeek。

使用预置服务

1. 在**MaaS控制台**左侧导航栏，选择“模型推理 > 在线推理”。
2. 在“预置服务”页签，在目标服务右侧的“操作”列，单击“开通服务”。
3. 在开通预置模型服务对话框，按需勾选预置服务，勾选“我已阅读并同意上述说明，及《MaaS 模型即服务声明》”，单击“一键开通”。
勾选目标预置服务后，会自动开通该服务下的所有模型版本。
4. 在已开通的预置服务右侧，单击“操作”列下的“API调用”，选择版本。
5. 在“API调用说明”页面，单击“创建 API Key”，配置相关参数，单击“确定”。

API Key用于接口的鉴权认证，最多可创建30个密钥。API Key的标签在创建完成后，不支持修改。每个密钥仅在创建时显示一次，请确保妥善保存。如果密钥丢失，无法找回，需要重新创建API Key以获取新的访问密钥。关于API Key参数说明，请参见[创建API Key](#)。

图 10-1 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

请输入标签

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

请输入API Key描述

权限设置

权限范围

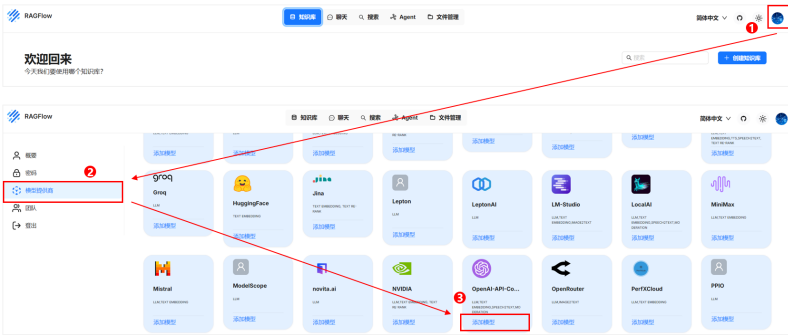
全部（访问所有资源不受限）

6. 在“您的API Key”对话框，复制密钥并保存至安全位置。保存完毕后，单击“我已保存，确认关闭”。
7. 在“API调用说明”页面，查看调用该服务需要的API地址和模型名称信息，在后续配置中使用。

步骤三：在 RAGFlow 中新增模型供应商

1. 在RAGFlow平台右上角，单击用户头像，选择“模型提供商”，在“OpenAI-API-Compatible”卡片中单击“添加模型”。

图 10-2 新增模型供应商



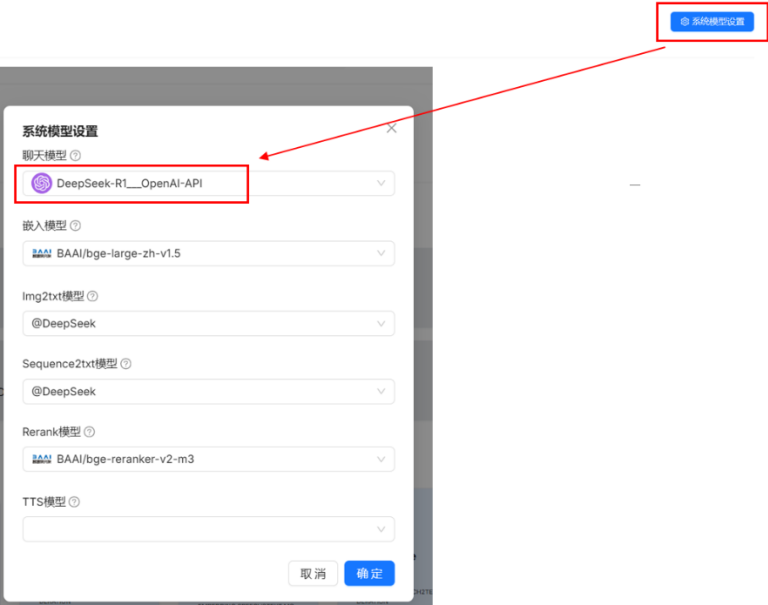
2. 在“添加 LLM”对话框，配置相关信息，然后单击“确定”。

表 10-1 添加 LLM 参数说明

参数	说明															
模型类型	选择“chat”。															
模型名称	步骤二获取的模型名称。															
基础 Url	步骤二获取的基础API地址，需要去掉地址尾部的“/chat/completions”后填入。															
API-Key	步骤二创建的API Key。															
最大 token数	即最大输出长度，您可以在“模型广场 > 模型详情”页面进行查看。具体操作，请参见在模型广场查看预置模型。 例如，最大输出长度为32K，K代表1024，32K=32*1024=32768，最大token数为32768。 图 10-3 最大输出长度 <table><tr><td colspan="3">基本信息</td></tr><tr><td>模型类型</td><td>支持语言</td><td>参数大小</td></tr><tr><td>文本生成</td><td>中文、英文</td><td>685B</td></tr><tr><td>支持输出类型</td><td>上下文长度</td><td>最大输入输出长度</td></tr><tr><td>文本</td><td>160K</td><td>128K/32K</td></tr></table>	基本信息			模型类型	支持语言	参数大小	文本生成	中文、英文	685B	支持输出类型	上下文长度	最大输入输出长度	文本	160K	128K/32K
基本信息																
模型类型	支持语言	参数大小														
文本生成	中文、英文	685B														
支持输出类型	上下文长度	最大输入输出长度														
文本	160K	128K/32K														

3. 修改系统模型设置。
- 在“模型供应商”页面右上角，单击“系统模型设置”，将“聊天模型”设置为已配置的MaaS模型。嵌入和Rerank模型可使用系统预置。

图 10-4 修改系统模型设置



步骤四：在 RAGFlow 中使用接入的 MaaS API 进行知识库问答

1. 创建知识库。
 - a. 在 RAGFlow 顶部导航栏，单击“知识库”，在右上角单击“创建知识库”，输入名称，单击“确定”。

图 10-5 创建知识库

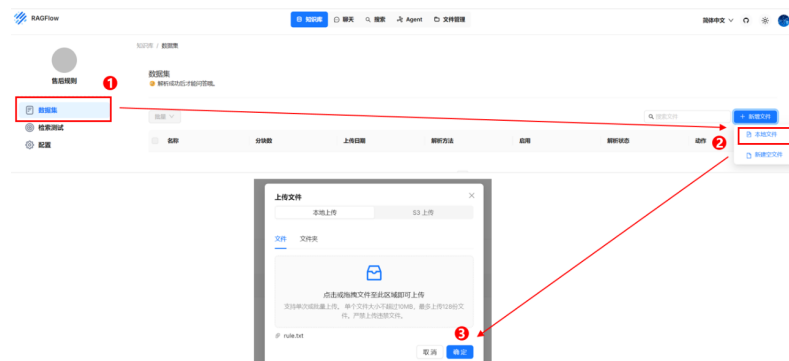


- b. 在知识库左侧导航栏，单击“配置”，按需配置文档解析、文档切分等规则。

图 10-6 配置知识库

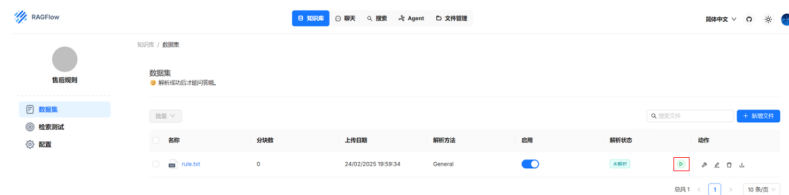


2. 上传文档。
 - a. 在知识库左侧导航栏，单击“数据集”，在右侧单击“新增文件 > 本地文件”，上传需要进行问答的知识文档，单击“确定”。



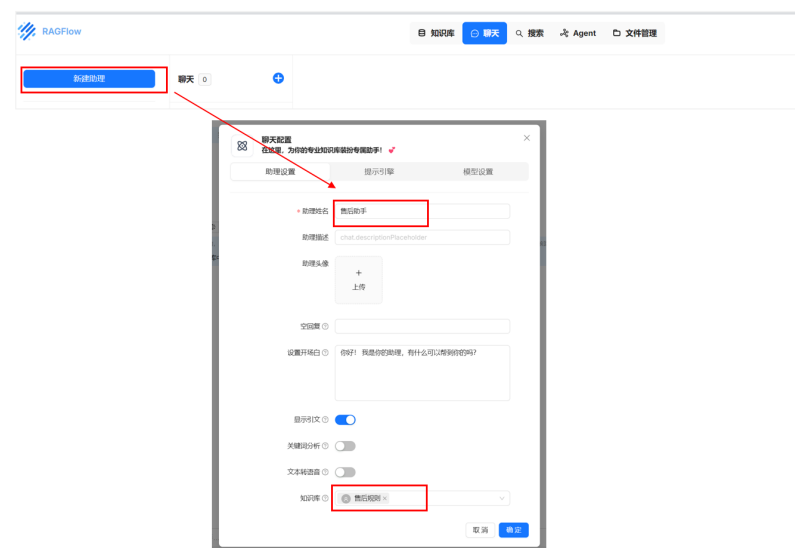
b. 上传完成后，在数据集列表中，单击目标数据集右侧的运行按钮。

图 10-7 运行数据集



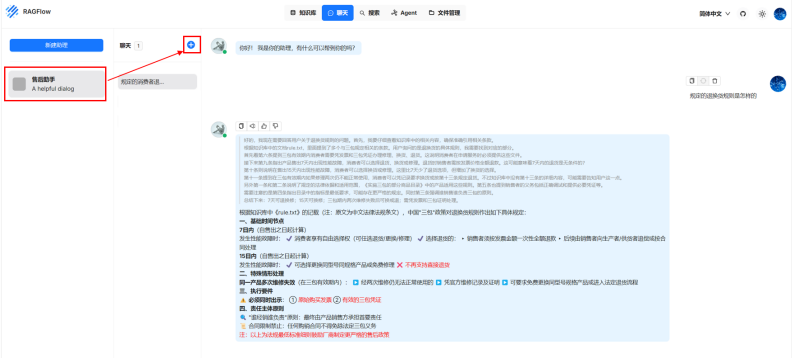
3. 创建助理。
- a. 在RAGFlow顶部导航栏，单击“聊天”。
 - b. 在“聊天”页面，单击“新建助理”，在“聊天配置”对话框，自定义助理姓名，选择已创建的知识库，然后单击“确定”。

图 10-8 配置助理信息



4. 在“聊天”页面，选择已创建的AI助理，单击 图标，进行文档问答。模型会基于已上传知识库中的内容进行问题答复。

图 10-9 文档问答示例



11 Deep Research

Deep Research是一款新型AI深度研究工具，它能够通过联网搜索和多步骤推理，快速生成专业级的研究报告，帮助用户在短时间内完成复杂的多步骤研究任务。该工具适用于金融、科学、政策、工程等领域，特别适用于需要深度分析和小众信息搜索的场景。

MaaS服务将多个系列模型部署到平台，基于华为云昇腾云服务的全栈优化适配，可获得持平全球高端GPU部署模型的效果，提供稳定的生产级服务能力，满足业务商用部署需求，支持广大开发者在线体验或端外调用。

本文介绍如何使用Deep Research调用部署在MaaS上的模型，快速实现行业洞察。

计费说明

本文支持使用预置服务进行调用。计费如下：

预置服务：提供商用级别的推理API服务，无需等待部署，开通后您可直接体验或调用模型服务，按Tokens数量计费。详细信息，请参见[推理服务计费项](#)。

约束限制

该功能仅支持“西南-贵阳一”区域。

前提条件

- 已注册华为云账号，并进行实名认证。具体操作，请参见[注册华为账号并开通华为云和实名认证](#)。
- 请参考[配置MaaS访问授权](#)进行委托授权。

步骤一：下载并安装 Deep Research

您可以通过[官方网站](#)或[Github](#)下载并安装Deep Research。

步骤二：MaaS 模型 API 接入准备

获取MaaS模型服务的关键信息，用于后续Deep Research对接MaaS模型API。

使用预置服务

1. 在MaaS控制台左侧导航栏，选择“模型推理 > 在线推理”。
2. 在“预置服务”页签，在目标服务右侧的“操作”列，单击“开通服务”。
3. 在开通预置模型服务对话框，按需勾选预置服务，勾选“我已阅读并同意上述说明，及《 MaaS 模型即服务声明》”，单击“一键开通”。
- 勾选目标预置服务后，会自动开通该服务下的所有模型版本。
4. 在已开通的预置服务右侧，单击“操作”列下的“API调用”，选择版本。
5. 在“API调用说明”页面，单击“创建 API Key”，配置相关参数，单击“确定”。

API Key用于接口的鉴权认证，最多可创建30个密钥。API Key的标签在创建完成后，不支持修改。每个密钥仅在创建时显示一次，请确保妥善保存。如果密钥丢失，无法找回，需要重新创建API Key以获取新的访问密钥。关于API Key参数说明，请参见[创建API Key](#)。

图 11-1 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

请输入标签

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

请输入API Key描述

权限设置

权限范围

全部（访问所有资源不受限）

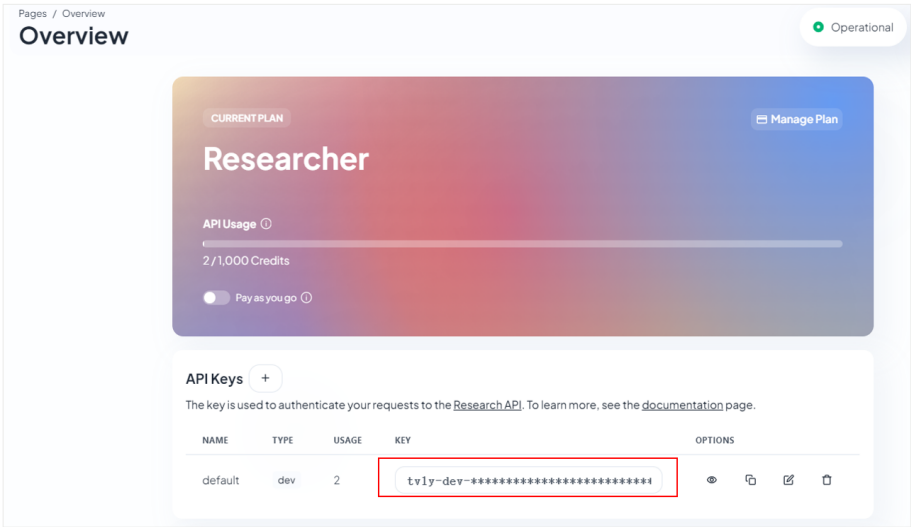
6. 在“您的API Key”对话框，复制密钥并保存至安全位置。保存完毕后，单击“我已保存，确认关闭”。
- 单击“我已保存，确认关闭”后将无法再次查看密钥。
7. 在“API调用说明”页面，查看调用该服务需要的API地址和模型名称信息，在后续配置中使用。

步骤三：联网搜索 API 接入准备

Tavily AI是专为AI大语言模型设计的搜索引擎，能够提供实时、基于事实的搜索结果。它通过API连接AI应用，获取可靠的网络知识。每个用户每月享有免费调用额度，具体额度请参考[Tavily AI官网](#)。

1. 访问[Tavily AI官网](#)并注册账号。
2. 在“Overview”页面的“API Keys”区域获取API Key，在后续Deep Research配置中使用。

图 11-2 获取 API Key



步骤四：在 Deep Research 中配置 MaaS API

1. 添加MaaS模型与联网搜索服务。
- a. 在Deep Research页面右上角，单击设置图标。
- b. 在“设置”对话框，配置相关信息，然后单击“保存”。

表 11-1 设置参数说明

参数	说明
AI服务	选择“OpenAI Compatible”。
API密钥	步骤二 创建的API Key。
API Base URL	步骤二 获取的基础API地址，需要去掉地址尾部的“/chat/completions”后填入。
模型名称	步骤二 获取的模型名称。
上下文长度	即最大输出长度，您可以在“模型广场 > 模型详情”页面进行查看。具体操作，请参见在模型广场查看预置模型。 例如，最大输出长度为32K，K代表1024，32K=32*1024=32768，最大token数为32768。 图 11-3 最大输出长度 基本信息 模型类型 文本生成 支持输出类型 文本 支持语言 中文、英文 上下文长度 160K 参数大小 685B 最大输入/输出长度 128K/32K
联网搜索服务	选择Tavily服务商，API密钥配置为 步骤三：联网搜索API接入准备 获取的API Key。

2. 在Deep Research中使用MaaS API。
- a. 在Deep Research页面的“研究主题”区域，填写想要研究的主题，设置问题数量、研究深度和研究广度，单击“开始研究”。

图 11-4 配置研究相关信息

Deep Researchv1.1.7



中文

Deep Research 是 [dzhng/deep-research](#) 的可视化 UI，可以让 AI 根据特定问题联网搜索并自行深挖，并输出研究报告。本项目可以流式传输 AI 的回答来实时反馈，并使用树状结构可视化搜索过程。
全部 API 请求都在浏览器本地完成。

1. 研究主题

研究主题 *

显卡行业的趋势预测

问题数量 *

1

AI 询问你的问题数量。这些问题能让 AI 更好地了解你的研究目标。

研究深度 (Depth) *

1

联网搜索的迭代轮数。

研究广度 (Breadth) *

1

第一次迭代中的搜索次数。后续每轮迭代的搜索次数为上一轮的一半。

开始研究

2. 模型反馈

AI 将会跟你确认一些细节，帮助你明确研究方向。

思考完毕

您更关注显卡行业的哪些细分领域或应用场景（例如 游戏显卡、AI 计算卡、数据中心 GPU、移动端 GPU 等）？这将帮助我聚焦在特定技术路线、市场动态和竞争格局的分析上。

游戏显卡

提交回答

- b. 在Deep Research页面的“模型反馈”区域，模型会基于输入信息进行研究方向的确认，回答后单击“提交回答”。
- c. 模型会根据用户的回答生成一个工作流，展示研究的具体流程和目标，然后模型会调用Tavily AI的联网搜索接口获取相关的信息。您可以单击节点查看搜索详情。

图 11-5 联网搜索



图 11-6 研究报告示例



12 TRAE

TRAE是一款AI原生集成开发环境（AI IDE），深度整合AI大模型与开发流程，实现从代码生成到项目部署的全流程自动化。TRAE支持接入MaaS模型API，从而提供更多可选择的模型。本文主要介绍如何配置MaaS模型API。

前提条件

- 已注册华为云账号，并进行实名认证。具体操作，请参见[注册华为账号并开通华为云和实名认证](#)。
- 已开通预置服务，或已创建接入点。具体操作，请参见[开通预置服务](#)或[创建自定义接入点](#)。

计费说明

在调用MaaS模型服务时，将根据实际使用的Tokens数量进行计费。计费详情请参见[MaaS文本生成模型](#)。

约束限制

该功能仅支持“西南-贵阳一”区域。

步骤一：安装 TRAE

在[TRAE官网](#)下载并安装对应操作系统的版本。

步骤二：接入 MaaS 模型 API

步骤1 选择“个人用户”入口登录后，单击左下角用户名，单击“设置”。

图 12-1 设置



- 步骤2** 进入设置页面，在左侧导航栏，选择“模型”，在模型管理页面单击“添加模型”。
- 步骤3** 在“添加模型”页面，单击“自定义模型”，进入“自定义模型”页面，如下图所示：

图 12-2 自定义模型页面



- 步骤4** 在自定义模型页面，参考下表配置。

表 12-1 自定义模型配置

参数	说明	示例
API格式	必填。 - OpenAI Chat Completions 格式 - Anthropic Messages 格式	选择OpenAI Chat Completions 格式

参数	说明	示例
自定义请求地址	必填。 完整URL开关关闭时，URL不带/chat/completions（OpenAI）或 /messages（Anthropic）。 - OpenAI Chat Completions 格式。完整路径如下： - https://api.modelarts-maas.com/v2/chat/completions - https://api.modelarts-maas.com/openai/v1/chat/completions - Anthropic Messages 格式。完整路径如下： https://api.modelarts-maas.com/anthropic/v1/messages	https://api.modelarts-maas.com/v2
模型ID	必填。 从模型列表中的“模型参数”中获取。	deepseek-v4.1-flash
API密钥	必填。 可使用已有的API Key，如果没有可获取API Key。	gA4ocs*****Tc-Kz4AQ
模型展示名称	选填。 可自定义模型名称。	maas-deepseek-v4.1-flash
高级配置	选填。可设置思考模式、采样参数、上下文长度等。	-

图 12-3 配置示例

编辑模型

API 格式

OpenAI Chat Completions 格式

自定义请求地址

完整 URL

请填写兼容 OpenAI API 的服务端点地址，不要以斜杠结尾。/chat/completions 将会被补充到你填写的地址末尾。

https://api.modelarts-maas.com/v2

模型 ID

deepseek-v4.1-flash

模型展示名称

在模型列表中展示的名称，未设置时默认显示 Model ID。

maas-deepseek-v4.1-flash24/64

API 密钥

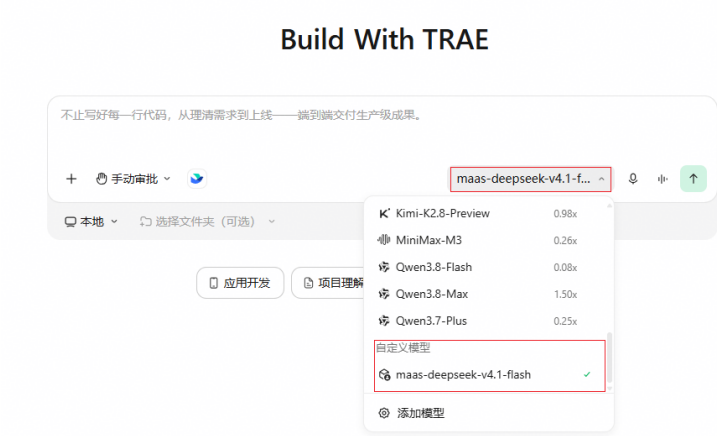
步骤5 单击“添加模型”，连通性测试通过之后，表示模型添加成功。

----结束

步骤三：使用 TRAE

步骤1 在TRAE主页面左侧导航栏，选择“新建任务”，进入对话框，在对话框下方的“自定义模型”区域选择**步骤二：接入MaaS模型API**配置的模型，如下图所示：

图 12-4 选择模型

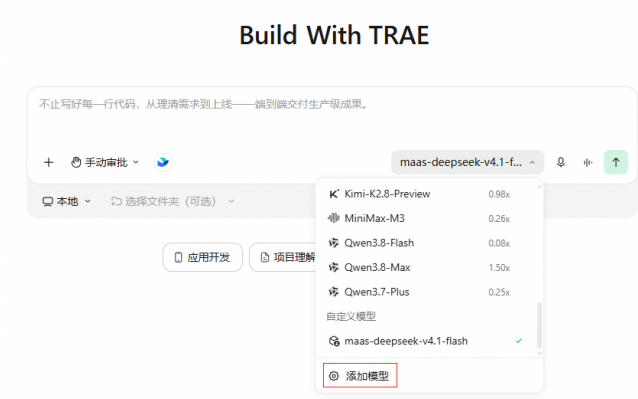


步骤2 在对话框进行对话，体验模型，如下图所示：

图 12-5 对话示例



步骤3 （可选）可在选择模型弹窗下方，单击“添加模型”，添加更多模型。



----结束

常见报错

表 12-2 常见报错排查

报错	原因	解决方法
404	Base URL 路径不完整。当打开了完整URL的开关时，需要完整路径。	补全 /chat/completions（OpenAI）或 /messages（Anthropic）后缀。
401 Unauthorized	API Key 错误或不完整。	检查 Key 是否完整复制、有无空格。
Connection refused	地址不可达或端口错误。	检查防火墙、内网地址是否可访问。
模型无法识别	模型 ID 拼写错误。	核对大小写、多余空格。