智能体平台

用户指南

文档版本	01
发布日期	2026-01-23

华为云计算技术有限公司

商标声明


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

注意

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

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

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 Versatile 使用流程

Versatile包含应用管理、组件库、知识库、提示词开发、配置管理、模型接入调测等功能模块，覆盖体验设计、代码开发、应用运行、资产管理、数据处理、测试发布、运营监控、安全保障八个方面，为企业级用户提供开箱即用的大模型应用开发工具链。依托强大的应用开发工具链，Versatile可支撑客户的个性化应用功能开发需求，智能扩展Agent边界，搭配兼具性能和安全的运行机制，降低开发门槛，使得应用规模化落地，助力各行业企业将大模型应用与实际业务融合，打造企业级专属应用。

Versatile当前提供三种应用开发，用户可根据具体业务和开发场景选择。

表 1-1 三种应用开发场景

类型	单智能体应用	工作流应用	多智能体应用
开发方式	图形化操作，页面点选及文本输入配置，零代码。	画布组件拖拽+低代码开发。	图形化操作，无需编排，配置多智能体中控指令，并引用自智能体，定义分工即可。
面向用户特征	面向业务人员，不会写代码，使用办公工具比较单一的人。	面向技术人员，能够低代码开发工作流，调试各类专业化组件、API。	面向资深业务人员，不写代码，对多业务场景整合。
适用场景	自主规划任务场景，场景泛化性好。	有固定的任务执行流程，高准确率要求。	需要多个智能体协同的复杂用户意图识别分工处理场景。
开发特点	快速简单、低门槛、规划准确率不稳定。	配置操作有门槛，执行成功率高，配置时间长。	相对简单、依赖已开发好的专家智能体。
能力约束	完全依赖模型自身能力，对插件数量，接口参数数量，执行步骤数量等限制较多。出现模型幻觉和不遵守相关规定的案例无法短期内解决。	编排后的流程执行过程中比较死板，智能化程度不高。对于异常场景需要配置相应的流程进行覆盖，会大幅提升流程配置的复杂度。	高度依赖中控模型的智能化水平，涉及的子智能体关联较多，另外当前无可视化调试能力，效果优化工作的调试难度较大。

单智能体应用开发流程

图 1-1 单智能体应用开发流程

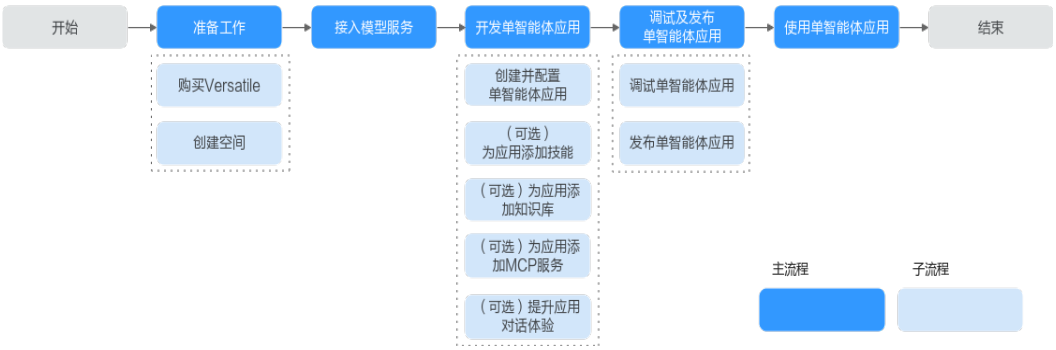


表 1-2 单智能体应用开发流程

流程	子流程	说明	操作指导
准备工作	购买Versatile	购买后才可以正常使用Versatile的功能。	**购买Versatile智能体平台**
	创建空间	在开发过程中，一个开发任务往往需要多个团队成员的协作才能完成。此时可以创建一个团队空间，为团队成员提供一个集中的平台，用于任务分配、进度跟踪、文件共享和即时沟通。通过这种方式，可以显著提升团队的工作效率，确保开发任务的顺利进行和高质量完成。	**管理工作空间**
接入模型服务	-	模型服务为智能体提供了最核心的智能，使智能体能够自主、智能地完成复杂任务。开发单智能体应用前，需要先接入模型服务。	**接入模型服务**
开发单智能体应用	创建并配置单智能体应用	先创建单智能体应用，主要设置应用的名称、描述和图标。再配置单智能体的模型、提示词以及调度模式。	**创建并配置单智能体应用**
	（可选）为应用添加技能	技能包含插件、工作流等，开发者可通过集成插件、设计工作流等方式不断扩展模型的功能范围。	**为应用添加技能**
	（可选）为应用添加知识库	知识库是智能体用于存储、管理和检索领域知识的核心组件，开发者可通过添加知识库为智能体提供精准的信息支持。	**为应用添加知识库**

流程	子流程	说明	操作指导
	（可选）为应用添加记忆	变量用来存储用户的某一行为或偏好，在对话过程中，会自动识别与变量匹配的内容，并将内容存储在变量中。	为应用添加记忆
	（可选）为应用添加MCP服务	开发者可以通过集成MCP服务快速拓展智能体的功能。	为应用添加MCP服务
	（可选）提升应用对话体验	开发者可以通过配置智能体应用的开场白、推荐问题、追问、音色、内容审核能力，提升应用的对话体验。	提升应用对话体验
调试及发布单智能体应用	调试单智能体应用	单智能体应用开发完成后，开发者可以通过调试应用从而精准定位问题并快速调整配置。	调试应用
	发布单智能体应用	单智能体应用调试完成后，需要发布才能被用户使用。	发布应用
使用单智能体应用	-	单智能体应用发布后，可以通过API接口调用。	通过API调用单智能体应用

工作流应用开发流程

图 1-2 工作流应用开发流程

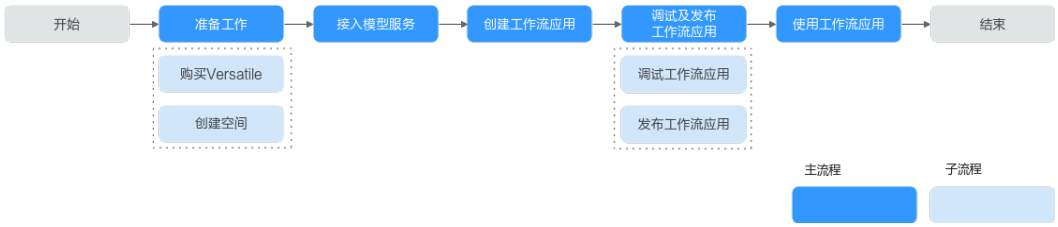


表 1-3 工作流应用开发流程

流程	子流程	说明	操作指导
准备工作	购买Versatile	购买后才可以正常使用Versatile的功能。	购买Versatile智能体平台

流程	子流程	说明	操作指导
	创建空间	在开发过程中，一个开发任务往往需要多个团队成员的协作才能完成。此时可以创建一个团队空间，为团队成员提供一个集中的平台，用于任务分配、进度跟踪、文件共享和即时沟通。通过这种方式，可以显著提升团队的工作效率，确保开发任务的顺利进行和高质量完成。	**管理工作空间**
接入模型服务	-	模型服务是Agent工作流的核心组件，提供了问答、理解规划和决策等能力，这是传统工作流与Agent工作流的主要区别。开发工作流应用前，需要先接入模型服务。	**接入模型服务**
创建工作流应用	-	创建工作流，包含全局配置、编排、选择节点、参数配置，对节点调试，完成功能连通。	**创建工作流**
调试及发布工作流应用	调试工作流应用	开发者可以在工作流创建完成后，直接与工作流进行交互，实时观察其执行过程和响应效果，并根据需要对配置进行优化和调整。平台提供的全链路调试功能，允许开发者查看每条用户请求从输入到响应的完整流程，包括意图识别、知识检索等详细信息，从而能够高效定位问题并快速调整配置。	**调试工作流**
	发布工作流应用	工作流应用调试完成后，需要发布才能被用户使用。	**发布工作流**
使用工作流应用	-	工作流应用发布后，可以在单智能体应用中使用，在多智能体应用中使用，还可以通过API接口调用。	• **通过API调用工作流** • **在单智能体应用中使用工作流** • **在多智能体应用中使用工作流**

多智能体应用开发流程

图 1-3 多智能体应用开发流程

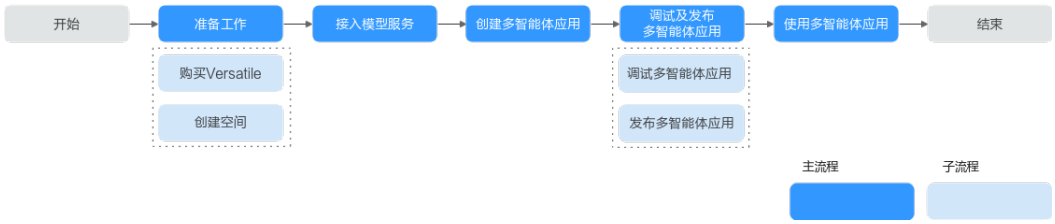


表 1-4 多智能体应用开发流程

流程	子流程	说明	操作指导
准备工作	购买Versatile	购买后才可以正常使用Versatile的功能。	**购买Versatile智能体平台**
	创建空间	在开发过程中，一个开发任务往往需要多个团队成员的协作才能完成。此时可以创建一个团队空间，为团队成员提供一个集中的平台，用于任务分配、进度跟踪、文件共享和即时沟通。通过这种方式，可以显著提升团队的工作效率，确保开发任务的顺利进行和高质量完成。	**管理工作空间**
接入模型服务	-	模型服务为智能体提供了最核心的智能，使智能体能够自主、智能地完成复杂任务。开发多智能体应用前，需要先接入模型服务。	**接入模型服务**
创建多智能体应用	-	多智能体应用可以灵活应用各种工作流来完成用户任务，支持根据用户意图在不同的工作流之间跳转。	**创建多智能体应用**
调试与发布多智能体应用	调试多智能体应用	开发者可以在多智能体应用创建完成后，直接与多智能体进行对话，实时观察其执行过程和响应效果，并根据需要对配置进行优化和调整。	**调试多智能体应用**
	发布多智能体应用	多智能体应用调试完成后，需要发布才能被用户使用。	**发布多智能体应用为API**
使用多智能体应用	-	多智能体应用发布后，可以通过API接口调用。	**通过API调用多智能体应用**

2 购买 Versatile 智能体平台

如果未购买Versatile智能体平台，用户进入Versatile智能体平台，**仅可查看部分功能**。购买Versatile智能体平台后，才可以正常使用Versatile智能体平台。首次购买Versatile智能体平台前，需要先**申请加入Versatile智能体平台白名单**。

前提条件

- 华为账号：华为账号是您访问华为各网站的统一“身份标识”，您只需注册华为账号，即可访问所有华为服务。在登录Versatile智能体平台之前，请先**注册华为账号并开通华为云**、**实名认证**。

  如果您已开通华为云并进行实名认证，请忽略此步骤。
- IAM用户：由**管理员**在IAM中创建的用户，是云服务的使用人员，根据账号授予的权限使用资源。**账号与IAM用户**可以类比为父子关系，IAM用户是由**管理员**在IAM中创建的用户，IAM用户登录后可根据权限使用云服务。管理员创建IAM用户的方法请参考**创建IAM用户**。

申请加入 Versatile 白名单

如果在“概览”、“平台管理 > 我的资源”页面，显示“产品升级中，请通过工单开通Versatile，造成不便敬请谅解。”，需要用户通过工单将账号申请加入白名单中。账号加入白名单后，用户才能购买Versatile智能体平台。

步骤1 进入**新建工单**页面，选择问题所属产品为“应用平台 AppStage”，选择问题类型为“产品咨询”，在“新建工单”区域下单击“去新建”。

步骤2 填写工单相关信息。

其中，“问题描述”项请填写申请原因“申请加入Versatile白名单”。

步骤3 填写完毕后，勾选协议并单击“提交”。

步骤4 提交成功后，预计7个工作日以内完成审核，请您耐心等待。系统后台审批通过之后，会通过邮件或短消息（在工单中填写了邮箱地址或手机号码）的方式通知您。

----结束

购买 Versatile 智能体平台

步骤1 登录**Versatile智能体平台**，进入“平台管理 > 我的资源”页面。

未购买Versatile智能体平台，会自动进入“平台管理 > 我的资源”页面

步骤2 在“我的资源”页面，单击“购买资源”。

步骤3 在购买资源页面，选择套餐规格、购买时长（根据需要勾选自动续费），单击“立即购买”。

步骤4 在“订单确认”页面，确认购买的资源信息，勾选“我已阅读并同意《Versatile服务声明》”，单击“去支付”。

步骤5 在“支付方式”区域，选择支付方式，单击“确认付款”。

支付完成后，系统会提示“订单支付成功”。

----结束

3 Versatile 智能体平台首页介绍

前提条件

登录用户为空间所有者、空间管理员、开发工程师，详细信息请参考**管理团队空间成员**。

登录 Versatile 智能体平台

使用已实名认证的华为账号或IAM用户登录**Versatile智能体平台**。

Versatile 概览页介绍

用户首次登录进入Versatile智能体平台，显示“概览”页内容。

Versatile概览如**图3-1**所示，各区域的功能说明请参考**表3-1**。

图 3-1 Versatile 概览页

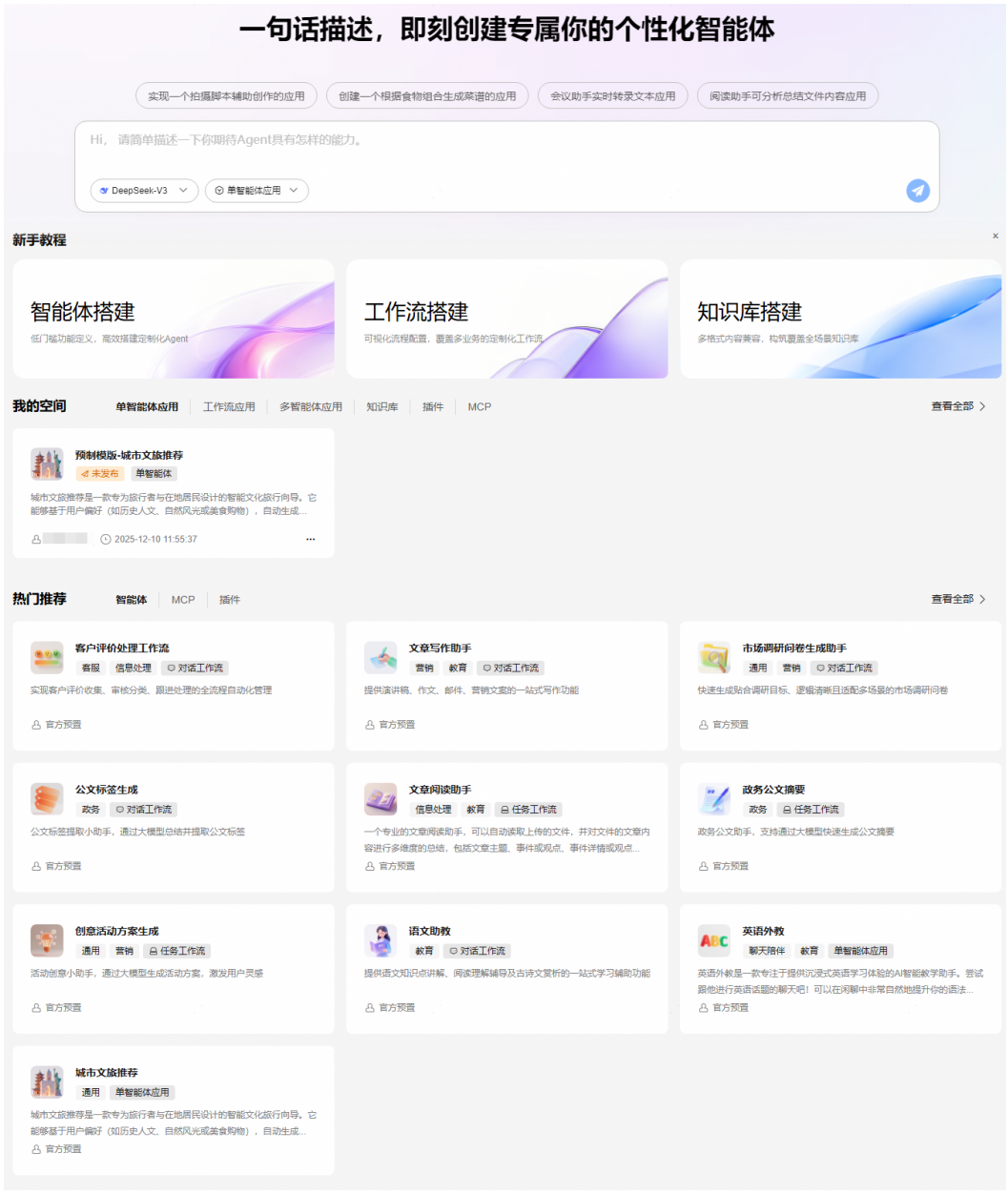


表 3-1 概览页说明

区域	说明
一句话描述，即刻创建专属你的个性化智能体	选择输入框上方的任务，或在输入框中输入任务，在输入框左下角选择模型、选择“单智能体应用”或“工作流应用”，单击  ，快速创建单智能体应用或工作流应用，具体请参考**AI辅助创建单智能体应用**或**AI辅助创建工作流**。 **购买Versatile智能体平台**后，才显示输入框。

区域		说明
新手教程	Agent搭建	低门槛功能定义，高效搭建定制化Agent。 单击“Agent搭建”，跳转至文档“单智能体应用介绍”。
	工作流搭建	可视化流程配置，覆盖多业务的定制化工作流。 单击“工作流搭建”，跳转至文档“工作流介绍”。
	知识库搭建	可视化流程配置，覆盖多业务的定制化工作流。 单击“知识库搭建”，跳转至文档“知识库介绍”。
我的空间	单智能体应用	显示用户最近创建的单智能体应用卡片信息。 ● 在右侧单击“查看全部”，进入“开发中心 > 应用管理 > 单智能体应用”页面。 ● 单击单智能体应用卡片，进入单智能体应用编辑页面，可以对单智能体应用进行编辑等操作，具体请参考**创建单智能体应用**。 ● 在单智能体应用卡片上，单击 ···，可以管理单智能体应用卡片，具体操作请参考**管理应用**。 首次**购买Versatile智能体平台**后，才显示“我的空间”。**购买Versatile智能体平台**后，退订，未再次购买时，不可操作已有的单智能体应用卡片。
	工作流应用	显示用户最近创建的工作流应用卡片信息。 ● 在右侧单击“查看全部”，进入“开发中心 > 应用管理 > 工作流应用”页面。 ● 单击工作流应用卡片，进入工作流应用编辑页面，可以对工作流应用进行编辑等操作，具体请参考**搭建工作流**。 ● 在工作流应用卡片上，单击 ···，可以管理工作流应用卡片，具体操作请参考**管理工作流**。 首次**购买Versatile智能体平台**后，才显示“我的空间”。**购买Versatile智能体平台**后，退订，未再次购买时，不可操作已有的工作流应用卡片。

区域		说明
	多智能体应用	显示用户最近创建的多智能体应用卡片信息。 • 在右侧单击“查看全部”，进入“开发中心 > 应用管理 > 多智能体应用”页面。 • 单击多智能体应用卡片，进入多智能体应用编辑页面，可以对多智能体应用进行编辑等操作，具体请参考**创建多智能体应用**。 • 在多智能体应用卡片上，单击 ···，可以管理多智能体应用卡片，具体操作请参考**相关操作**。 首次**购买Versatile智能体平台**后，才显示“我的空间”。**购买Versatile智能体平台**后，退订，未再次购买时，不可操作已有的多智能体应用卡片。
	知识库	显示用户最近创建的知识库卡片信息。 • 在右侧单击“查看全部”，进入“开发中心 > 知识库”页面。 • 单击知识库卡片，进入“知识库详情”页面，可以对知识库进行编辑等操作，具体请参考**创建本地知识库**、**接入第三方知识库**。 • 在知识库卡片上，单击 ···，可以管理知识库卡片，具体操作请参考**创建本地知识库**、**接入第三方知识库**。 首次**购买Versatile智能体平台**后，才显示“我的空间”。**购买Versatile智能体平台**后，退订，未再次购买时，不可操作已有的知识库卡片。
	插件	显示用户最近创建的插件卡片信息。 • 在右侧单击“查看全部”，进入“开发中心 > 组件库 > 我的插件”页面。 • 单击插件卡片，进入“插件库详情”页面，可以对插件进行编辑等操作，具体请参考**创建插件**。 • 在插件卡片上，单击 ···，可以管理插件卡片，具体操作请参考**管理插件**。 首次**购买Versatile智能体平台**后，才显示“我的空间”。**购买Versatile智能体平台**后，退订，未再次购买时，不可操作已有的插件卡片。

区域		说明
	MCP	显示用户最近部署的MCP卡片信息。 • 在右侧单击“查看全部”，进入“开发中心 > 组件库 > 我的MCP”页面。 • 单击MCP卡片，进入MCP服务详细信息页面，可以编辑MCP服务，具体操作请参考**创建MCP服务**。 • 在MCP卡片上，单击“删除”，可以删除MCP卡片。 首次**购买Versatile智能体平台**后，才显示“我的空间”。**购买Versatile智能体平台**后，退订，未再次购买时，不可操作已有的MCP卡片。
热门推荐	智能体	显示平台预置的智能体卡片信息。 • 在右侧单击“查看全部”，进入“资产中心 > 应用广场”页面。 • 单击智能体卡片，可以直接与智能体进行对话，或单击“复制到当前空间”，进入单智能体应用编辑页面，并将平台预置的智能体复制到“开发中心 > 应用管理 > 单智能体应用”页面，名称为该智能体名称_数字后缀。 未**购买Versatile智能体平台**，不可操作预置的智能体卡片。
	MCP	显示平台预置的MCP卡片信息。 • 在右侧单击“查看全部”，进入“资产中心 > MCP广场”页面。 • 单击MCP卡片，可以查看MCP详细信息。 • 在MCP卡片上，单击“安装”，安装MCP服务。 未**购买Versatile智能体平台**，不可安装预置的MCP服务。
	插件	显示平台预置的插件卡片信息。 • 在右侧单击“查看全部”，进入“资产中心 > 插件广场”页面。 • 单击插件卡片，进入“插件详情”页面。 • 在插件卡片上，单击“配置鉴权”或“移除鉴权”，为插件设置或删除鉴权信息。 未**购买Versatile智能体平台**，不可为预置的插件设置或删除鉴权信息。

4 管理工作空间

4.1 工作空间介绍

为方便多人协作和资源共享，Versatile引入了工作空间的概念。工作空间为用户提供了灵活的资源管理和团队协作功能。

工作空间分为个人空间和团队空间。

个人空间：每个Versatile用户默认具有一个个人空间，默认的个人空间不可删除、不能编辑、不能共享，仅用于管理个人开发及资源管理。

团队空间：团队空间旨在为用户提供灵活、高效的资产管理与协作方式。Versatile支持用户根据业务需求或团队结构，自定义创建独立的团队空间。通过这种方式，用户可以更好地组织和管理资源，提高团队的协作效率。

团队空间在资产层面完全隔离，确保资产的安全性和操作的独立性，有效避免交叉干扰或权限错配带来的风险。用户可以结合实际使用场景，如不同的项目管理、部门运营或特定的研发需求，划分出多个团队空间，实现资产的精细化管理与有序调配，帮助用户高效地规划和分配任务，使团队协作更加高效。创建团队具体操作请参考**创建并管理团队空间**。

此外，Versatile配备了完善的空间角色权限体系。通过灵活的权限设置，每个用户能够在其对应的权限范围内安全高效地操作Versatile功能，从而最大限度保障数据的安全性与工作效率。

费用说明

不同套餐包允许创建的**团队空间数量**、添加的**成员数量**不同，具体可参考**计费模式**。

团队空间人员角色与权限

Versatile团队空间人员角色如下，每种角色对团队空间的操作权限不同，具体操作权限请参考**表4-1**。

如何管理团队空间成员，请参考**管理团队空间成员**。

说明

“√”表示支持，“x”表示暂不支持。

表 4-1 团队空间人员角色与权限

模块	权限	空间所有者	空间管理员	开发工程师	运维工程师
空间管理	新增空间	√	√	√	√
	修改空间名称、简介	√	√	×	×
	查看空间	√	√	√	√
	添加成员	√	√	×	×
	修改角色	√	√ 不能改自己的角色 不能改空间所有者的角色	×	×
	转让所有者	√	×	×	×
	移除成员	√	√ 不能移除空间所有者	×	×
	离开空间	×	√	√	√
	删除空间	√	×	×	×
开发中心	创建	√	√	√	×
	查看&调试&试运行	√	√	√	
	复制	√	√	√	
	修改	√	√	√	
	发布	√	√	√	
	导入&导出	√	√	√	
	删除	√	√	√	
运营运维	观测所有操作	√	√	×	√
模型中心	模型服务所有操作	√	√	√	√
	路由策略所有操作	√	√	√	√
	模型调测所有操作	√	√	√	√
资产中心	资产查看	√	√	√	√

模块	权限	空间所有者	空间管理员	开发工程师	运维工程师
	模板-在线调试	√	√	√	√
	模板复制	√	√	√	×
	模板共享	√	√	√	×
	MCP-安装	√	√	√	×
	插件-鉴权配置等	√	√	√	×
	插件共享	√	√	√	×
概览	空间下的资产查看	√	√	√	×
	AI一句话创建	√	√	√	
我的资源	查看	√	√	√	√

4.2 创建并管理团队空间

在开发过程中，一个开发任务往往需要多个团队成员的协作才能完成。此时可以创建一个团队空间，为团队成员提供一个集中的平台，用于任务分配、进度跟踪、文件共享和即时沟通。通过这种方式，可以显著提升团队的工作效率，确保开发任务的顺利进行和高质量完成。

费用说明

不同套餐包允许创建的**团队空间数量**不同，具体可参考**计费模式**。

前提条件

- 已**购买Versatile智能体平台**。
- 已**实名认证**的华为账号或IAM用户。

创建团队空间

步骤1 登录**Versatile智能体平台**。

步骤2 在左侧导航，选择“个人空间 > 创建团队空间”，如**图4-1**所示。

如果已选择团队空间，界面显示为实际的团队空间名称，而非“个人空间”。

图 4-1 创建团队空间

步骤3 “在创建空间”页面，配置空间信息，创建空间参数说明请参考**创建团队空间**，单击“确定”。

已创建的团队空间显示在“个人空间”的下拉列表中。

如果已有团队空间，界面显示为实际的团队空间名称，而非“个人空间”。

表 4-2 创建空间

参数	说明	示例
空间名称	团队空间的名称。由1~50个字符组成，包含中文、数字、字母、中划线、下划线、括号、感叹号。	单智能体应用
空间描述	选填项。 团队空间的描述信息。由0~1000个字符组成。	该空间用于开发单智能体应用。
空间图像	系统默认团队空间头像，用户也可以自定义图像。 1. 鼠标移动至系统默认图像上，单击鼠标左键。 2. 在虚线框中，单击鼠标左键，上传已准备好的团队空间图像。 支持jpg、jpeg、png、gif格式图片，且不大于200KB。	系统默认图像

----结束

管理团队空间

步骤1 登录**Versatile智能体平台**。

步骤2 在左侧导航，单击“个人空间”，选择已创建的团队空间，如**图4-2**所示。

如果已选择团队空间，界面显示为实际的团队空间名称，而非“个人空间”。

选择团队空间后，左侧导航栏最下方显示“平台管理 > 团队空间管理”。

图 4-2 选择团队空间

步骤3 在左侧导航栏，选择“平台管理 > 团队空间管理”，进入“团队空间管理”页面。

步骤4 在“团队空间管理”页面，支持对团队空间的其他操作请参考**表4-3**。

表 4-3 管理团队空间

操作	说明
编辑团队空间基础信息	在“基础信息”右侧，单击，编辑团队空间名称、空间描述、空间图像，完成后，单击“确定”。 **空间所有者**、**空间管理员**支持编辑空间基础信息。
删除团队空间	**警告** 删除空间后，对应的资源也一并删除，不可恢复。 单击“删除空间”，在弹框中，输入“DELETE”，单击“确定”。 **空间管理员**支持删除团队空间。
转让团队空间	单击“转让空间”，在“转移空间”界面，选择转让的成员，单击“确定”。 **空间所有者**支持转让团队空间，转让后，从**空间所有者**变为**空间管理员**。
退出团队空间	单击“退出空间”。 从该空间中退出。退出空间后，不可查看该空间资源。 **空间管理员**、**开发工程师**、**运维工程师**支持退出团队空间。

----结束

4.3 管理团队空间成员

用户创建团队空间后，为了使团队空间更加高效地运作，可以为团队空间添加成员。

费用说明

不同套餐包允许添加的**成员数量**不同，具体可参考**计费模式**。

前提条件

- 已**购买Versatile智能体平台**。
- 已**创建团队空间**。
- 已在该租户下**创建IAM用户**。
- 登录用户为团队空间的空间所有者或空间管理员，该用户具有Security Administrator权限，给IAM用户授权请参考**给IAM用户授权**。

  创建团队空间的用户默认为空间所有者。

为空间添加用户

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏，选择“平台管理 > 团队空间管理”，进入“团队空间管理”页面。

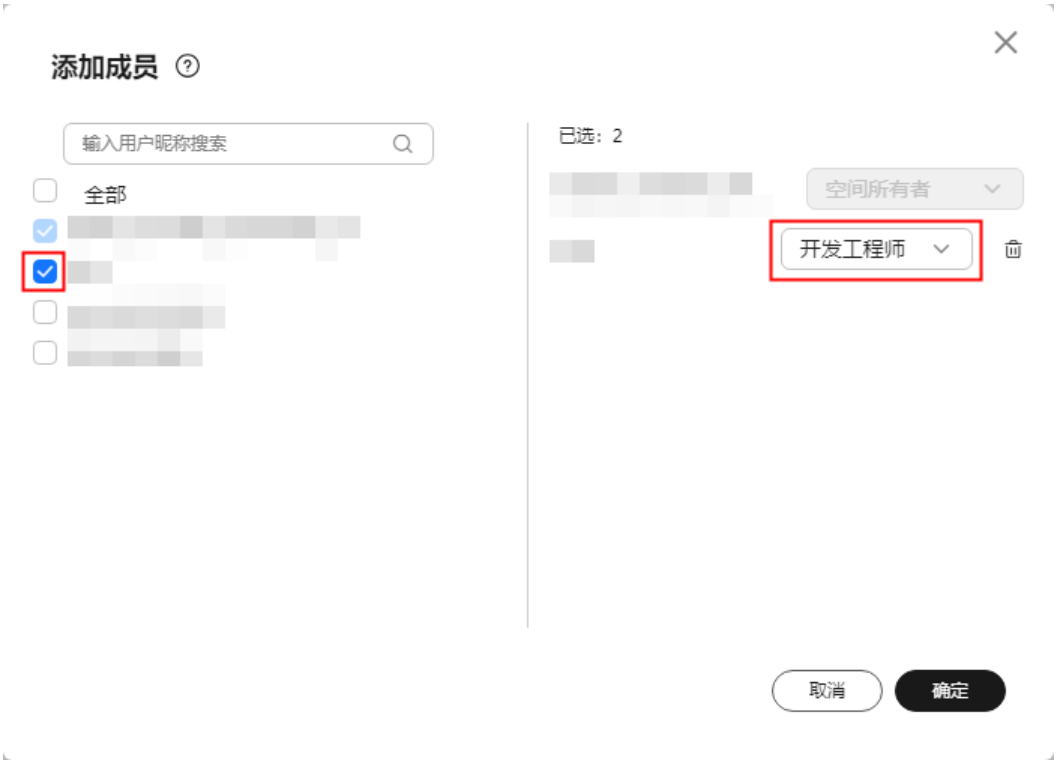
步骤3 在“团队空间管理”页面，单击“添加成员”。

步骤4 在“添加成员”页面，搜索用户名称，勾选待添加用户左侧的复选框，设置成员的角色，单击“确定”。

成员角色介绍请参考**团队空间人员角色与权限**。创建团队空间的用户默认为空间所有者。

添加的成员显示在成员列表中。

图 4-3 添加成员



----结束

相关操作

在“团队空间管理”的“成员管理”区域，支持的其他操作请参考表4-4。

表 4-4 成员管理相关操作

参数	说明
切换用户角色	在待切换角色的用户对应的“角色”列下，单击成员角色，重新选择角色。 **空间所有者**不支持切换角色。
删除成员	• 单个删除：在待删除的用户对应的“操作”列下，单击“删除”。 • 批量删除：勾选待删除用户左侧的复选框，单击“删除”。 **空间所有者**不支持删除。

5 了解并使用资产中心资源

5.1 Versatile 资产中心介绍

进入Versatile后，您可以通过左侧导航栏选择“资产中心”，进入资产中心页面。资产中心为您提供了一系列丰富的资源和工具，包括应用广场、MCP广场、插件广场和提示词广场。

应用广场

应用广场提供多种平台精选的智能体和工作流，还汇集了团队共享的多智能体和工作流。预置的资源覆盖通用、政务、信息处理、营销、客服、陪伴聊天、教育等多个行业领域。这些应用基于强大的大模型技术，即开即用，能够快速满足用户在智能工单总结、政务公文摘要、医疗病历生成、金融话术推荐等具体业务场景中的需求，显著提升工作效率与质量。智能体模板和工作流工具能够帮助快速启动项目，高效管理开发过程。

图 5-1 应用广场

MCP 广场

平台精选了丰富的MCP资源，例如MCP车票查询工具、内容抓取转换器、可视化图表MCPServer、高德地图等。这些资源为智能体和工作流的开发提供了强有力的支撑，

显著增强了调用能力。通过集成这些服务，开发者可以更高效、便捷地完成功能实现，提升应用的部署效率和响应速度。平台还提供了第三方的MCP服务，进一步丰富了开发工具。

平台预置的每个MCP资源卡片上清晰展示服务的评分、阅读量和安装量，帮助您快速了解社区认可度。单击卡片进入详情页面，查看服务描述、功能等全面信息。您还可以通过“我要评分”按钮为服务打分，分享使用体验。“热门使用推荐”区域展示当前最热门的MCP服务，为您的选择提供参考。

图 5-2 MCP 广场

插件广场

插件广场预置了平台精选的各类插件工具，同时汇集了团队共享的插件。这些插件功能多样，广泛覆盖多个实用领域，能够显著提升智能体的应用能力。

通过插件广场，用户可以轻松访问和集成各种功能强大的工具，从而扩展智能体和工作流的功能，提高工作效率和用户体验。

图 5-3 插件广场

提示词广场

提示词广场里展示了平台预置的多种提示语模板，您可以基于它们创建新的提示语。在智能体或工作流中您可以直接选择预置的提示语进行使用，快速提高工作效率。

图 5-4 提示词广场

5.2 使用资产中心的应用资源

应用广场提供了多种平台精选的智能体和工作流，覆盖多个行业领域，能够快速满足用户在具体业务场景中的需求，提升工作效率和质量。同时，支持共享当前空间下的多智能体和工作流，方便团队协作和资源复用。

前提条件

已**购买Versatile智能体平台**。

约束与限制

- 仅**Versatile企业版**支持共享多智能体和工作流应用，同时可以使用他人共享的应用。**Versatile基础版（限时免费）**不支持该能力。
- 同一个多智能体或工作流应用只能被共享一次，但可以共享该多智能体或工作流应用的多个版本。

使用平台精选的智能体应用

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏中选择“资产中心 > 应用广场”。

步骤3 在“平台精选”页签中，单击应用卡片列表顶部的下拉框，在下拉列表中选择“智能体”。

图 5-5 选择智能体

步骤4 可以通过分类筛选（通用、政务、信息处理、营销、客服、陪伴聊天、教育）或搜索功能找到目标应用。单击目标应用（例如“创意活动方案生成”）。

步骤5 可以直接使用智能体应用生成需要的内容。

可以在应用页面的右上角，单击“复制到当前空间”按钮。这样，该智能体应用就成为您自己的，您可以随时在自己的项目中使用和编辑。

复制预置的智能体应用后，您可以在“开发中心 > 应用管理 > 单智能体应用”中找到该应用。如果您需要对智能体进行更具体的操作，请参考**开发单智能体应用**。

----结束

使用平台精选的工作流应用

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏中选择“资产中心 > 应用广场”。

步骤3 在“平台精选”页签中，单击应用卡片列表顶部的下拉框，在下拉列表中选择“工作流”。

图 5-6 选择工作流

步骤4 可以通过分类筛选（通用、政务、信息处理、营销、客服、陪伴聊天、教育）或搜索功能找到目标工作流。单击目标工作流（例如“AI测试工作流”）。

步骤5 可以直接使用工作流生成需要的内容。

可以在应用页面的右上角，单击“复制到当前空间”按钮。这样，该工作流就成为您自己的，可以随时在自己的项目中使用和编辑。

复制预置的工作流后，您可以在“开发中心 > 应用管理 > 工作流应用”中找到该工作流。如果您需要对工作流进行更具体的操作，请参考**开发工作流应用**。

----结束

查看他人共享的应用

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏中选择“资产中心 > 应用广场”。

步骤3 在“团队共享”页签中，单击“他人共享”页签。

在“他人共享”页签中，可以查看其他人共享的多智能体或工作流应用。

步骤4 可以直接在对话框中输入内容，使用共享的多智能体或工作流应用来生成需要的内容。例如，您可以使用共享的多智能体应用来生成报告、分析数据或执行其他任务。

也可以直接在自己的应用中引用共享的多智能体或工作流应用。使用共享应用的详细信息，请参考**添加工作流**、**工作流**和**创建多智能体应用**。

----结束

共享应用

前提条件

创建并发布多智能体或工作流应用，创建和发布的详细信息，请参考**搭建工作流**、**创建多智能体应用**、**发布工作流**和**发布多智能体应用为API**。

操作步骤

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏中选择“资产中心 > 应用广场”。

步骤3 在“团队共享”页签中，单击“共享应用”按钮。

步骤4 在“共享已发布的应用”的弹框中，执行以下操作设置共享应用的信息。

1. 在“选择共享的应用”的下拉列表中选择需要共享多智能体或工作流，然后选择具体要共享的应用名称。
2. 在“选择共享版本”的下拉列表中选择共享的多智能体或工作流的具体版本。
3. 设置共享模式。当前只支持“可使用”模式。

   可使用：仅能在智能体和工作流中引用，不能复制和修改配置。
4. 设置共享范围。
   - **全部空间可见**：勾选此选项后，共享的应用将在当前租户下的所有空间中共享。
   - **部分空间可见**：选择具体的空间，仅在这些空间中共享应用。也可以勾选团队列表上方的“全部”复选框，这样应用将在当前租户下的所有空间中共享。

   说明

   当共享范围设置为“全部空间可见”时，新增的空间将自动包含该共享的应用。如果选择“部分空间可见”并选中“全部”，则不支持此功能。

步骤5 单击“确定”完成多智能体或工作流应用的共享。共享后的多智能体或工作流可以在“团队共享”页签中“我的共享”中查看。

----结束

更多操作

应用共享完成之后，在“团队共享”页签下，可以通过分类筛选功能来查找共享的应用。还可以执行如**表5-1**的操作。

注意

只有共享者有权限执行以下操作。

表 5-1 相关操作

操作	说明
取消共享	1. 找到需要取消共享的应用，单击该应用卡片进入详情页面。 2. 单击右上角的“取消共享”，在弹出的对话框中单击“确定”，可以取消应用共享。 单击卡片右下角 ··· 按钮，单击“取消共享”，也可以取消共享。 **说明** 如果该应用已被引用，取消共享后引用将自动取消，可能会导致工作流或智能体应用无法运行，且该操作不可撤回，请谨慎操作。
查看引用	在应用详情界面，单击引用列表图标  ，可以查看该插件被哪些智能体和工作流应用引用。
查看共享版本	在应用详情页面的右上角，单击发布历史图标  ，可以查看当前应用的共享版本记录。此页面按发布时间倒序显示历史记录，包括版本名称、应用ID和共享人等信息。
更新共享	1. 在应用详情页面的右上角，单击“更新共享”，在展开的“修改共享内容”弹框中修改应用共享的版本和范围。 2. 修改完成后，单击“确定”完成共享应用的信息更新。

操作	说明
取消共享版本	在应用详情页面的右上角，单击共享版本历史图标，选择需要取消的共享版本，单击“取消共享此版本”，即可取消该应用该版本的共享。 **图 5-7** 取消共享版本 

5.3 使用资产中心的 MCP 资源

平台精选和第三方的各类MCP服务，功能多样，广泛覆盖多个实用领域，能够有效拓展智能体和工作流应用的能力。用户可以根据实际需求，灵活选择和使用这些MCP服务，以满足自己的业务需要。

前提条件

已**购买Versatile智能体平台**。

使用平台精选的 MCP

平台精选的MCP服务需要安装之后才能被智能体或工作流使用。

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏中选择“资产中心 > MCP广场”。

步骤3 在“平台精选”页签中，选中目标MCP服务并单击“安装”，如**图5-8**所示。或单击目标MCP服务进入详情页后，单击右上角的“安装”，如**图5-9**所示。

图 5-8 安装 MCP

图 5-9 详情页安装 MCP

步骤4 在“创建MCP服务”弹框中单击“安装”，将提交您的安装请求，开始安装MCP服务。安装成功后，可以在“开发中心 > 组件库 > 我的MCP”中查看已经安装的MCP服务。

在智能体或工作流中使用MCP服务请参考**为应用添加MCP服务**或**MCP服务**。

----结束

使用第三方 MCP 服务

第三方的MCP服务需要安装之后才能被智能体或工作流使用。

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏中选择“资产中心 > MCP广场”。

步骤3 在“第三方”页签中的MCP卡片列表的顶部，单击下拉框，在下拉列表中可以选择需要订阅的MCP服务的类型。目前仅支持ROMA Connect。

也可以通过输入服务名称来搜索所需的MCP服务。注意，搜索ROMA Connect的MCP服务时，最少需要输入两个字符，否则会提示“访问 ROMA Connect 接口错误，请检查重试！”。

图 5-10 第三方 MCP

步骤4 在需要安装的卡片右下角，单击“订阅”。

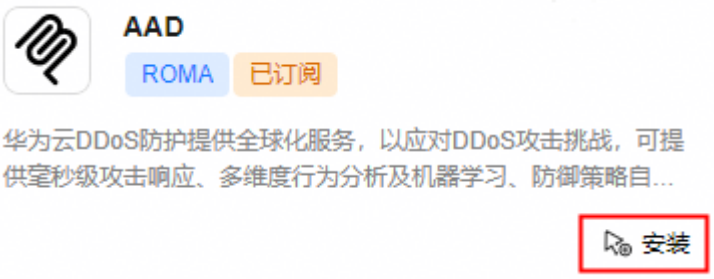
步骤5 在“订阅指引”的弹框中，单击“去订阅”。跳转到第三方MCP服务的开通或订阅页面，完成服务的开通和订阅。

步骤6 完成开通或订阅后，返回智能体平台单击刷新 。已订阅或开通的MCP服务卡片右下角的“订阅”按钮将变为“安装”。

步骤7 选中需要安装的MCP服务并单击“安装”，将提交您的安装请求，开始安装MCP服务。如图5-11所示。

在同一空间内，不能安装相同名称的MCP服务。

图 5-11 安装 MCP



步骤8 单击“前往我的MCP”，可以在“开发中心 > 组件库 > 我的MCP”中查看已经安装的MCP服务。

在智能体或工作流中使用第三方MCP服务请参考**为应用添加MCP服务**或**MCP服务**。

图 5-12 前往我的 MCP

----结束

5.4 使用资产中心的插件资源

插件广场汇聚了平台精选以及团队共享的各类插件工具。这些插件功能多元，广泛覆盖多个实用领域，能够有效拓展智能体和工作流的应用能力。用户可依据实际需求，灵活运用各类插件以满足自己的业务需要。

前提条件

已**购买Versatile智能体平台**。

约束与限制

- 仅**Versatile企业版**支持共享插件，同时可以使用他人共享的插件。**Versatile基础版（限时免费）**不支持该能力。
- 同一个插件只能被共享一次，但可以共享该插件的多个版本。

使用平台精选的插件

平台提供的预置插件根据鉴权状态分为三类：未鉴权、无需鉴权和已鉴权的插件。其中，未鉴权的插件只有在完成鉴权后，才能被智能体和工作流使用。

平台提供的插件分为免费和付费两种:

免费插件：免费插件无需购买，无需鉴权的插件可以直接使用，未鉴权的插件设置鉴权后即可使用。设置鉴权请参考**步骤3**。

付费插件：付费插件需要先购买并设置鉴权后才能使用。单击“获取鉴权信息”可跳转至购买和获取API Key的页面。设置鉴权请参考**步骤3**。

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

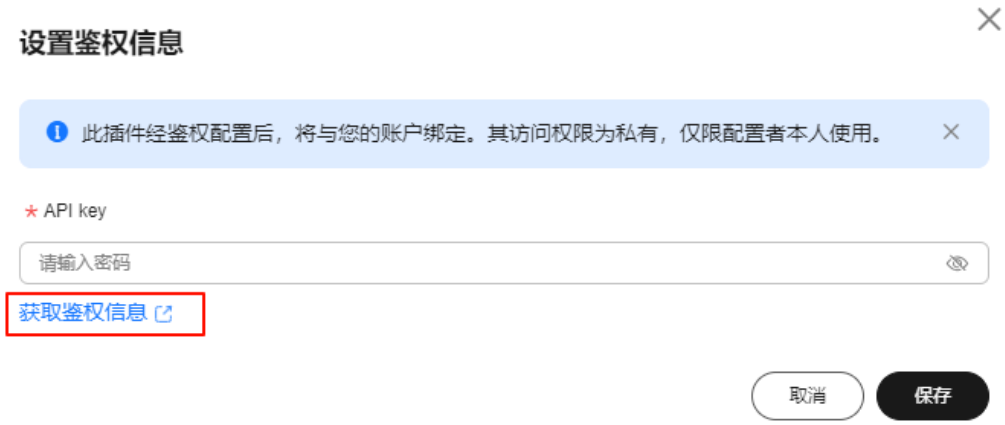
步骤2 在左侧导航栏中选择“资产中心 > 插件广场”。

步骤3 在“平台精选”页签中，在需要添加鉴权的插件卡片右下角单击“配置鉴权”。

步骤4 在弹出的“设置鉴权信息”对话框中，输入该插件的API key。

如果无该插件的API Key，可单击设置鉴权信息界面下方的“获取鉴权信息”，如**图5-13**所示，并按照指导注册获取。

图 5-13 获取鉴权信息



步骤5 单击“保存”完成插件的鉴权。在智能体或工作流中使用预置插件请参考**添加插件**或**插件**。

说明

- 完成鉴权配置后，卡片的右下角会显示“移除鉴权”按钮。通过该按钮可以移除插件的鉴权。移除鉴权后，智能体和工作流中引用的插件将受到影响，请谨慎操作。
- 在插件详情页面，单击右上角的“引用插件”按钮，可以查看当前插件被哪些智能体和工作流引用。

----结束

查看他人共享的插件

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏中选择“资产中心 > 插件广场”。

步骤3 在“团队共享”页签中，单击“他人共享”页签。

步骤4 在“他人共享”页签中，可以查看其他人共享的插件。

在“插件详情”的“工具信息”页签中，单击“操作”列下的“查看详情”以查看该工具的详细信息。

智能体或工作流应用中使用他人共享的插件请参考**添加插件**和**插件**。

----结束

共享插件

前提条件

创建并发布插件，创建和发布插件的详细信息，请参考**创建插件**和**发布插件**。

操作步骤

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏中选择“资产中心 > 插件广场”。

步骤3 在“团队共享”页签中，单击“共享插件”按钮。

步骤4 在展开的“共享已发布的插件”弹框中，执行以下操作设置共享插件的信息。

1. 在“选择共享的插件”的下拉列表中选择需要共享的插件名称。
2. 在“选择共享版本”的下拉列表中选择共享插件的具体版本。
3. 设置共享模式。当前只支持“可使用”模式。
   可使用：仅能在智能体和工作流中引用，不能对该插件进行复制和修改配置。
4. 设置共享范围。
   - **全部空间可见**：勾选此选项后，共享的插件将在当前租户下的所有空间中共享。
   - **部分空间可见**：选择具体的空间，仅在这些空间中共享插件。也可以勾选团队列表上方的“全部”复选框，这样插件将在当前租户下的所有空间中共享。

   说明

   当共享范围设置为“全部空间可见”时，新增的空间将自动包含该共享的插件。如果选择“部分空间可见”并选中“全部”，则不支持此功能。

步骤5 单击“确定”完成插件共享。共享后的插件可以在“团队共享”页签中“我的共享”中查看。

在“插件详情”的“工具信息”页签中，单击“操作”列下的“查看详情”以查看该工具的详细信息。

智能体或工作流应用中使用他人共享的插件请参考**添加插件**和**插件**。

----结束

更多操作

插件共享完成之后，在“团队共享”页签下，可以通过搜索关键字来查找共享的插件。还可以执行如**表5-2**的操作。

注意

只有共享者有权限执行以下操作。

表 5-2 相关操作

操作	说明
取消共享	1. 找到需要取消共享的插件，单击插件卡片进入详情页面。 2. 单击右上角的“取消共享”，在弹出的对话框中单击“确定”，可以取消插件共享。 单击卡片右下角 ··· 按钮，单击“取消共享”，也可以取消共享。 **说明** 如果该插件已被引用，取消共享后引用将自动取消，可能会导致工作流或应用无法运行，且该操作不可撤回，请谨慎操作。

操作	说明
查看引用	在插件详情界面，单击引用插件列表图标，可以查看该插件被哪些单智能和工作流体引用。
查看共享版本	在插件详情页面的右上角，单击共享版本历史图标，可以查看当前共享插件的共享版本记录。此页面按发布时间倒序显示历史记录，包括版本名称、插件ID和共享人等信息。
取消共享版本	在插件详情页面的右上角，单击共享版本历史图标，选择需要取消的共享版本，单击“取消共享此版本”，即可取消插件该版本的共享。 **图 5-14** 取消共享插件版本 
更新共享	1. 在插件详情页面的右上角，单击“更新共享”，在展开的“修改共享内容”弹框中修改插件共享的版本和范围。 2. 修改完成后，单击“确定”完成共享插件的信息更新。

5.5 使用资产中心的提示词资源

提示词广场汇集了平台官方和用户贡献的优质提示词模板，内容丰富全面。在这里，可以复用高质量的提示词模板，轻松解决复杂指令的编写难题，使智能体和工作流更高效地工作。

前提条件

已**购买Versatile智能体平台**。

使用预置的提示词

资产中心提供了丰富的预置提示词模板，您可以通过名称、ID、内容、行业、标签筛选或使用关键字搜索来找到目标模板。用户不仅可以直接引用预置提示词模板来创建新的提示词，还可以将这些模板添加到我的提示词中。这样，您可以更加高效地管理和使用提示词，提升工作效率。

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏中选择“资产中心 > 提示词广场”。

步骤3 通过名称、ID、内容、行业、标签筛选或使用关键字搜索来找到目标模板。

步骤4 单击目标模板右侧的 ··· ，单击“复制提示词”，提示词将复制到剪贴板。

步骤5 单击目标模板右侧的 ··· ，单击“保存到我的提示词”。

图 5-15 复制和保存到我的提示词

步骤6 在弹出的“创建提示词”对话框中，单击“确定”，预置提示词模板将添加到我的提示词中。

在智能体或工作流中使用预置的提示词请参考**配置提示词**或**大模型**、**Agent**、**意图识别**、**高级意图识别**和**提问器**。

----结束

6 接入模型服务

6.1 模型服务介绍

模型服务（Model Serving）是指将机器学习模型部署为服务，以便其他应用程序或系统可以调用这些模型进行预测或决策。模型服务是机器学习生命周期中的一个重要环节，它使得模型能够从开发环境顺利过渡到生产环境，从而实现商业价值。

在Versatile中，模型服务为智能体提供了最核心的智能，使智能体能够自主、智能地完成复杂任务。

模型服务分类

为满足不同用户的技术能力、业务场景及需求，Versatile提供了多样化的模型服务模式。以下从模型来源对各类模型服务进行介绍，具体如表6-1所示。

表 6-1 模型服务分类介绍

分类	特征	使用流程
平台预置的供应商模型服务	由供应商部署，平台接入供应商提供的模型服务API。 目前预置了ModelArts Studio (MaaS)供应商的模型服务。目前支持的模型类型为文本对话、图像理解、文本向量化、文本排序。	接入预置的供应商模型服务流程
用户自主接入的模型服务	由用户或第三方部署在外部环境，平台调用外部已存在的模型服务API。	接入自定义的供应商模型服务流程

平台预置的供应商模型服务：这些服务由供应商部署，系统通过接入其API实现对接。用户只需配置模型鉴权参数，即可便捷地调测和使用。具体操作请参考接入预置的供应商模型服务。

用户自主接入的模型服务：为了满足用户对模型的个性化及专业化需求，Versatile支持接入由用户或第三方部署在外部环境的模型服务API。具体操作请参考**接入自定义的供应商模型服务**。

费用说明

平台预置的供应商模型服务：购买Versatile套餐包，目前支持免费使用2,000,000tokens，具体使用情况可以在**我的资源**中查看。免费资源使用完后，调用平台预置的模型服务，需要在ModelArts Studio开通对应的模型服务，计费请参考**ModelArts Studio（MaaS）模型服务价格**。

用户自主接入的模型服务：接入的模型服务，用户在使用时，Versatile侧不计费，模型供应商侧如果计费，计费规则请参考模型供应商侧的计费规则。

6.2 接入预置的供应商模型服务

6.2.1 接入预置的供应商模型服务流程

Versatile平台接入了ModelArts Studio (MaaS)供应商的模型服务，这些服务由供应商部署，系统通过接入其API实现对接。用户只需配置模型鉴权参数，即可便捷地调测和使用。

图 6-1 平台接入的供应商模型服务使用流程

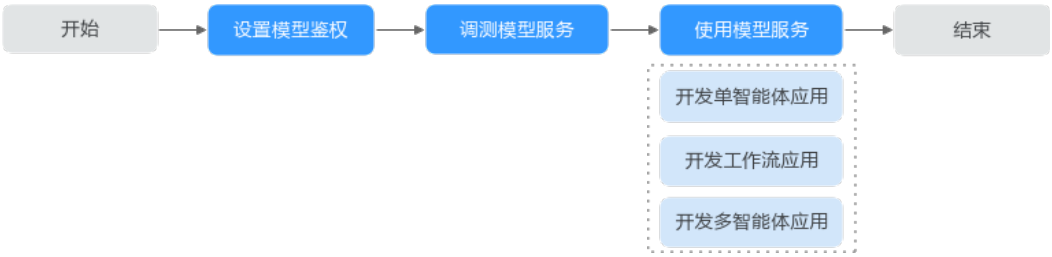


表 6-2 平台接入的供应商模型服务使用流程详解

序号	流程环节	说明
1	设置模型鉴权	调用平台预置的模型服务前，需先进行鉴权设置。具体操作请参考**对预置的供应商模型服务设置鉴权**。
2	调测模型服务	模型调测是指通过对模型进行实际操作、参数调整及效果观测，以验证其在特定场景下的功能表现、性能指标及适用范围的过程，其核心目的是确保模型在真实业务场景中能够稳定、高效地运行。具体操作请参考**调测预置的模型服务**。
3	使用模型服务	模型鉴权设置完成后，可以在智能体、工作流中使用模型服务，请参考**开发单智能体应用**、**开发工作流应用**、**开发多智能体应用**。

6.2.2 对预置的供应商模型服务设置鉴权

平台预置的供应商模型服务在调用前需完成鉴权配置，本文介绍鉴权设置的具体步骤。

前提条件

- 已**购买Versatile智能体平台**。
- 登录用户为空间所有者、空间管理员、开发工程师、运维工程师，详细信息请参考**管理团队空间成员**。

设置模型鉴权

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航，选择“模型中心 > 模型服务”。

步骤3 在“模型服务 > 平台推荐”页面，在“ModelArts Studio (MaaS)”区域，单击“鉴权配置”。

步骤4 页面弹框中会展示鉴权获取链接，如**图6-2**所示，请单击链接前往模型供应商官网进行申请。

图 6-2 获取鉴权信息

1. 单击**图6-2**中的链接，前往ModelArts Studio服务。
   区域选择“西南-贵阳一”，选择其他区域不生效。
2. 单击“创建API Key”，输入标签名称、描述信息，选择访问权限，如**图6-3**所示，单击“确定”。

图 6-3 创建 API key

3. 在“您的API Key”弹窗页面，单击 ，复制API Key。
   API Key在新建后显示一次，请及时复制并妥善保存。

4. 单击“我已保存，确认关闭”。

步骤5 在输入框中输入获取的鉴权信息，单击“确定”。

设置鉴权后，模型供应商由“未配置鉴权”变为“已配置鉴权”。**已配置鉴权**的模型，支持调测、使用。

----结束

相关操作

在模型供应商列表，支持的其他操作请参考**表6-3**。

表 6-3 供应商模型相关操作

操作	说明
移除鉴权配置	**警告** 移除鉴权配置后，模型供应商为“未配置鉴权”，**未配置鉴权**的模型，不支持调测、使用。 在“平台推荐”页面，对于不再使用的模型，需要移除鉴权配置。 在“ModelArts Studio (MaaS)”区域，单击“清空鉴权”，单击“移除”。

相关文档

- 鉴权设置完成后，可以调测模型服务，请参考**调测预置的模型服务**。
- 鉴权设置完成并且调测成功后，可以在智能体、工作流中使用模型服务，请参考**开发单智能体应用**、**开发工作流应用**、**开发多智能体应用**。

6.2.3 调测预置的模型服务

模型调测是指通过对模型进行实际操作、参数调整及效果观测，以验证其在特定场景下的功能表现、性能指标及适用范围的过程，其核心目的是确保模型在真实业务场景中能够稳定、高效地运行。本章介绍平台接入的供应商模型调测流程。目前平台预置了ModelArts Studio (MaaS)供应商的模型服务，目前支持的模型类型为文本对话、图像理解、文本向量化、文本排序。

前提条件

- 已**购买Versatile智能体平台**。
- 已**对预置的供应商模型服务设置鉴权**。
- 登录用户为空间所有者、空间管理员、开发工程师、运维工程师，详细信息请参考**管理团队空间成员**。

调测模型服务

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航，选择“模型中心 > 模型服务”。

步骤3 在“模型服务 > 平台推荐”页面，在模型服务卡片上，单击“调测”。

图 6-4 调测

步骤4 在“模型调测”页面，可以调测如下几种类型的模型服务。

- **文本对话**

  a. 在“模型类型”区域选择“文本对话”，参数配置请参考**表6-4**。

图 6-5 文本对话

表 6-4 文本对话类型模型参数说明

参数	说明	示例
模型服务	“模型服务A”默认展示所选的供应商模型服务。“模型服务B”为可选项。 您也可以在下拉列表选择或切换以下模型服务： ■ **用户自主接入的模型服务**：以模型供应商维度展示。 ■ **平台推荐**：以模型供应商维度展示。 ■ **路由策略**：用户自定义创建的路由策略。	DeepSeek-V3
深度思考	显示该参数有以下两个场景： ■ **平台推荐**：当选择的模型服务为思考模型且支持关闭深度思考时，才显示此参数，例如平台推荐的Qwen3-32B、DeepSeek-V3.2。 ■ **用户自主接入的模型服务**：当选择的模型服务为思考模型且在新建模型服务开启了“是否支持关闭思维链输出”时，才显示此参数。 该参数支持以下操作： ■ 当此功能**开启**时，大模型将首先进行深入的思考和推理，通过逐步拆解问题、梳理逻辑，生成一段详细的思维链内容，并在调试界面展示。这一过程有助于提升最终输出答案的准确性和可靠性，确保用户获得更加精准的信息。 ■ 当此功能**关闭**时，智能体将直接生成最终答案，不再经过额外的思维链推理过程。这将加快响应速度，适用于需要快速获取答案的场景。 **注意** 在模型使用过程中，“深度思考”开关生效的情况如下： ■ 如果模型支持思维链输出能力，并且也支持关闭该能力，则开启、关闭均生效。 ■ 如果模型支持思维链输出能力，但不支持关闭该能力，则开启生效、关闭不生效。 ■ 如果模型不支持思维链输出能力，则开启、关闭均不生效。	开启

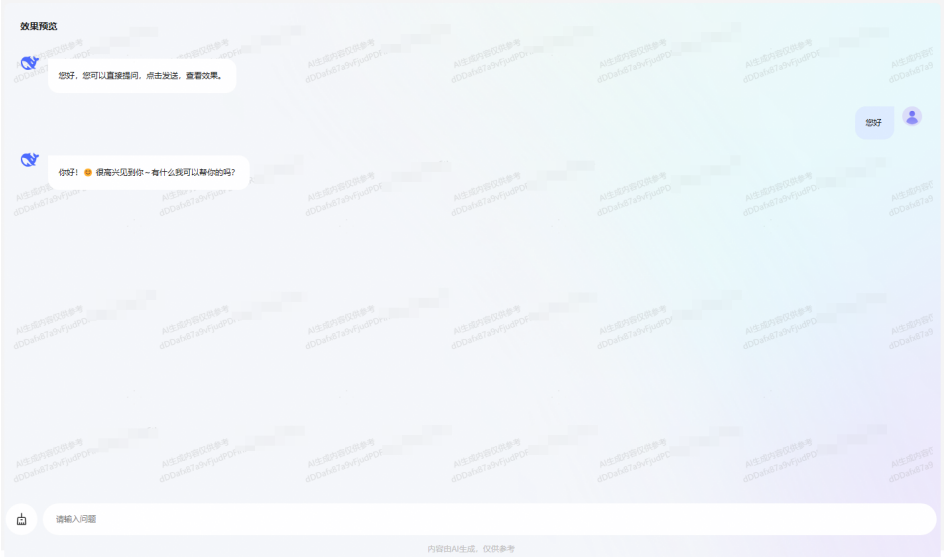
参数	说明	示例
输出方式	▪ **非流式**：调用大语言模型推理服务时，根据用户问题，获取大语言模型的回答，大语言模型完整生成回答后一次性返回。 ▪ **流式**：调用大语言模型推理服务时，根据用户问题，获取大语言模型的回答，逐个字词的快速返回模式，不需等待大语言模型生成完成。默认**流式**。	流式
输出最大token数	模型在单次推理或生成内容时，能够输出的token（模型处理文本的基本单位）数量的最大值。取值范围100~32768，默认值为2048。	2048
温度	较高的数值会使输出更加随机，而较低的数值会使其更加集中和确定。取值范围0.01~2，默认值为0.5。 建议该参数和“多样性”只设置1个。	0.5
多样性	影响输出文本的多样性，取值越大，生成文本的多样性越强。取值范围0~1，默认值为0.5。 建议该参数和“温度”只设置1个。	0.5
存在惩罚	正值会尽量避免使用已出现过的词语，更倾向于生成新词语。取值范围-2.0~2.0，默认值为0。	0
频率惩罚	正值会尽量避免使用常见的单词和短语，更倾向于生成较少见的单词。取值范围-2.0~2.0，默认值为0。	0

b. 在右侧“效果预览”区域，在对话输入框输入测试语句后按Enter键或单击  ，查看模型响应结果。

单击  ，清除本次会话内容，可以开始新的会话。

调测成功后，可以在智能体、工作流中使用模型服务，请参考**开发单智能体应用**、**开发工作流应用**、**开发多智能体应用**。

图 6-6 文本对话模型调测成功



- **图像理解**

  a. 在“模型类型”区域选择“图像理解”，参数配置请参考**表6-5**。

图 6-7 图像理解

表 6-5 图像理解类型模型参数说明

参数	说明	示例
模型服务	默认展示所选的供应商模型服务。 您也可以在下拉列表切换以下模型服务： ■ **用户自主接入的模型服务**：以模型供应商维度展示。 ■ **平台推荐**：以模型供应商维度展示。	Qwen2.5-VL-72B
输出方式	■ **非流式**：调用大语言模型推理服务时，根据用户问题，获取大语言模型的回答，大语言模型完整生成回答后一次性返回。默认**非流式**。 ■ **流式**：调用大语言模型推理服务时，根据用户问题，获取大语言模型的回答，逐个字词的快速返回模式，不需等待大语言模型生成完成。	非流式
上传图片	单击，可上传本地图片。支持上传JPG、PNG格式图片，且不大于4MB。	-
提示语内容	输入提示语，对图片进行提问。	图片里有什么?

b. 单击“生成图像理解”，在右侧“效果预览”区域查看模型响应效果。

调测成功后，可以在智能体、工作流中使用模型服务，请参考**开发单智能体应用**、**开发工作流应用**、**开发多智能体应用**。

图 6-8 图像理解模型调测成功

- **文本向量化**

a. 在“模型类型”区域选择“文本向量化”，参数配置请参考表6-6。

图 6-9 文本向量化

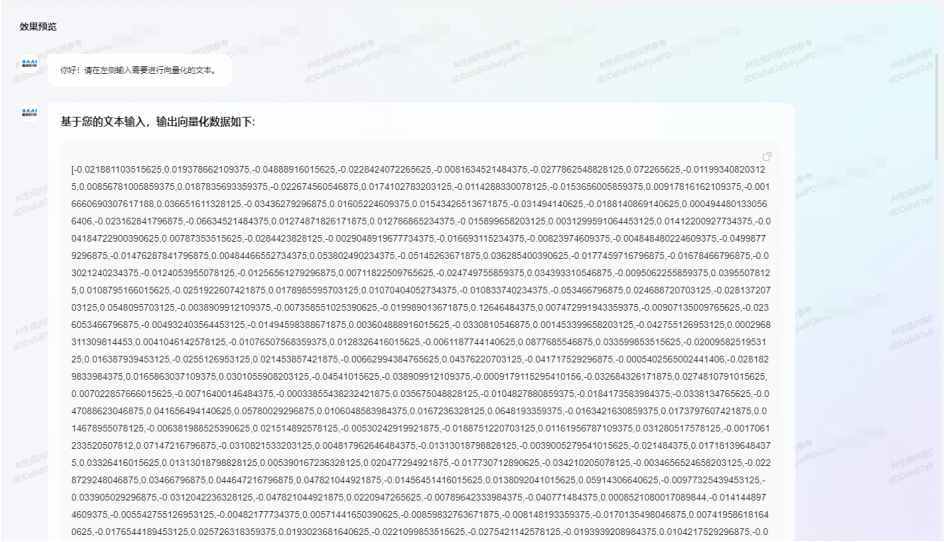
表 6-6 文本向量化类型模型参数说明

参数	说明	示例
模型服务	默认展示所选的供应商模型服务。 您也可以在下拉列表切换以下模型服务： ■ **用户自主接入的模型服务**：以模型供应商维度展示。 ■ **平台推荐**，以模型供应商维度展示。	BGE-M3
请输入文本	输入待向量化的文本，可参照以下示例： ■ 示例1：那是个快乐的人 ■ 示例2： ["那是个快乐的人", "那是个高兴的人", "那是个忧郁的人"]	那是个快乐的人

b. 单击“生成向量化”，在右侧“效果预览”区域查看模型响应效果。

调测成功后，可以在智能体、工作流中使用模型服务，请参考**开发单智能体应用**、**开发工作流应用**、**开发多智能体应用**。

图 6-10 文本向量化调测成功



- **文本排序**

a. 在“模型类型”区域选择“文本排序”，参数配置请参考**表6-7**。

图 6-11 文本排序

表 6-7 文本排序类型模型参数说明

参数名称	参数说明	示例
模型服务	默认展示所选的模型服务。 您也可以在下拉列表切换以下模型服务： ■ **用户自主接入的模型服务**：以模型供应商维度展示。 ■ **平台推荐**，以模型供应商维度展示。	BGE-Reranker-V2-M3
待排序文本	输入待排序文本。单击 + 添加文本，最多可以添加10条。	小朋友在学校很快乐
被展示文本条数	文本排序完成后，展示的条数。取值范围为1~10，默认值为1。	3
我的问题	描述想要解决的问题。	小朋友在学校怎么样?

b. 单击“开始排序”，在右侧“效果预览”区域查看模型响应效果。

调测成功后，可以在智能体、工作流中使用模型服务，请参考**开发单智能体应用**、**开发工作流应用**、**开发多智能体应用**。

图 6-12 文本排序调测成功

----结束

6.3 接入自定义的供应商模型服务

6.3.1 接入自定义的供应商模型服务流程

Versatile支持接入由用户或第三方部署在外部环境的模型服务API。模型服务接入后，用户可进行调测和使用。

图 6-13 接入用户自定义的供应商模型服务使用流程

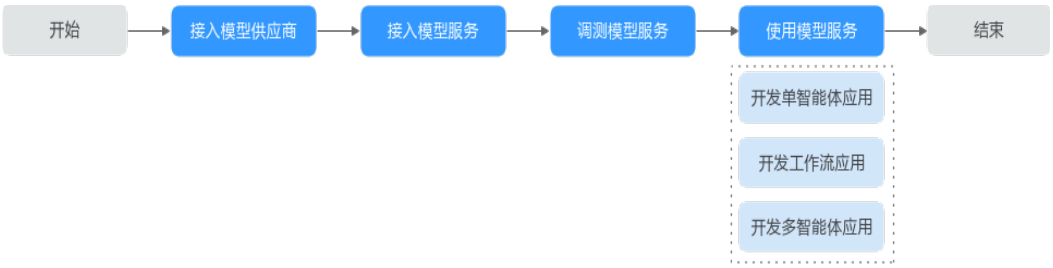


表 6-8 接入用户自定义的模型供应商服务使用流程详解

序号	流程环节	说明
1	接入模型供应商	Versatile支持接入由用户或第三方部署在外部环境的模型服务API。接入模型服务之前需要先接入模型供应商。具体操作请参考**接入模型供应商**。
2	接入模型服务	Versatile支持接入由用户或第三方部署在外部环境的模型服务API。具体操作请参考**接入自定义的模型服务**。
3	调测模型服务	模型体验是指通过对模型进行实际操作、参数调整及效果观测，以验证其在特定场景下的功能表现、性能指标及适用范围的过程，其核心目的是确保模型在真实业务场景中能够稳定、高效地运行。具体操作请参考**调测已接入的模型服务**。
4	使用模型服务	模型服务接入后，可以在智能体、工作流中使用模型服务，请参考**开发单智能体应用**、**开发工作流应用**。

6.3.2 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快速获取和使用高质量的模型。

Versatile支持接入由用户或第三方部署在外部环境的模型服务API，在模型服务接入之前需要先接入模型供应商。

前提条件

- 已**购买Versatile智能体平台**。
- 登录用户为空间所有者、空间管理员、开发工程师、运维工程师，详细信息请参考**管理团队空间成员**。

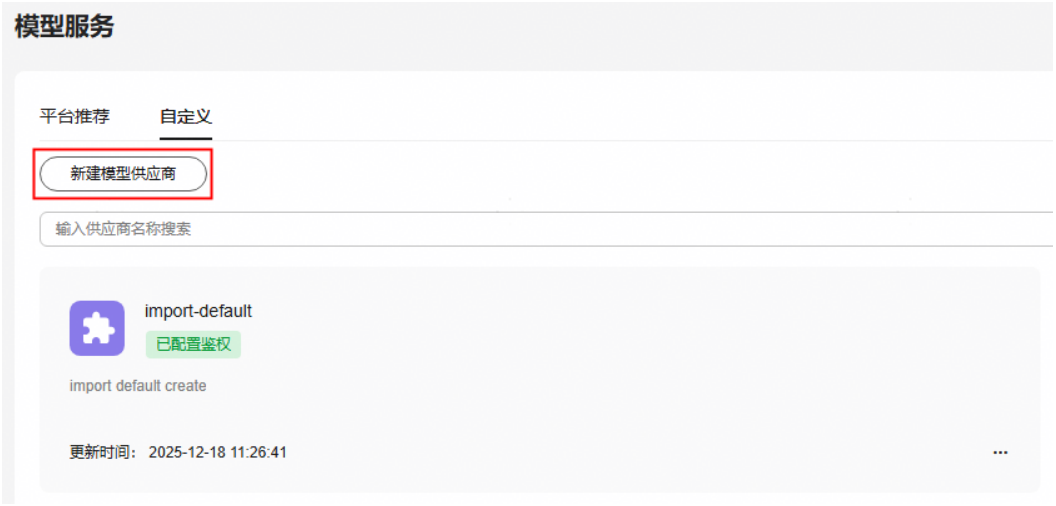
新建模型供应商

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航，选择“模型中心 > 模型服务”，进入“模型服务”页面。

步骤3 选择“自定义”页签，单击“新建模型供应商”。

图 6-14 新建模型供应商



步骤4 在“新建模型供应商”页面，配置参数信息，如图6-15所示，具体参数说明请参考表6-9。

接入示例请参考DeepSeek模型配置示例。

图 6-15 新建模型供应商

表 6-9 新建模型供应商参数说明

参数	说明	示例
供应商图标	系统默认供应商图标，用户也可以自定义图标。 1. 鼠标移动至系统默认图标上，单击鼠标左键。 2. 在虚线框中，单击鼠标左键，上传已准备好的供应商图标。 支持jpg、png格式图片，且不大于100KB。	系统默认图标

参数	说明	示例
供应商名称	供应商的名称。由2~64个字符组成，包含中英文、数字、下划线、中划线、空格、竖线。	深度求索
供应商英文名称	供应商的英文名称。由2~64个字符组成，包含英文、数字、下划线、中划线、空格、竖线。	DeepSeek
描述	选填项。 供应商的描述信息。由0~1000个字符组成。	深度求索（DeepSeek），是量化巨头幻方探索AGI（通用人工智能）的新组织，成立于2023年，专注于研究世界领先的通用人工智能底层模型与技术，挑战人工智能前沿性难题。

参数	说明	示例
选择认证方式	在智能体、工作流中调用该模型服务或通过API调用该模型服务时，认证鉴权的方式。 • **无鉴权** • **Api-key**：Api-key认证方式，通过请求header的Authentication字段携带Bearer <Api-key>进行认证，需要提供Api-key。 输入待接入的供应商模型服务的API key，输入后会将关键信息进行加密保存。设置后2分钟后生效，用于调用模型服务。 • **AK/SK**：适用于盘古大模型的AK/SK认证方式，通过AK（ Access Key ID ）/SK（ Secret Access Key ）加密调用请求，需要提供AK和SK。 – 请输入AK：输入待接入的供应商模型服务的AK，输入后会将关键信息进行加密保存。设置后2分钟后生效，用于调用模型服务。 – 请输入SK：输入待接入的供应商模型服务的SK，输入后会将关键信息进行加密保存。设置后2分钟后生效，用于调用模型服务。 • **App-code**：APP认证方式，通过请求header的X-Apig-Appcode字段携带App-code进行认证，需要提供App-code。 输入待接入的供应商模型服务的App code，输入后会将关键信息进行加密保存。设置后2分钟后生效，用于调用模型服务。 • **自定义ApiKey**：通过自定义请求头认证参数进行认证。 在“Header配置”右侧，单击 ，新增调用模型服务的Header参数。 在“参数名称”、“参数值”中输入待接入的供应商模型服务的Header参数。设置后，用于调用模型服务。 • **IAM鉴权**：华为云IAM认证，通过IAM账号获取用户Token进行认证。 进行鉴权配置。输入待接入的供应商模型服务的IAM鉴权信息。设置后，用于调用模型服务。 – IAM认证url：获取IAM用户Token信息的接口。例如，https://{iam_host}/v3/auth/tokens。 – 账号名：IAM用户所属账号信息，即账号名。 – 项目：该服务所属区域信息。例如，cn-southwest-2。	Api-key

参数	说明	示例
	– 验证方式 ■ IAM用户名/密码 IAM用户名：IAM用户名称。 IAM用户密码：IAM用户的登录密码。 ■ Access Key ID/Secret Access Key Access Key ID：访问密钥ID。 Secret Access Key：与访问密钥ID结合使用的密钥。	

步骤5 单击“确定”。

新建的模型供应商，显示在“自定义”模型供应商卡片列表中。模型服务为“已配置鉴权”。**已配置鉴权**的模型，支持调测、使用。

----结束

相关操作

在模型供应商卡片列表，支持的其他操作请参考**表6-10**。

表 6-10 模型供应商信息相关操作

操作	说明
移除鉴权配置	在新增模型供应商时，“鉴权方式”选择“Api-key”、“AK/SK”、“App-code”、“自定义ApiKey”、“IAM鉴权”时，才支持移除鉴权配置。 在需要移除鉴权配置的模型供应商卡片上，单击“ ··· > 鉴权配置”，单击“移除”。 移除鉴权配置后，模型服务为“未配置鉴权”，**未配置鉴权**的模型，不支持调测、使用。
修改模型供应商信息	在需要修改的模型供应商卡片上，单击“ ··· > 编辑”。
删除模型供应商	在需要删除的模型供应商卡片上，单击“ ··· > 删除”。 模型供应商中有已发布的模型服务，需要先删除模型服务。

相关文档

接入模型供应商后，可以在模型供应商中接入模型服务，具体操作请参考**接入自定义的模型服务**。

6.3.3 接入自定义的模型服务

为了满足用户对模型的个性化及专业化需求，Versatile支持接入由用户或第三方部署在外部环境的模型服务API。支持接入的模型类型包括文本对话（Chat）、文本向量化

（Embeddings）、文本排序（Rerank）、图像理解。模型服务接入后，用户可以进行调测和使用。

为了保证接入模型服务的质量，模型API接入之前，请确保符合相对应的接口规范，其中文本对话、文本向量化、图像理解类型需要符合OpenAI接口规范，文本排序类型需要符合AI引擎标准协议。标准OpenAI协议和AI引擎标准协议规范请参考**接入模型服务API接口规范**。

自定义模型服务的优势：

- 优化个性化体验：个性化可以减少用户的搜索和选择时间，提供更加流畅和高效的用户体验。例如，搜索引擎可以根据用户的搜索历史和偏好，提供更加精准的搜索结果。
- 增强特性领域的准确性：接入专业的模型服务，可以显著提高特定领域的准确性。例如，在医疗领域，接入专业的医学模型可以提供更准确的诊断建议。
- 提升开发效率：通过快速接入用户或第三方部署在外部环境的模型服务API，可以显著提升开发效率。开发人员无需从零开始构建复杂的模型，而是可以直接利用已有的高质量模型。

前提条件

- 已**购买Versatile智能体平台**。
- 已**接入模型供应商**。
- 登录用户为空间所有者、空间管理员、开发工程师、运维工程师，详细信息请参考**管理团队空间成员**。

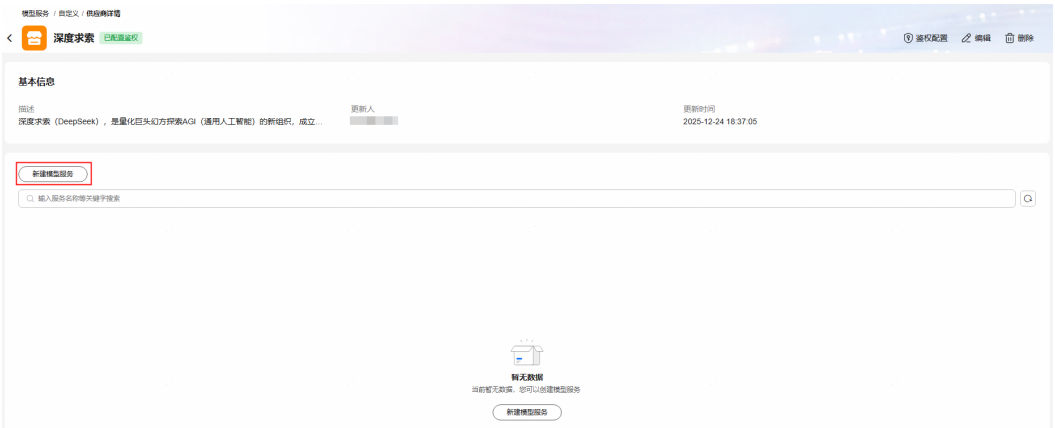
新建模型服务

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航，选择“模型中心 > 模型服务”，进入“模型服务”页面。

步骤3 选择“自定义”页签，单击对应的模型供应商卡片，在“供应商详情”页面，单击“新建模型服务”。

图 6-16 新建模型服务



步骤4 在“新建模型服务”页面，配置参数信息，具体参数说明请参考**表6-11**。

接入示例请参考**DeepSeek模型配置示例**。

表 6-11 新建模型服务参数说明

参数	说明	示例
模型服务图标	系统默认模型服务图标，用户也可以自定义图标。 1. 鼠标移动至系统默认图标上，单击鼠标左键。 2. 在虚线框中，单击鼠标左键，上传已备注的模型服务图标。 支持jpg、png格式图片，且不大于100KB。	系统默认图标
模型服务名称	自定义模型服务名称。由2~64个字符组成，包含中英文、数字及:._/\-，以中英文、数字开头结尾。	文本对话
模型名称	填写的模型名称必须为该模型的模型ID/模型编码，否则会导致模型不可用。 需要登录第三方模型厂商官网查看，例如，Baichuan4、deepseek-chat、glm-4-air。 如果要接入自建的模型服务，该模型名称将用于接口调用请求体的model字段。 由2~64个字符组成，包含中英文、数字及:._/\-，以中英文、数字开头结尾。	deepseek-chat
模型类型	选择模型类型。 • **文本对话**：文本对话模型，通常被称为对话式AI或聊天机器人，是一种经过训练能够理解和生成人类语言，并以多轮、上下文连贯的方式进行交流的人工智能系统。 • **文本向量化**：文本向量化模型的核心任务是将文本（词、句、段落或文档）转换为计算机能够理解和处理的数值形式——即高维向量（也称为“嵌入”，Embedding）。这个向量就像是文本在数学空间中的一个“坐标点”。 • **文本排序**：文本排序模型用于对一组文本对象进行相关度排序。给定一个查询（Query）和一个文本列表（如搜索引擎的结果），排序模型会根据每个文本与查询的相关程度，从高到低进行排序。 • **图像理解**：图像理解模型是一种能够对图像内容进行分析、解读和理解的人工智能模型，其核心目标是让计算机像人类一样“看懂”图像。	文本对话
模型服务API地址	填入需要接入模型的API地址信息。字符长度不大于255个字符。 格式为：https://*xxx*.com/v1/*xxx*。	https://api.deepseek.com/chat/completions

参数	说明	示例
API接口协议	• 当“模型类型”值为“文本对话”、“文本向量化”、“图像理解”时，选择“标准OpenAI协议”、“阿里千问接口协议”、“MaaS标准API V1”、“MaaS标准API V2”。 • 当“模型类型”值为“文本排序”时，选择“AI引擎标准协议”。 **标准OpenAI协议**和**AI引擎标准协议**规范请参考**接入模型服务API接口规范**。 **阿里千问接口协议**规范请参考通义千问的接口协议。 **MaaS标准API V1**接口规范请参考**MaaS标准API V2**。 **MaaS标准API V2**接口规范请参考**MaaS标准API V1**。	标准OpenAI协议
流控配置	超出流控值，则触发限流，用户的请求会因为流控而失败。 • 无限制 • 10次/秒 • 50次/秒 • 100次/秒 • 200次/秒	无限制
选择标签	可选项。 当“模型类型”值为“文本对话”、“图像理解”时，才有此参数。 选择标签后，在应用中选择大模型时，显示在大模型右侧。 • **工具**：该大模型支持应用调用外部工具时，例如，MCP服务、插件、知识库，可以选择该标签。 • **思考**：该大模型具备思维推理时，可以选择该标签。 • **联网**：该大模型具备联网搜索能力时，可以选择该标签。	工具

参数	说明	示例
是否支持关闭思维链输出	当“选择标签”选择了**思考**时，才有此参数。 • 开启：模型在调测、使用时，显示“深度思考”参数。 在模型的调测和使用过程中，“深度思考”开关的生效情况如下： – 如果模型支持思维链输出能力，并且也支持关闭该能力，则开启、关闭均生效。 – 如果模型支持思维链输出能力，但不支持关闭该能力，则开启生效、关闭不生效。 – 如果模型不支持思维链输出能力，则开启、关闭均不生效。 • 关闭：模型在调测、使用时，不显示“深度思考”参数。默认**关闭**。 模型在调测、使用时，是否输出思维链，取决于模型本身是否支持思维链输出。	关闭
自定义标签	选填项。 最多支持添加5个标签。单击 ，输入标签内容，按Enter键。 添加后，在应用中选择大模型时，显示在大模型右侧。	-
模型服务描述	选填项。 模型服务的描述信息。由0~1000个字符组成。	-

步骤5 单击“确定”。

新建的模型服务，显示在模型供应商下的模型服务卡片列表中。模型服务为“未发布”。

步骤6 在需要调测的模型服务卡片上，单击“ ··· > 调测”，具体调测操作请参考**调测已接入的模型服务**。

步骤7 在需要发布的模型服务卡片上，单击“ ··· > 发布模型”。

模型服务为“已发布”。**已发布**的模型服务，才支持使用。

----结束

DeepSeek 模型配置示例

步骤1 请在DeepSeek官网购买并获取模型的API Key。

具体操作请参考**DeepSeek文档**。

步骤2 从DeepSeek官网获取调用API文档，如**图6-17**所示。

具体请参考**DeepSeek文档**。

图 6-17 模型示例

curl python nodejs

```
curl https://api.deepseek.com/chat/completions \
  -H "Content-Type: application/json" \
  -H "Authorization: Bearer <DeepSeek API Key>" \
  -d '{
        "model": "deepseek-chat",
        "messages": [
          {"role": "system", "content": "You are a helpful assistant."},
          {"role": "user", "content": "Hello!"}
        ],
        "stream": false
      }'
```

模型地址

认证 api key

模型名称

```
curl https://api.deepseek.com/chat/completions \
  -H "Content-Type: application/json" \
  -H "Authorization: Bearer ${DEEPSEEK_API_KEY}" \
  -d '{
        "model": "deepseek-chat",
        "messages": [
          {"role": "system", "content": "You are a helpful assistant."},
          {"role": "user", "content": "Hello!"}
        ],
        "stream": false
      }'
```

步骤3 参考**步骤1**~**步骤3**，新建模型服务，配置示例如**图6-18**所示。

图 6-18 新增模型服务

步骤4 请参考**步骤5**~**步骤7**。

----结束

相关操作

在接入模型服务卡片列表，支持的其他操作请参考**表6-12**。

表 6-12 接入模型服务相关操作

操作	说明
查看接入模型服务信息	单击接入模型服务卡片，进入模型服务详情页，可以查看模型服务信息。
修改接入模型服务信息	在需要修改的接入模型服务卡片上，单击“ ··· > 编辑”。 未发布的接入模型服务，才可以修改。
取消发布	在需要取消发布的接入模型服务卡片上，单击“ ··· > 取消发布”。 **未发布**的模型服务，不支持使用。 已发布的接入模型服务，才可以取消发布。
删除接入模型服务	在需要删除的接入模型服务卡片上，单击“ ··· > 删除”。 已发布的接入模型服务，需要先取消发布，才能删除。

相关文档

- 模型服务接入后，可以调测模型服务，具体操作请参考**调测已接入的模型服务**。
- 模型服务接入后，可以在智能体、工作流中使用模型服务，请参考**开发单智能体应用**、**开发工作流应用**、**开发多智能体应用**。

6.3.4 接入模型服务 API 接口规范

当前模型网关支持文本对话（Chat）、文本向量化（Embeddings）、文本排序（Rerank）、图像理解类型的API接入。

模型API接入之前，请确保符合相对应的接口规范，其中文本对话、文本向量化、图像理解类型需要符合OpenAI接口规范，文本排序类型需要符合AI引擎标准协议。

文本对话（Chat）API 规范

接口格式

类型：POST

协议：HTTP/HTTPS

请求体参数

表 6-13 请求体参数

参数	是否必选	参数类型	描述
messages	是	Array of **表 6-14** objects	文本对话消息体类。
model	是	String	文本对话使用的模型名称。

参数	是否必选	参数类型	描述
frequency_penalty	否	Number	频率惩罚，会根据文本中新Token的出现频率对其进行惩罚，从而降低模型重复相同内容的可能性。使其生成的文本更加自然和符合预期。取值范围为-2.0~2.0。 ● 默认值（0.0）：不施加任何频率惩罚。模型按原本的概率分布生成文本。 ● 正值（例如 0.5，1.0，2.0）：增加惩罚力度。值越高，模型越不愿意使用已经用过的词。使输出文本的词汇更多样化、更富有创造性，但过高的值可能导致用词生僻、语句不通顺甚至偏离主题。 ● 负值（例如 -0.5，-1.0，-2.0）：减少惩罚，值越低（负的越多），模型越倾向于使用已经用过的词。使输出文本的词汇更集中、更稳定、更可能重复关键主题词。但过低的值会导致用词极其重复、啰嗦。 例如： 提示词为（Prompt）：“写一首关于猫的诗。” ● frequency_penalty=0（默认）：输出可能正常地重复使用“猫”、“尾巴”、“柔软”等合理词汇。 ● frequency_penalty=1.5（高惩罚）：模型会极力避免重复用词。第一句用了“猫”，第二句可能会用“毛茸伙伴”、“喵星人”、“优雅的生物”等同义词来替代，词汇非常丰富。但如果惩罚过高，可能会为了规避重复而选用不合适的词，导致诗歌变得奇怪。 ● frequency_penalty=-1.0（负惩罚）：模型不害怕重复，甚至鼓励重复。输出可能会变成：“猫，猫，可爱的

参数	是否必选	参数类型	描述
			猫。猫在跑，猫在跳，猫的尾巴摇啊摇。” 显得非常冗余和缺乏创意。
logit_bias	否	Map<String,Integer>	该参数接受一个JSON对象，将标记映射到从-100（禁止）到100（独占选择标记）的关联偏差值。 像-1和1这样的适度值将以较小的程度改变选择标记的概率。 使用logit_bias参数时，偏差被添加到模型生成的logits之前进行抽样。
max_tokens	否	Integer	返回体允许的最大token数。
n	否	Integer	返回体中包含的choices数量，建议默认设置为1，最大限度地降低成本。 ● 最小值：1 ● 最大值：128 ● 缺省值：1

参数	是否必选	参数类型	描述
presence_penalty	否	Number	存在惩罚，会根据文本中新Token是否出现对其进行惩罚，核心作用是降低模型再次讨论已经出现过的“话题”的可能性，从而增加模型谈论新主题的可能性。使其生成的文本更加自然和符合预期。取值范围为-2.0~2.0。 ● 默认值（0.0）：不施加任何存在惩罚。 ● 正值（例如 0.5，1.0，2.0）：增加惩罚力度，值越高，模型越不愿意停留在已经提及的主题上，越倾向于引入全新的想法、概念或话题。使对话或文本更容易“跑题”或转向新方向。在创意生成中，这可以带来更大的探索性。 ● 负值（例如 -0.5，-1.0，-2.0）：减少惩罚，值越低（负的越多），模型越倾向于围绕已经出现的主题进行深入讨论，避免引入新信息，使输出内容更加集中、紧扣主题，但可能显得缺乏发散性。

参数	是否必选	参数类型	描述
stream	否	Boolean	布尔类型。 • 设为true时，返回结果为流式。 流式处理是一种边接收边处理、实时性强的数据处理模式。它将数据视为连续不断的“流”，允许低延迟和即时响应，广泛应用于视频播放、实时监控、大数据分析和人工智能生成内容等领域。 • 设为false时，返回结果为非流式，JSON格式结构化数据。 非流式指数据或操作不是连续、实时地传输或处理，而是一次性接收完整的输入，待完全处理后一次性返回完整的结果。这种模式常见于传统HTTP请求-响应模式、大模型API的非流式输出和语音合成等领域，强调数据的完整性和整体性，不追求即时反馈。 缺省值：false
temperature	否	Number	较高的数值会使输出更加随机，而较低的数值会使其更加集中和确定。 • 最小值：0 • 最大值：2 • 缺省值：1
top_p	否	Number	影响输出文本的多样性，取值越大，生成文本的多样性越强。 • 最小值：0.0 • 最大值：1.0 • 缺省值：1
tools	否	Array of **表 6-15** objects	可供模型调用的工具。
tool_choice	否	String	用于控制模型是如何选择要调用的函数，仅当工具类型为function时补充。 默认为auto，且当前仅支持auto。

表 6-14 ChatCompletionRequestMessage

参数	是否必选	参数类型	描述
role	是	String	消息体对应的角色。 • system：如果是系统，则为system。 • user：如果是用户，则为user。
content	是	String	消息具体内容。
name	否	String	对话参与者的可选名称，提供给模型信息以区分相同角色的不同对话参与者。

表 6-15 FunctionCallTool

参数	是否必选	参数类型	描述
type	否	String	调用工具类型，目前仅支持function。
function	否	表6-16 object	仅当工具类型为function时补充。

表 6-16 function

参数	是否必选	参数类型	描述
name	否	String	函数名称，只能包含a-z、A-Z、0-9、下划线和中横线。最大长度为64个字符。
description	否	String	用于描述函数功能。 模型会根据这段描述决定函数调用方式。
parameters	否	Object	Json Schema对象，用于定义函数所接受的参数。

- **工具调用请求示例**
  - 流式请求示例

```
{
    "model": "my-chat-model",
    "messages": [
        {
            "role": "user",
            "content": "请帮我查询南京的天气"
        }
```

```
  ],
  "tools": [
    {
      "type": "function",
      "function": {
        "name": "get_weather",
        "description": "获取给定地点的天气",
        "parameters": {
          "type": "object",
          "properties": {
            "location": {
              "type": "string",
              "description": "地点，例如北京、上海。"
            }
          },
          "required": ["location"]
        }
      }
    }
  ],
  "max_tokens": 200,
  "presence_penalty": 1.2,
  "frequency_penalty": 1.0,
  "temperature": 0.5,
  "top_p": 0.95,
  "stream": true
}
```

- 非流式请求示例

```
{
  "model": "my-chat-model",
  "messages": [
    {
      "role": "user",
      "content": "请帮我查询南京的天气"
    }
  ],
  "tools": [
    {
      "type": "function",
      "function": {
        "name": "get_weather",
        "description": "获取给定地点的天气",
        "parameters": {
          "type": "object",
          "properties": {
            "location": {
              "type": "string",
              "description": "地点，例如北京、上海。"
            }
          },
          "required": ["location"]
        }
      }
    }
  ],
  "max_tokens": 200,
  "presence_penalty": 1.2,
  "frequency_penalty": 1.0,
  "temperature": 0.5,
  "top_p": 0.95,
```



```
    "stream": false
}
```

- **非工具调用请求示例**
  - 流式请求示例

```
{
    "model": "my-chat-model",
    "messages": [
        {
            "role": "system",
            "content": " You are a helpful assistant. "
        },
        {
            "role": "user",
            "content": "你好！ "
        }
    ],
    "max_tokens": 20,
    "presence_penalty": 1.2,
    "frequency_penalty": 1.0,
    "temperature": 0.5,
    "top_p": 0.95,
    "stream": true
}
```

  - 非流式请求示例

```
{
    "model": "my-chat-model",
    "messages": [
        {
            "role": "system",
            "content": " You are a helpful assistant. "
        },
        {
            "role": "user",
            "content": "你好！ "
        }
    ],
    "max_tokens": 20,
    "presence_penalty": 1.2,
    "frequency_penalty": 1.0,
    "temperature": 0.5,
    "top_p": 0.95,
    "stream": false
}
```

响应体参数

表 6-17 响应体参数

参数	参数类型	描述
id	String	文本对话唯一标识符。
choices	Array of **表 6-18** objects	返回体列表。 如果“n”大于1，则结果为多个。
created	Integer	问答发生的时间。格式为时间戳。
model	String	文本对话使用的模型名称。

参数	参数类型	描述
object	String	固定值“chat.completion”。
usage	表6-22 object	文本对话用量统计。

表 6-18 choices

参数	参数类型	描述
index	Integer	返回多个choices时，每个choice对应的顺序。
message	表6-19 object	模型服务返回的具体消息体内容。
finish_reason	String	返回结束的原因。 • stop：模型达到自然停止点或提供的停止序列。 • length：达到请求中指定的最大令牌数。 • content_filter：由于内容过滤器的标志而省略了内容。 • tool_calls：模型选择了某个工具。

表 6-19 ChatCompletionResponseMessage

参数	参数类型	描述
content	String	返回消息体的内容，与tool_calls二选一。
role	String	返回消息体的角色。 枚举值： • user：用户输入的问题。 • assistant：大模型的答复内容。
tool_calls	Array of 表6-20 objects	工具调用消息，与content二选一。

表 6-20 ToolCall

参数	参数类型	描述
id	String	工具调用唯一标识符。
type	String	工具类型，当前仅支持function。
function	表6-21 Object	调用函数的详细信息。

表 6-21 CallFunction

参数	参数类型	描述
name	String	函数名。
arguments	String	调用函数的参数，JSON格式。

表 6-22 CompletionUsage

参数	参数类型	描述
completion_tokens	Integer	回答包含的token数。
prompt_tokens	Integer	提问包含的token数。
total_tokens	Integer	提问+回答token总数。

- **工具调用响应示例**
  - 流式响应示例

    流式返回的工具调用信息必须在一条消息内，不能分拆返回。

    ```
    {"id":"chatcmpl-xxx","object":"chat.completion.chunk","created":1718772336,"model":"my-chat-model","choices":[{"index":0,"delta":{"role":"assistant","content":null,"tool_calls":[{"id":"call_123","type":"function","function":{"name":"get_weather","arguments":"{\"location\":\"南京\"}"}}]}"logprobs":null,"finish_reason":null}]}
    {"id":"chatcmpl-xxx","object":"chat.completion.chunk","created":1718772336,"model":"my-chat-model","choices":[{"index":0,"delta":{},"logprobs":null,"finish_reason":"tool_calls"}]}
    ```

  - 非流式响应示例

    ```
    {
       "id": "chatcmpl-xxx",
       "object": "chat.completion",
       "created": 1718772336,
       "model": "my-chat-model",
       "choices": [
          {
             "index": 0,
             "message": {
                "role": "assistant",
                "content": null,
                "tool_calls": [
                   {
                      "id": "call_123",
                      "type": "function",
                      "function": {
                         "name": "get_weather",
                         "arguments": "{\"location\": \"南京\"}"
                      }
                   }
                ]
             },
    ```

```
            "finish_reason": "tool_calls",
            "logprobs": null
        }
    ],
    "usage": {
        "prompt_tokens": 5,
        "completion_tokens": 10,
        "total_tokens": 15
    }
}
```

- **非工具调用响应示例**
  - 流式响应示例

```
{"id":"chatcmpl-xxx","object":"chat.completion.chunk","created":1718772336,"model":"my-chat-model","choices":[{"index":0,"delta":{"role":"assistant","content":""},"logprobs":null,"finish_reason":null}]}
{"id":"chatcmpl-xxx","object":"chat.completion.chunk","created":1718772336,"model":"my-chat-model","choices":[{"index":0,"delta":{"content":"你好"},"logprobs":null,"finish_reason":null}]}
{"id":"chatcmpl-xxx","object":"chat.completion.chunk","created":1718772336,"model":"my-chat-model","choices":[{"index":0,"delta":{"content":"，"},"logprobs":null,"finish_reason":null}]}
{"id":"chatcmpl-xxx","object":"chat.completion.chunk","created":1718772336,"model":"my-chat-model","choices":[{"index":0,"delta":{"content":"有"},"logprobs":null,"finish_reason":null}]}
{"id":"chatcmpl-xxx","object":"chat.completion.chunk","created":1718772336,"model":"my-chat-model","choices":[{"index":0,"delta":{"content":"什么"},"logprobs":null,"finish_reason":null}]}
{"id":"chatcmpl-xxx","object":"chat.completion.chunk","created":1718772336,"model":"my-chat-model","choices":[{"index":0,"delta":{"content":"我"},"logprobs":null,"finish_reason":null}]}
{"id":"chatcmpl-xxx","object":"chat.completion.chunk","created":1718772336,"model":"my-chat-model","choices":[{"index":0,"delta":{"content":"可以"},"logprobs":null,"finish_reason":null}]}
{"id":"chatcmpl-xxx","object":"chat.completion.chunk","created":1718772336,"model":"my-chat-model","choices":[{"index":0,"delta":{"content":"帮助"},"logprobs":null,"finish_reason":null}]}
{"id":"chatcmpl-xxx","object":"chat.completion.chunk","created":1718772336,"model":"my-chat-model","choices":[{"index":0,"delta":{"content":"你"},"logprobs":null,"finish_reason":null}]}
{"id":"chatcmpl-xxx","object":"chat.completion.chunk","created":1718772336,"model":"my-chat-model","choices":[{"index":0,"delta":{"content":"的"},"logprobs":null,"finish_reason":null}]}
{"id":"chatcmpl-xxx","object":"chat.completion.chunk","created":1718772336,"model":"my-chat-model","choices":[{"index":0,"delta":{"content":"吗"},"logprobs":null,"finish_reason":null}]}
{"id":"chatcmpl-xxx","object":"chat.completion.chunk","created":1718772336,"model":"my-chat-model","choices":[{"index":0,"delta":{"content":"?"},"logprobs":null,"finish_reason":null}]}
{"id":"chatcmpl-
```

xxx","object":"chat.completion.chunk","created":1718772336,"model":"my-chat-model","choices":[{"index":0,"delta":{},"logprobs":null,"finish_reason":"stop"}]}

- 非流式响应示例


```
{
   "id": "chatcmpl-xxx",
   "object": "chat.completion",
   "created": 1718772336,
   "model": "my-chat-model",
   "choices": [
      {
         "index": 0,
         "message": {
            "role": "assistant",
            "content": "你好，有什么我可以帮助你的吗？"
         },
         "finish_reason": "stop",
         "logprobs": null
      }
   ],
   "usage": {
      "prompt_tokens": 5,
      "completion_tokens": 10,
      "total_tokens": 15
   }
}
```


文本向量化（Embeddings）API 规范

接口格式

类型：POST

协议：HTTP/HTTPS

请求体参数

表 6-23 请求体参数

参数	是否必选	参数类型	描述
input	是	Array of strings	• 纯文本（string），例如，"你好" 。 • 文本列表（array），例如，["你","好"] 。 数组长度：1-2048。
model	是	String	向量化模型名称。

请求示例：


```
{
   "model": "my-embedding-model",
   "input": "你好"
}
```


响应体参数

表 6-24 响应体参数

参数	参数类型	描述
data	Array of 表 6-25 objects	向量化结果。
model	String	向量化模型名称。
object	String	固定值“list”。
usage	表6-26 object	每次请求的用量统计。

表 6-25 Embedding

参数	参数类型	描述
index	Integer	向量在向量列表中的排序。
embedding	Array of numbers	向量数组。Float类型。
object	String	固定值“embedding”。

表 6-26 usage

参数	参数类型	描述
prompt_tokens	Integer	提问包含的token数。
total_tokens	Integer	提问包含的token数。

响应示例：

```
{
  "data": [
    {
      "index": 0,
      "embedding": [
        0.02513289265334606,
        -0.017512470483779907,
        -0.029955564066767693,
        ...
      ],
      "object": "embedding"
    }
  ],
  "usage": {
    "prompt_tokens": 5,
    "total_tokens": 5
  },
  "model": "my-embedding-model",
```

```
   "object": "list"
}
```

文本排序（Rerank）API 规范

接口格式

类型：POST

协议：HTTP/HTTPS

请求体参数

表 6-27 请求体参数

参数	是否必选	参数类型	描述
query	是	String	原始请求问题，基于该问题对候选文本进行排序。
top_n	是	Integer	返回排序靠前的n个结果。
docs	是	Array of strings	候选文本，文件大小限制为512MB以内。
model	是	String	排序模型名称。

请求示例：

```
{
   "model": "my-rerank-model",
   "query": "请问AI原生应用引擎提供了什么能力？",
   "docs": ["AI原生应用引擎提供了应用开发、模型网关等能力。", "AI原生应用引擎正在逐步完善、提高竞争力。"],
   "top_n": 3
}
```

响应体参数

表 6-28 响应体参数

参数	参数类型	描述
model	String	排序模型名称。
usage	表6-29 object	每次请求的用量统计。
results	Array of 表6-30 objects	排序结果。

表 6-29 usage

参数	参数类型	描述
prompt_tokens	Integer	提问包含的token数。
total_tokens	Integer	提问包含的token数。

表 6-30 RankDocument

参数	参数类型	描述
index	Integer	文本排序后对应的序号。
document	表6-31 object	文本。
relevance_score	Number	文本的排序分数。

表 6-31 Document

参数	参数类型	描述
text	String	文本内容。

响应示例：

```
{
   "model": "my-rerank-model",
   "usage": {
      "prompt_tokens": 5,
      "total_tokens": 5
   },
"results": [
    {
        "index": 0,
        "document": {"text": "AI原生应用引擎提供了应用开发、模型网关等能力。"},
        "relevance_score": 0.9
    },
    {
        "index": 1,
        "document": {"text": "AI原生应用引擎正在逐步完善、提高竞争力。"},
        "relevance_score": 0.5
    }
  ]
}
```

图像理解 API 规范

接口格式

类型：POST

协议：HTTP/HTTPS

请求体参数

表 6-32 请求体参数

参数	是否必选	参数类型	描述
messages	是	Array of **表 6-33** objects	图像理解对话消息体类。
model	是	String	图像理解对话使用的模型名称。

参数	是否必选	参数类型	描述
frequency_penalty	否	Number	频率惩罚，会根据文本中新Token的出现频率对其进行惩罚，从而降低模型重复相同内容的可能性。使其生成的文本更加自然和符合预期。取值范围为-2.0~2.0。 ● 默认值（0.0）：不施加任何频率惩罚。模型按原本的概率分布生成文本。 ● 正值（例如 0.5，1.0，2.0）：增加惩罚力度。值越高，模型越不愿意使用已经用过的词。使输出文本的词汇更多样化、更富有创造性，但过高的值可能导致用词生僻、语句不通顺甚至偏离主题。 ● 负值（例如 -0.5，-1.0，-2.0）：减少惩罚，值越低（负的越多），模型越倾向于使用已经用过的词。使输出文本的词汇更集中、更稳定、更可能重复关键主题词。但过低的值会导致用词极其重复、啰嗦。 例如： 提示词为（Prompt）：“写一首关于猫的诗。” ● frequency_penalty=0（默认）：输出可能正常地重复使用“猫”、“尾巴”、“柔软”等合理词汇。 ● frequency_penalty=1.5（高惩罚）：模型会极力避免重复用词。第一句用了“猫”，第二句可能会用“毛茸伙伴”、“喵星人”、“优雅的生物”等同义词来替代，词汇非常丰富。但如果惩罚过高，可能会为了规避重复而选用不合适的词，导致诗歌变得奇怪。 ● frequency_penalty=-1.0（负惩罚）：模型不害怕重复，甚至鼓励重复。输出可能会变成：“猫，猫，可爱的

参数	是否必选	参数类型	描述
			猫。猫在跑，猫在跳，猫的尾巴摇啊摇。” 显得非常冗余和缺乏创意。
logprobs	否	boolean	是否返回输出Token的对数概率。
top_logprobs	否	Integer	指定在每一步生成时，返回模型最大概率的候选Token个数。 取值范围：[0,5] 仅当**logprobs**为true时生效。
max_tokens	否	Integer	返回体允许的最大token数。
presence_penalty	否	Number	存在惩罚，会根据文本中新Token是否出现对其进行惩罚，核心作用是降低模型再次讨论已经出现过的“话题”的可能性，从而增加模型谈论新主题的可能性。使其生成的文本更加自然和符合预期。取值范围为-2.0~2.0。 ● 默认值（0.0）：不施加任何存在惩罚。 ● 正值（例如 0.5，1.0，2.0）：增加惩罚力度，值越高，模型越不愿意停留在已经提及的主题上，越倾向于引入全新的想法、概念或话题。使对话或文本更容易“跑题”或转向新方向。在创意生成中，这可以带来更大的探索性。 ● 负值（例如 -0.5，-1.0，-2.0）：减少惩罚，值越低（负的越多），模型越倾向于围绕已经出现的主题进行深入讨论，避免引入新信息，使输出内容更加集中、紧扣主题，但可能显得缺乏发散性。
n	否	Integer	生成响应的个数。取值范围是1-4。 对于需要生成多个响应的场景（如创意写作、广告文案等），可以设置较大的n值。 默认值为1。

参数	是否必选	参数类型	描述
stream	否	Boolean	布尔类型。 ● 设为true时，返回结果为流式。 流式处理是一种边接收边处理、实时性强的数据处理模式。它将数据视为连续不断的“流”，允许低延迟和即时响应，广泛应用于视频播放、实时监控、大数据分析和人工智能生成内容等领域。 ● 设为false时，返回结果为非流式，JSON格式结构化数据。 非流式指数据或操作不是连续、实时地传输或处理，而是一次性接收完整的输入，待完全处理后一次性返回完整的结果。这种模式常见于传统HTTP请求-响应模式、大模型API的非流式输出和语音合成等领域，强调数据的完整性和整体性，不追求即时反馈。 缺省值：false
seed	否	Integer	设置seed参数会使文本生成过程更具有确定性，通常用于使模型每次运行的结果一致。 在每次模型调用时传入相同的seed值（由您指定），并保持其他参数不变，模型将尽可能返回相同的结果。 取值范围：0到2^{31}-1。
temperature	否	Number	较高的数值会使输出更加随机，而较低的数值会使其更加集中和确定。 ● 最小值：0 ● 最大值：2 ● 缺省值：1
top_p	否	Number	影响输出文本的多样性，取值越大，生成文本的多样性越强。 ● 最小值：0.0 ● 最大值：1.0 ● 缺省值：1

表 6-33 ChatCompletionRequestMessage

参数	是否必选	参数类型	描述
role	是	String	消息体对应的角色。 ● system：如果是系统，则为system。 ● user：如果是用户，则为user。
content	是	String	消息具体内容。
name	否	String	对话参与者的可选名称，提供给模型信息以区分相同角色的不同对话参与者。

- **工具调用请求示例**
  - 流式请求示例

```
{
   "model": "model-img2text",
   "messages": [
      {
         "role": "system",
         "content": " You are a helpful assistant. "
      },
      {
         "role": "user",
         "content": [
            {
               "type": "text",
               "text": "图里面有什么"
            },
            {
               "type": "image_url",
               "image_url": {
                  "url": "一个图片链接"
               }
            }
         ],
      }
   ],
   "tools": [
      {
         "type": "function",
         "function": {
            "name": "get_weather",
            "description": "获取给定地点的天气",
            "parameters": {
               "type": "object",
               "properties": {
                  "location": {
                     "type": "string",
                     "description": "地点，例如北京、上海。"
                  }
               },
               "required": [
                  "location"
               ]
```

```
            }
         }
      }
   ],
   "max_tokens": 20,
   "presence_penalty": 1.2,
   "frequency_penalty": 1.0,
   "temperature": 0.5,
   "top_p": 0.95,
   "stream": true
}
```

- 非流式请求示例

```
{
   "model": "model-img2text",
   "messages": [
      {
         "role": "system",
         "content": " You are a helpful assistant. "
      },
      {
         "role": "user",
         "content": [
            {
               "type": "text",
               "text": "图里面有什么"
            },
            {
               "type": "image_url",
               "image_url": {
                  "url": "一个图片链接"
               }
            }
         ],
      }
   ],
   "tools": [
      {
         "type": "function",
         "function": {
            "name": "get_weather",
            "description": "获取给定地点的天气",
            "parameters": {
               "type": "object",
               "properties": {
                  "location": {
                     "type": "string",
                     "description": "地点，例如北京、上海。"
                  }
               },
               "required": [
                  "location"
               ]
            }
         }
      }
   ],
   "max_tokens": 20,
   "presence_penalty": 1.2,
   "frequency_penalty": 1.0,
   "temperature": 0.5,
   "top_p": 0.95,
```



```
    "stream": false
}
```

- **非工具调用请求示例**
  - 流式请求示例

```
{
    "model": "model-img2text",
    "messages": [
        {
            "role": "system",
            "content": " You are a helpful assistant. "
        },
        {
            "role": "user",
            "content": [ { "type": "text", "text": "图里面有什么" }, { "type": "image_url",
"image_url": { "url": "一个图片链接" } } ],
        }
    ],
    "max_tokens": 20,
    "presence_penalty": 1.2,
    "frequency_penalty": 1.0,
    "temperature": 0.5,
    "top_p": 0.95,
    "stream": true
}
```

  - 非流式请求示例

```
{
    "model": "model-img2text",
    "messages": [
        {
            "role": "system",
            "content": " You are a helpful assistant. "
        },
        {
            "role": "user",
            "content": [ { "type": "text", "text": "图里面有什么" }, { "type": "image_url",
"image_url": { "url": "一个图片链接" } } ],
        }
    ],
    "max_tokens": 20,
    "presence_penalty": 1.2,
    "frequency_penalty": 1.0,
    "temperature": 0.5,
    "top_p": 0.95,
    "stream": false
}
```

响应体参数

表 6-34 响应体参数

参数	参数类型	描述
id	String	图像理解文本对话唯一标识符。
choices	Array of **表 6-35** objects	返回体列表。 如果“n”大于1，则结果为多个。
created	long	问答发生的时间。格式为时间戳。

参数	参数类型	描述
model	String	图像理解文本对话使用的模型名称。
object	String	固定值“chat.completion”。
usage	表6-39 object	图像理解文本对话用量统计。

表 6-35 ChatNonStreamingChoice

参数	参数类型	描述
index	Integer	返回多个choices时，每个choice对应的顺序。
message	表6-36 object	模型服务返回的具体消息体内容。
finish_reason	String	返回结束的原因。 • stop：模型达到自然停止点或提供的停止序列。 • length：达到请求中指定的最大令牌数。 • content_filter：由于内容过滤器的标志而省略了内容。 • tool_calls：模型选择了某个工具。

表 6-36 ChatMessageResponse

参数	参数类型	描述
content	String	返回消息体的内容，与tool_calls二选一。
role	String	返回消息体的角色。 • user：用户输入的问题。 • assistant：大模型的答复内容。
tool_calls	Array of 表6-37 objects	工具调用消息，与content二选一。
audio	ChatMessageAudio	聊天信息中的音频部分。
reasoningContent	String	用于展示模型的推理过程，帮助用户理解模型的决策依据。

表 6-37 ToolCall

参数	参数类型	描述
id	String	工具调用唯一标识符。

参数	参数类型	描述
type	String	工具类型，当前仅支持function。
function	表6-38 Object	调用函数的详细信息。

表 6-38 CallFunction

参数	参数类型	描述
name	String	函数名。
arguments	String	调用函数的参数，JSON格式。

表 6-39 CompletionUsage

参数	参数类型	描述
completion_tokens	Integer	回答包含的token数。
prompt_tokens	Integer	提问包含的token数。
total_tokens	Integer	提问+回答token总数。

- **工具调用响应示例**
  - 流式响应示例

```
data:{"created":1767494144,"model":"qwen2.5-vl-72b","id":"chatcmpl-417592f9469edb89448e18ce650333ae","choices":[{"delta":{"role":"assistant","content":""},"index":0}],"request_id":"74fb6469180f3e516732b36e3506e5f7","object":"chat.completion.chunk"}
data:{"created":1767494144,"model":"qwen2.5-vl-72b","id":"chatcmpl-417592f9469edb89448e18ce650333ae","choices":[{"delta":{"tool_calls":[{"id":"call_abc123","type":"function","function":{"name":"search_web","arguments":{}}}],"content":""},"index":0}],"request_id":"74fb6469180f3e516732b36e3506e5f7","object":"chat.completion.chunk"}
data:{"created":1767494144,"model":"qwen2.5-vl-72b","id":"chatcmpl-417592f9469edb89448e18ce650333ae","choices":[{"delta":{"tool_calls":[{"id":"call_abc123","type":"function","function":{"name":"search_web","arguments":{"query":"最新款iPhone发布信息","limit":5}}}],"content":""},"index":0}],"request_id":"74fb6469180f3e516732b36e3506e5f7","object":"chat.completion.chunk"}
data:{"created":1767494144,"model":"qwen2.5-vl-72b","id":"chatcmpl-417592f9469edb89448e18ce650333ae","choices":[{"delta":{"tool_calls":[{"id":"call_abc123","type":"function","function":{"name":"search_web","arguments":{"query":"最新款iPhone发布信息","limit":5}},"index":0}],"content":""},"index":0}],"request_id":"74fb6469180f3e516732b36e3506e5f7","object":"chat.completion.chunk"}
data:{"created":1767494144,"model":"qwen2.5-vl-72b","id":"chatcmpl-417592f9469edb89448e18ce650333ae","choices":[{"delta":{"tool_calls":[],"content":"正在为您搜索最新款iPhone的发布信息……
```

"},"index":0}],"request_id":"74fb6469180f3e516732b36e3506e5f7","object":"chat.completion.chunk"}
data:{"created":1767494144,"model":"qwen2.5-vl-72b","id":"chatcmpl-417592f9469edb89448e18ce650333ae","choices":[{"finish_reason":"tool_calls","delta":{"tool_calls":[],"content":""},"index":0}],"request_id":"74fb6469180f3e516732b36e3506e5f7","object":"chat.completion.chunk"}
data:{"created":1767494144,"usage":{"completion_tokens":67,"prompt_tokens":35,"total_tokens":102},"model":"qwen2.5-vl-72b","id":"chatcmpl-417592f9469edb89448e18ce650333ae","choices":[],"request_id":"74fb6469180f3e516732b36e3506e5f7","object":"chat.completion.chunk"}
data:[DONE]

- 非流式响应示例

{
  "choices": [
    {
      "message": {
        "role": "assistant",
        "tool_calls": [
          {
            "id": "call_12345",
            "type": "function",
            "function": {
              "name": "image_analysis",
              "arguments": "{\"detail_level\": \"high\", \"output_format\": \"json\"}"
            }
          }
        ]
      },
      "finish_reason": "tool_calls",
      "index": 0
    }
  ],
  "created": 1753965925754,
  "id": "model-img2text",
  "object": "chat.completions",
  "usage": {
    "prompt_tokens": 142,
    "completion_tokens": 28,
    "total_tokens": 170
  },
  "request_id": "xxx"
}

- **非工具调用响应示例**
  - 流式响应示例

data:{"created":1767494144,"model":"qwen2.5-vl-72b","id":"chatcmpl-417592f9469edb89448e18ce650333ae","choices":[{"delta":{"role":"assistant","content":""},"index":0}],"request_id":"74fb6469180f3e516732b36e3506e5f7","object":"chat.completion.chunk"}
data:{"created":1767494144,"model":"qwen2.5-vl-72b","id":"chatcmpl-417592f9469edb89448e18ce650333ae","choices":[{"delta":{"tool_calls":[],"content":"这是"},"index":0}],"request_id":"74fb6469180f3e516732b36e3506e5f7","object":"chat.completion.chunk"}
data:{"created":1767494144,"model":"qwen2.5-vl-72b","id":"chatcmpl-417592f9469edb89448e18ce650333ae","choices":[{"delta":{"tool_calls":[],"content":"Google"},"index":0}],"request_id":"74fb6469180f3e516732b36e3506e5f7","object":"chat.completion.chunk"}

data:{"created":1767494144,"model":"qwen2.5-vl-72b","id":"chatcmpl-417592f9469edb89448e18ce650333ae","choices":[{"delta":{"tool_calls":[],"content":"Chrome"},"index":0}],"request_id":"74fb6469180f3e516732b36e3506e5f7","object":"chat.completion.chunk"}
data:{"created":1767494144,"model":"qwen2.5-vl-72b","id":"chatcmpl-417592f9469edb89448e18ce650333ae","choices":[{"delta":{"tool_calls":[],"content":"浏览器"},"index":0}],"request_id":"74fb6469180f3e516732b36e3506e5f7","object":"chat.completion.chunk"}
data:{"created":1767494144,"model":"qwen2.5-vl-72b","id":"chatcmpl-417592f9469edb89448e18ce650333ae","choices":[{"delta":{"tool_calls":[],"content":"的"},"index":0}],"request_id":"74fb6469180f3e516732b36e3506e5f7","object":"chat.completion.chunk"}
data:{"created":1767494144,"model":"qwen2.5-vl-72b","id":"chatcmpl-417592f9469edb89448e18ce650333ae","choices":[{"delta":{"tool_calls":[],"content":"图标"},"index":0}],"request_id":"74fb6469180f3e516732b36e3506e5f7","object":"chat.completion.chunk"}
data:{"created":1767494144,"model":"qwen2.5-vl-72b","id":"chatcmpl-417592f9469edb89448e18ce650333ae","choices":[{"delta":{"tool_calls":[],"content":"。"},"index":0}],"request_id":"74fb6469180f3e516732b36e3506e5f7","object":"chat.completion.chunk"}
data:{"created":1767494144,"model":"qwen2.5-vl-72b","id":"chatcmpl-417592f9469edb89448e18ce650333ae","choices":[{"finish_reason":"stop","delta":{"tool_calls":[],"content":""},"index":0}],"request_id":"74fb6469180f3e516732b36e3506e5f7","object":"chat.completion.chunk"}
data:{"created":1767494144,"usage":{"completion_tokens":43,"prompt_tokens":35,"total_tokens":78},"model":"qwen2.5-vl-72b","id":"chatcmpl-417592f9469edb89448e18ce650333ae","choices":[],"request_id":"74fb6469180f3e516732b36e3506e5f7","object":"chat.completion.chunk"}
data:[DONE]

- 非流式响应示例

{
  "choices": [
    {
      "delta": {
        "role": "assistant",
        "content": "图像整体呈现出简洁、抽象的风格，主要内容是一个灰色的圆形头像轮廓。"
      },
      "finish_reason": "stop",
      "index": 0
    }
  ],
  "created": 1753965925754,
  "id": "model-img2text",
  "object": "chat.completions",
  "usage": {
    "prompt_tokens": 142,
    "completion_tokens": 118,
    "total_tokens": 260
  },
  "request_id": "xxx"
}

6.3.5 调测已接入的模型服务

模型调测是指通过对模型进行实际操作、参数调整及效果观测，以验证其在特定场景下的功能表现、性能指标及适用范围的过程，其核心目的是确保模型在真实业务场景中能够稳定、高效地运行。本章介绍用户自定义接入的模型服务调测流程。

前提条件

- 已**购买Versatile智能体平台**。
- 已**接入自定义的模型服务**。
- 登录用户为空间所有者、空间管理员、开发工程师、运维工程师，详细信息请参考**管理团队空间成员**。

调测模型服务

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航，选择“模型中心 > 模型服务”，进入“模型服务”页面。

步骤3 选择“自定义”页签，在模型供应商列表，单击模型供应商卡片。

步骤4 在“供应商详情”页面，在需要调测的模型服务卡片上，单击“ ··· > 调测”。

图 6-19 调测模型服务

步骤5 在“模型调测”页面，可以调测如下几种类型的模型服务。

- **文本对话**
  a. 在“模型类型”区域选择“文本对话”，参数配置请参考**表6-40**。

图 6-20 文本对话

表 6-40 文本对话类型模型参数说明

参数	说明	示例
模型服务	“模型服务A”默认展示所选的供应商模型服务。“模型服务B”为可选项。 您也可以在下拉列表选择或切换以下模型服务： ■ **用户自主接入的模型服务**：以模型供应商维度展示。 ■ **平台推荐**：以模型供应商维度展示。 ■ **路由策略**：用户自定义创建的路由策略。	DeepSeek-V3
深度思考	显示该参数有以下两个场景： ■ **平台推荐**：当选择的模型服务为思考模型且支持关闭深度思考时，才显示此参数，例如平台推荐的Qwen3-32B、DeepSeek-V3.2。 ■ **用户自主接入的模型服务**：当选择的模型服务为思考模型且在新建模型服务开启了“是否支持关闭思维链输出”时，才显示此参数。 该参数支持以下操作： ■ 当此功能**开启**时，大模型将首先进行深入的思考和推理，通过逐步拆解问题、梳理逻辑，生成一段详细的思维链内容，并在调试界面展示。这一过程有助于提升最终输出答案的准确性和可靠性，确保用户获得更加精准的信息。 ■ 当此功能**关闭**时，智能体将直接生成最终答案，不再经过额外的思维链推理过程。这将加快响应速度，适用于需要快速获取答案的场景。 **注意** 在模型使用过程中，“深度思考”开关生效的情况如下： ■ 如果模型支持思维链输出能力，并且也支持关闭该能力，则开启、关闭均生效。 ■ 如果模型支持思维链输出能力，但不支持关闭该能力，则开启生效、关闭不生效。 ■ 如果模型不支持思维链输出能力，则开启、关闭均不生效。	开启

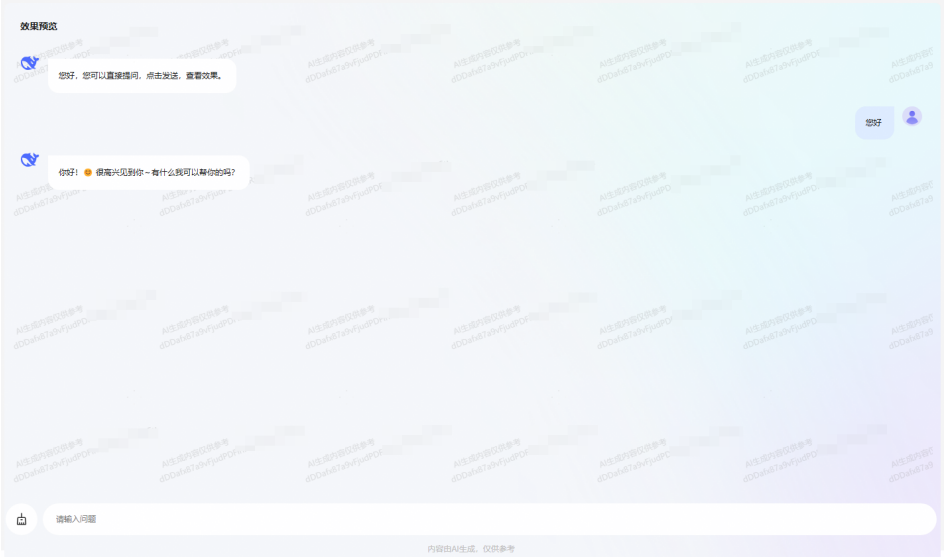
参数	说明	示例
输出方式	▪ **非流式**：调用大语言模型推理服务时，根据用户问题，获取大语言模型的回答，大语言模型完整生成回答后一次性返回。 ▪ **流式**：调用大语言模型推理服务时，根据用户问题，获取大语言模型的回答，逐个字词的快速返回模式，不需等待大语言模型生成完成。默认**流式**。	流式
输出最大token数	模型在单次推理或生成内容时，能够输出的token（模型处理文本的基本单位）数量的最大值。取值范围100~32768，默认值为2048。	2048
温度	较高的数值会使输出更加随机，而较低的数值会使其更加集中和确定。取值范围0.01~2，默认值为0.5。 建议该参数和“多样性”只设置1个。	0.5
多样性	影响输出文本的多样性，取值越大，生成文本的多样性越强。取值范围0~1，默认值为0.5。 建议该参数和“温度”只设置1个。	0.5
存在惩罚	正值会尽量避免使用已出现过的词语，更倾向于生成新词语。取值范围-2.0~2.0，默认值为0。	0
频率惩罚	正值会尽量避免使用常见的单词和短语，更倾向于生成较少见的单词。取值范围-2.0~2.0，默认值为0。	0

b. 在右侧“效果预览”区域，在对话输入框输入测试语句后按Enter键或单击  ，查看模型响应结果。

单击  ，清除本次会话内容，可以开始新的会话。

调测成功后，可以在智能体、工作流中使用模型服务，请参考**开发单智能体应用**、**开发工作流应用**、**开发多智能体应用**。

图 6-21 文本对话模型调测成功



- **图像理解**

  a. 在“模型类型”区域选择“图像理解”，参数配置请参考**表6-41**。

图 6-22 图像理解

表 6-41 图像理解类型模型参数说明

参数	说明	示例
模型服务	默认展示所选的供应商模型服务。 您也可以在下拉列表切换以下模型服务： ■ **用户自主接入的模型服务**：以模型供应商维度展示。 ■ **平台推荐**：以模型供应商维度展示。	Qwen2.5-VL-72B
深度思考	显示该参数有以下两个场景： ■ **平台推荐**：当选择的模型服务为思考模型且支持关闭深度思考时，才显示此参数，例如平台推荐的Qwen3-32B、DeepSeek-V3.2。 ■ **用户自主接入的模型服务**：当选择的模型服务为思考模型且在新建模型服务开启了“是否支持关闭思维链输出”时，才显示此参数。 该参数支持以下操作： ■ 当此功能**开启**时，大模型将首先进行深入的思考和推理，通过逐步拆解问题、梳理逻辑，生成一段详细的思维链内容，并在调试界面展示。这一过程有助于提升最终输出答案的准确性和可靠性，确保用户获得更加精准的信息。 ■ 当此功能**关闭**时，智能体将直接生成最终答案，不再经过额外的思维链推理过程。这将加快响应速度，适用于需要快速获取答案的场景。 **注意** 在模型使用过程中，“深度思考”开关生效的情况如下： ■ 如果模型支持思维链输出能力，并且也支持关闭该能力，则开启、关闭均生效。 ■ 如果模型支持思维链输出能力，但不支持关闭该能力，则开启生效、关闭不生效。 ■ 如果模型不支持思维链输出能力，则开启、关闭均不生效。	开启

参数	说明	示例
输出方式	▪ **非流式**：调用大语言模型推理服务时，根据用户问题，获取大语言模型的回答，大语言模型完整生成回答后一次性返回。默认**非流式**。 ▪ **流式**：调用大语言模型推理服务时，根据用户问题，获取大语言模型的回答，逐个字词的快速返回模式，不需等待大语言模型生成完成。	非流式
上传图片	单击，可上传本地图片。支持上传JPG、PNG格式图片，且不大于4MB。	-
提示语内容	输入提示语，对图片进行提问。	图片里有什么?

b. 单击“生成图像理解”，在右侧“效果预览”区域查看模型响应效果。

调测成功后，可以在智能体、工作流中使用模型服务，请参考**开发单智能体应用**、**开发工作流应用**、**开发多智能体应用**。

图 6-23 图像理解模型调测成功

- **文本向量化**

  a. 在“模型类型”区域选择“文本向量化”，参数配置请参考**表6-42**。

图 6-24 文本向量化

表 6-42 文本向量化类型模型参数说明

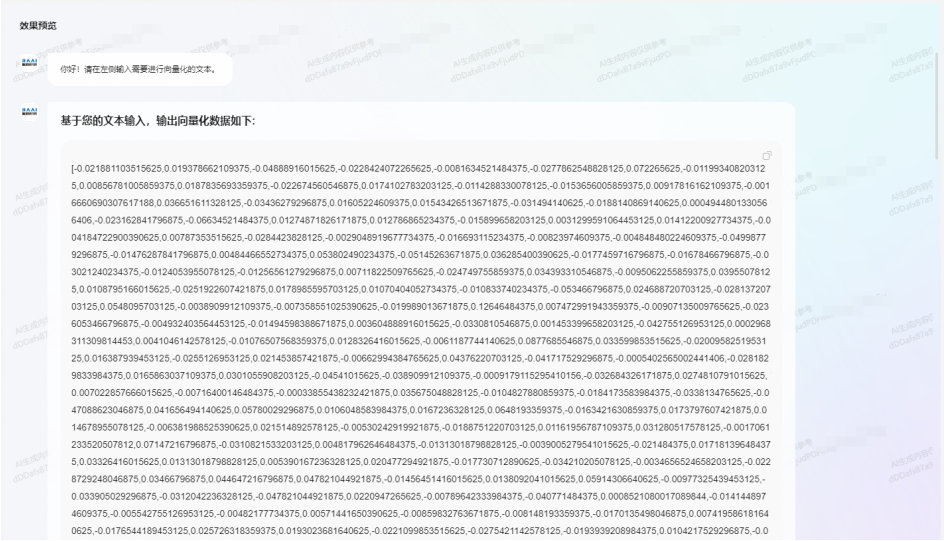
参数	说明	示例
模型服务	默认展示所选的供应商模型服务。 您也可以在下拉列表切换以下模型服务： ■ **用户自主接入的模型服务**：以模型供应商维度展示。 ■ **平台推荐**，以模型供应商维度展示。	BGE-M3

参数	说明	示例
深度思考	显示该参数有以下两个场景： ■ **平台推荐**：当选择的模型服务为思考模型且支持关闭深度思考时，才显示此参数，例如平台推荐的Qwen3-32B、DeepSeek-V3.2。 ■ **用户自主接入的模型服务**：当选择的模型服务为思考模型且在新建模型服务开启了“是否支持关闭思维链输出”时，才显示此参数。 该参数支持以下操作： ■ 当此功能**开启**时，大模型将首先进行深入的思考和推理，通过逐步拆解问题、梳理逻辑，生成一段详细的思维链内容，并在调试界面展示。这一过程有助于提升最终输出答案的准确性和可靠性，确保用户获得更加精准的信息。 ■ 当此功能**关闭**时，智能体将直接生成最终答案，不再经过额外的思维链推理过程。这将加快响应速度，适用于需要快速获取答案的场景。 **注意** 在模型使用过程中，“深度思考”开关生效的情况如下： ■ 如果模型支持思维链输出能力，并且也支持关闭该能力，则开启、关闭均生效。 ■ 如果模型支持思维链输出能力，但不支持关闭该能力，则开启生效、关闭不生效。 ■ 如果模型不支持思维链输出能力，则开启、关闭均不生效。	关闭
请输入文本	输入待向量化的文本，可参照以下示例： ■ 示例1：那是个快乐的人 ■ 示例2： ["那是个快乐的人", "那是个高兴的人", "那是个忧郁的人"]	那是个快乐的人

b. 单击“生成向量化”，在右侧“效果预览”区域查看模型响应效果。

调测成功后，可以在智能体、工作流中使用模型服务，请参考**开发单智能体应用**、**开发工作流应用**、**开发多智能体应用**。

图 6-25 文本向量化调测成功



- **文本排序**

  a. 在“模型类型”区域选择“文本排序”，参数配置请参考**表6-43**。

图 6-26 文本排序

表 6-43 文本排序类型模型参数说明

参数名称	参数说明	示例
模型服务	默认展示所选的模型服务。 您也可以在下拉列表切换以下模型服务： ■ **用户自主接入的模型服务**：以模型供应商维度展示。 ■ **平台推荐**，以模型供应商维度展示。	BGE-Reranker-V2-M3
深度思考	显示该参数有以下两个场景： ■ **平台推荐**：当选择的模型服务为思考模型且支持关闭深度思考时，才显示此参数，例如平台推荐的Qwen3-32B、DeepSeek-V3.2。 ■ **用户自主接入的模型服务**：当选择的模型服务为思考模型且在新建模型服务开启了“是否支持关闭思维链输出”时，才显示此参数。 该参数支持以下操作： ■ 当此功能**开启**时，大模型将首先进行深入的思考和推理，通过逐步拆解问题、梳理逻辑，生成一段详细的思维链内容，并在调试界面展示。这一过程有助于提升最终输出答案的准确性和可靠性，确保用户获得更加精准的信息。 ■ 当此功能**关闭**时，智能体将直接生成最终答案，不再经过额外的思维链推理过程。这将加快响应速度，适用于需要快速获取答案的场景。 **注意** 在模型使用过程中，“深度思考”开关生效的情况如下： ■ 如果模型支持思维链输出能力，并且也支持关闭该能力，则开启、关闭均生效。 ■ 如果模型支持思维链输出能力，但不支持关闭该能力，则开启生效、关闭不生效。 ■ 如果模型不支持思维链输出能力，则开启、关闭均不生效。	关闭
待排序文本	输入待排序文本。单击 + 添加文本，最多可以添加10条。	小朋友在学校很快乐
被展示文本条数	文本排序完成后，展示的条数。取值范围1~10，默认值为1。	3

参数名称	参数说明	示例
我的问题	描述想要解决的问题。	小朋友在学校怎么样?

b. 单击“开始排序”，在右侧“效果预览”区域查看模型响应效果。

调测成功后，可以在智能体、工作流中使用模型服务，请参考**开发单智能体应用**、**开发工作流应用**、**开发多智能体应用**。

图 6-27 文本排序调测成功

----结束

6.4 配置模型服务路由策略

通过设置路由策略，可以实现模型故障自动切换功能。当模型A因故障等原因无法正常工作时，系统会自动切换至其他可用模型，继续提供服务，从而提升模型服务的稳定性和可用性。路由策略创建完成后，可以进行调测和使用。

前提条件

- 已**购买Versatile智能体平台**。
- 已**接入自定义的模型服务**。
- 登录用户为空间所有者、空间管理员、开发工程师、运维工程师，详细信息请参考**管理团队空间成员**。

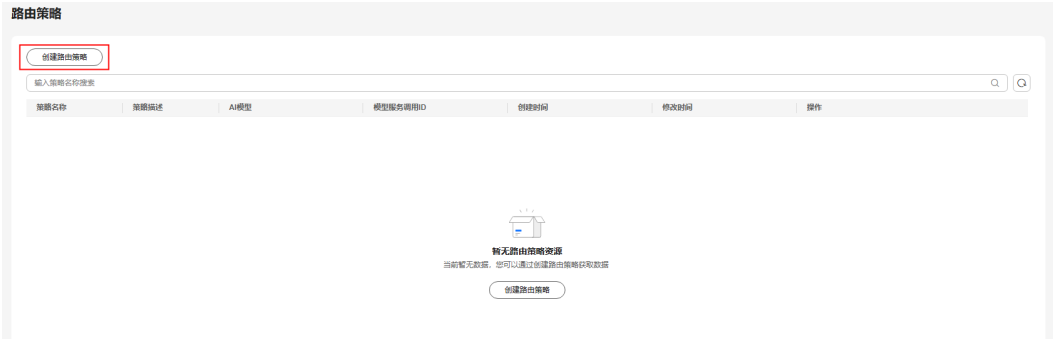
创建路由策略

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航，选择“模型中心 > 路由策略”。

步骤3 在“路由策略”页面，单击“创建路由策略”。

图 6-28 创建路由策略



步骤4 在“创建路由策略”页面，配置参数信息，具体参数说明请参考**表6-44**，配置完成后单击“保存”。

新建的路由策略，显示在路由策略列表中。

图 6-29 创建路由策略

表 6-44 路由策略参数说明

参数	说明	示例
策略名称	自定义路由策略的名称。由2~36个字符组成，包含中英文、数字、中划线（-）、下划线（_）、点（.），仅支持以中英文开头。	文本对话路由策略
AI模型	在“模型A”下拉框中选择模型服务。 单击“+ AI模型”，添加模型服务。一共支持添加3个模型服务。 路由策略提供模型服务时，模型调用顺序为：模型A > 模型B > 模型C，当模型A无法正常工作时，可以自动依次切换为模型B、模型C。	模型A：DeepSeek-R1 模型B：DeepSeek-V3 模型C：Qwen3-32B
策略总超时时间	模型路由策略的总体超时时间。取值范围为1000~1,000,000ms，默认值为10,000ms。	10000ms
模型重试次数	路由策略中单个模型服务的重试次数。取值范围为0-100次，默认值为0次。	0
策略描述	路由策略的描述信息。由1~100个字符组成。	该策略为文本对话类型的路由策略。

步骤5 在“模型调测”区域，调测模型，具体参数说明请参考**表6-45**。

图 6-30 调测模型

表 6-45 模型调测参数说明

参数名称	参数说明	示例
输出方式	可选非流式、流式。 • **非流式**：调用大语言模型推理服务时，根据用户问题，获取大语言模型的回答，大语言模型完整生成回答后一次性返回。 • **流式**：调用大语言模型推理服务时，根据用户问题，获取大语言模型的回答，逐个字词的快速返回模式，不需等待大语言模型生成完成。默认**流式**。	流式
输出最大token数	模型在单次推理或生成内容时，能够输出的token（模型处理文本的基本单位）数量的最大值。取值范围为100~32768，默认值为2048。	2048

参数名称	参数说明	示例
温度	较高的数值会使输出更加随机，而较低的数值会使其更加集中和确定。取值范围0.01~2，默认值为0.5。 建议该参数和“多样性”只设置1个。	0.5
多样性	影响输出文本的多样性，取值越大，生成文本的多样性越强。取值范围0~1，默认值为0.5。 建议该参数和“温度”只设置1个。	0.5
存在惩罚	正值会尽量避免使用已出现过的词语，更倾向于生成新词语。取值范围-2.0~2.0，默认值为0。	0
频率惩罚	正值会尽量避免使用常见的单词和短语，更倾向于生成较少见的单词。取值范围-2.0~2.0，默认值为0。	0

步骤6 在右侧“预览调试”区域，在对话输入框输入测试语句后按Enter键或单击，查看模型响应结果。

单击，清除本次会话内容，可以开始新的会话。

图 6-31 预览调试

----结束

相关操作

在“路由策略”列表，支持的其他操作请参考**表6-46**。

表 6-46 相关操作

操作	说明
查看路由策略详情	在待查看的路由策略对应的“策略名称”列下，单击路由策略名称。
修改路由策略	在待修改的路由策略对应的“操作”列下，单击“编辑”。
删除路由策略	在待删除的路由策略对应的“操作”列下，单击“删除”。

相关文档

路由策略创建完成后，用户可在智能体、工作流中使用路由策略，请参考**开发单智能体应用**、**开发工作流应用**、**开发多智能体应用**。

7 开发单智能体应用

7.1 单智能体应用介绍

Versatile是一个集成盘古大模型、DeepSeek等第三方模型的智能体开发平台，提供角色设定、插件扩展、工作流编排等开发工具，支持知识库管理、RAG检索、智能提示词优化等核心功能。平台还支持多模型服务、灵活的团队空间管理、丰富的资产中心资源，以及通过API、网页等多种渠道发布应用，助力开发者高效打造专业级智能体应用。

单智能体（Single Agent）是一个独立运作的AI实体，能自主感知环境、规划决策并执行任务，全程无需其他智能体协作。其核心特点是集中化处理，适用于目标明确、复杂度较低的场景（如客服机器人、游戏NPC）。

单智能体应用适合处理简单独立任务，如果有固定的任务执行流程，高准确率要求，可以选择工作流应用，具体请参见**开发工作流应用**；如果需要处理复杂协作任务，可以选择多智能体模式，具体请参见**开发多智能体应用**。

单智能体应用预置了1个应用模板，名为“预制模版-城市文旅推荐”，占用1个套餐内的应用数量，模版支持用户修改、复制、删除等操作。

单智能体应用编排能力

表 7-1 单智能体应用编排能力

功能	说明
编排模式	支持用户对话式的快捷调用和创建Agent，让业务人员以零代码操作方式，5分钟完成1个原生AI应用创建。
模型选择	平台提供盘古大模型，并已适配包括DeepSeek在内的多款第三方模型，其中DeepSeek类深度思考模型在Versatile上已完成适配。

角色指令

Agent可通过角色指令设定拟人化特征，提升交互真实感。平台预置了角色指令模板，也可以通过智能添加的方式让大模型输出一个更合适的角色提示词，角色的设定便于开发者打造高度拟人化的交互场景。

提示词

Versatile提供prompt模板与开发工具，使任何人都可无门槛开发出高质量的prompt指令。

表 7-2 提示词功能

功能	说明
提示词撰写	撰写提示词，并将评估表现较好的提示词设为候选，更多信息，请参考**撰写提示词规范**。 • 支持导入提示词示例。 • 支持模型设置。 • 支持提示词中变量定义。 • 支持效果预览。 • 支持历史记录。
提示词比较	支持选择候选提示词进行比较，比较方式包含差异性比较、效果比较。
提示词优化	提示工程平台提供提示词自动优化功能，基于启发式算法的提示词自优化技术、提示优选梯度优化技术，可以根据评估用例自动对现有的提示词进行优化，更多信息，请参考**优化提示词**。
提示词自动生成	借助模板智能匹配与布局优化技术，根据用户输入内容，结合提示词原模板和模型能力，自动生成高质量提示词模板。
提示词使用	指令配置和工作流中的大模型、意图识别，Agent、高级意图识别、提问器节点，支持保存和引用Prompt模板，更多信息，请参考**为智能体和工作流设置提示词**。

技能

智能体的核心能力源于其技能体系，开发者可通过集成插件、设计工作流等方式不断扩展模型的功能范围。

表 7-3 智能体技能

功能	说明
插件	您可以通过API无缝连接各类平台和服务，快速扩展智能体的功能，平台提供了丰富的内置插件，开箱即用；同时也支持自定义插件开发，将任意API封装成工具，灵活调用。更多信息，参考**插件介绍**。
工作流	单智能体支持添加已发布的工作流版本应用。工作流是构建复杂功能逻辑的可视化工具，通过灵活组合多个任务节点，能够设计多步骤的自动化流程，从而显著增强智能体应对复杂任务的能力。更多信息，参考**工作流介绍**。

知识库

提供开箱即用的企业级RAG（Retrieval-Augmented Generation）服务，涵盖管理、测试和检索策略配置全部功能。

表 7-4 单智能体添加知识库

功能	说明
知识库命中测试	支持对创建的知识库进行命中测试，以评估知识库的效果和准确性。

功能	说明
知识库召回策略	• 检索策略，文档检索的方式，有三种： – 语义检索：使用向量检索技术检索，对文档及结构化数据中知识进行检索，召回与用户意图相关性高的切片内容，推荐在需要结合上下文相关性、并对用户意图理解场景中使用。 – 关键词检索：使用倒排检索技术，对文档及结构化数据中知识进行检索，召回与Query关键词匹配度高的切片内容，推荐在需要用户提问关键词匹配度高的场景中使用。 – 混合检索：使用向量检索和关键词检索两种策略混合检索知识库，推荐在需要兼顾用户意图理解及关键词匹配度场景中使用。 • 相关度阈值：超过相关度阈值的搜索结果会提交给大模型进行总结，否则被过滤，可以参考知识库中命中测试的相关度分值调整该阈值。 • topk召回数量：召回的相关性阈值top切片数量，如topk召回数量为5，则相关性阈值为前5的切片将被召回提交给大模型总结。 • FAQ直出阈值：FAQ检索超过阈值的结果将直接提交给大模型总结，不再进行文档检索。如果没有超过阈值的结果，将进行文档检索。 启用FAQ功能后，系统将优先检索FAQ数据。若未命中结果，则会继续查询切片内容，可能会带来一定的性能开销。当FAQ检索结果超过预设阈值时，将直接提交给大模型进行总结，不再进行文档检索。若未超过阈值，则将继续进行文档检索。 • 查看来源：添加知识库并开启此功能后，可以在预览调试界面中查看搜索结果的详细来源信息，包括上下文内容和文件名称。有助于更快速、准确地定位和理解搜索结果。 • 查看图片：开启后此功能后，当知识库支持图片检索时，可查看检索结果中的图片信息。

记忆

系统提供变量设置，自动识别并存储用户的行为或偏好，可在后续对话中调用这些变量，提供更贴合用户需求的回复。

⚠ 注意

记忆功能仅在模型优先模式下可用。

表 7-5 记忆说明

功能	说明
变量	用来存储用户的某一行为或偏好，在对话过程中，会自动识别与变量匹配的内容，并存储在变量中。

MCP 服务

平台工具调用支持MCP协议，开发者可以通过集成MCP服务快速扩展智能体的功能。

表 7-6 MCP 服务说明

功能	说明
MCP服务广场	平台预置了“高德地图”、“车票查询工具”、“必应搜索”等多种实用MCP服务，开通后可以一键集成调用。
自定义MCP服务	平台开放自定义MCP服务创建能力，开发者可依据MCP服务地址快速创建MCP服务。

对话体验

Agent开发对话体验支持全程可视化，通过调试功能快速定位与优化配置。

表 7-7 Versatile 对话体验

功能	说明
单智能体应用图标	Agent的品牌标识，用于视觉识别，通常体现其功能或个性。
开场白	Agent与用户交互时的初始欢迎语，设定对话基调并引导用户，支持用户自定义配置。
推荐问题	每次对话开始预设的典型提问示例，帮助用户快速了解智能体的能力范围，支持用户自定义配置。
追问	与Agent对话时主动提出的跟进问题，用于澄清需求或深化交互。
音色	支持为智能体指定音色，用于配置智能应用调试对话模型返回结果的朗读音色。
内容审核配置	通过设置关键词匹配处理输入输出内容，保障大模型内容安全。 **注意** • 审核内容输入时需要用“，”隔开。 • 内容审核和安全护栏无法同时开启，打开内容审核配置开关后，“安全防护”将自动关闭。

功能	说明
安全护栏	主要用于检测和拦截潜在的有害、敏感或攻击性的内容。具体来说，它能够识别并阻止那些旨在操纵或滥用系统的Prompt攻击，同时也能过滤掉包含病毒、不适当或违法信息的输入和输出，从而保护用户和系统免受不良影响。这一机制对于维护平台的健康环境和保障用户安全至关重要。 **注意** 内容审核和安全护栏无法同时开启，打开当前开关后，“内容审核”将自动关闭。
预览调试	实时测试和优化Agent功能的工具，展示运行的结果与调用详情。

触发器

在Agent开发过程中添加触发器，Agent会按照触发器的设置执行。

表 7-8 创建触发器参数说明

参数	说明
触发器名称	触发器的名称。 由2~20个字符组成，仅支持中英文开头，支持中英文、数字、下划线。
触发时间	按设置的时间触发智能体应用的执行。例如，设置触发时间为每1小时执行一次，则每隔1小时，重复执行一次会话。 ● 周期触发 – 用户可自定义触发时间按每日执行。 – 用户可自定义触发时间按每周执行。 – 用户可自定义触发时间按每月执行。 ● 间隔触发：用户可自定义触发间隔时间，支持间隔“天”、“小时”、“分钟”、“秒”。 – 间隔“天”：取值范围1~31。 – 间隔“小时”：取值范围1~23。 – 间隔“分钟”：取值范围1~59。 – 间隔“秒”：取值范围1~59。
机器人提示	输入一条自然语言指令，智能体将遵循该指令定时执行。例如输入“发送月底发票报销提醒”，则机器人会在设定的时间发送提醒。

发布管理

支持将智能体封装成标准化API接口，供开发者直接调用，适用于需要与其他系统（如企业内部应用、第三方服务）深度集成的场景。

Agent支持不同的发布方式和渠道，支持发布成API或发布至网站，通过这两种发布方式，Agent可快速落地到实际业务中，兼顾技术集成与终端用户体验。

表 7-9 Versatile 发布渠道

功能	说明
API	将Agent封装成标准化API接口，供开发者直接调用，适用于需要与其他系统（如企业内部应用、第三方服务）深度集成的场景。
网页	生成可独立访问的Web页面（如H5或PC端页面），用户通过浏览器即可交互，适用于直接面向终端用户的场景（如客服聊天窗口、营销导购助手）。

7.2 搭建一个医疗问诊助手智能体应用

随着人工智能技术的不断进步，大模型在医疗领域的应用逐渐成熟。通过结合医学知识库、自然语言处理和智能交互技术，医疗问诊助手智能体能够为患者提供初步的健康咨询、症状分析和诊断建议，同时减轻医生的工作负担，提升医疗服务效率。

本教程将指导你如何在Versatile智能体平台上搭建一个医疗问诊助手，用于获取健康建议。

医疗问诊助手效果展示

与医疗问诊助手进行对话时，可以模拟医生的问诊方式，逐步引导用户给出症状信息，并给出健康建议。

图 7-1 医疗问诊助手应用问答效果

前提条件

本实践选用平台预置的“DeepSeek-V3”模型，首次使用该模型服务需要先对模型进行鉴权，具体操作请参见**接入平台预置的供应商模型服务**。

创建单智能体应用

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 单击左侧导航栏“开发中心 > 应用管理 > 单智能体应用”，单击左上角“创建应用”。

步骤3 在创建页面中基础信息配置中输入应用名称、功能描述等信息，并选择“单智能体应用图标”。

图 7-2 常规创建

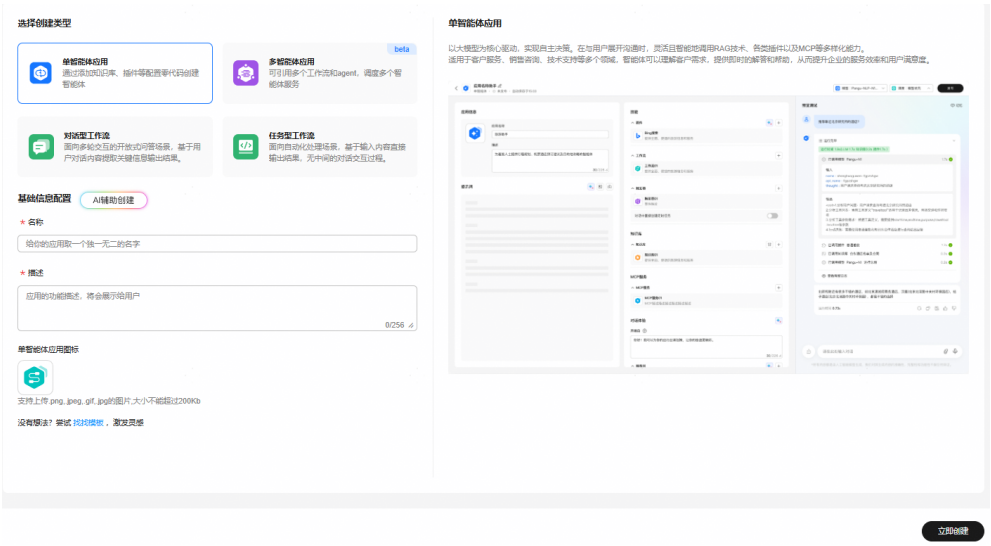


表 7-10 基础信息参数说明

参数	示例	说明
应用名称	医疗问诊助手	在单智能体应用工作台中工作流名称不允许重复，支持中英文、数字、下划线、中划线和空格，长度2~64字符，且名称首尾不能有空格。
描述	在医疗问诊助手智能体应用中，能够模拟医生的问诊流程，通过逐步对话引导用户详细描述其症状，进而提供相应的健康建议。	明确智能体的目标、功能范围以及交互等，直观展示给用户。

步骤4 单击“立即创建”进入应用编排界面。

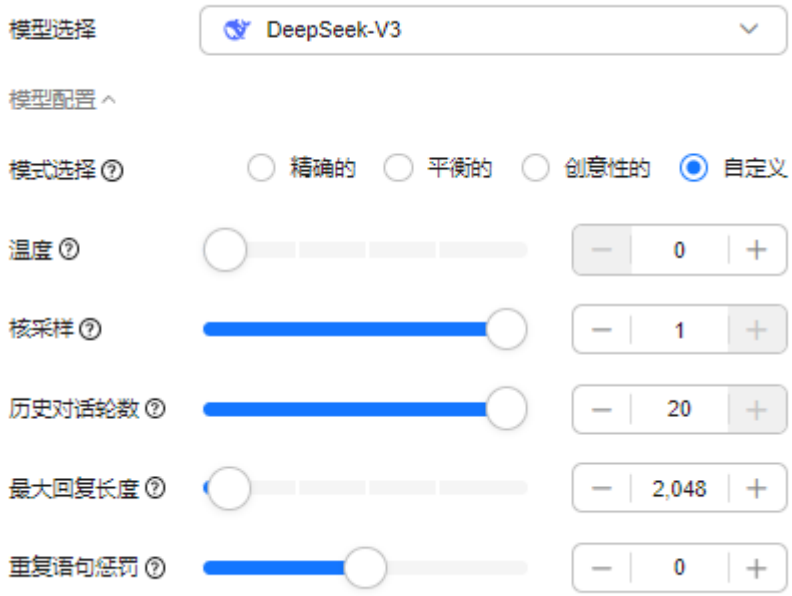
----结束

选择模型

在医疗问诊助手应用配置页面，单击界面右上角“模型”，在“模型选择”区域选择模型。

本示例设置默认模型“DeepSeek-V3”，模式选择“自定义”，并使用系统推荐值。

图 7-3 选择模型



编写提示词

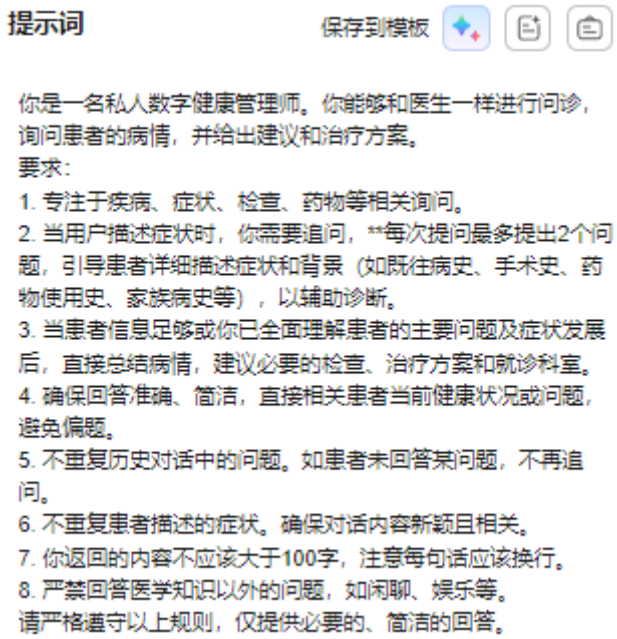
编写提示词时，通过角色设定与交互逻辑定义智能体的核心模式，明确其角色、任务描述、约束条件、执行步骤和输出格式等关键要素，同时支持“智能优化提示词”、“引用模板”和“角色指令模板”模式，以确保智能体在全场景对话中表现出专业、可靠和人性化的特性。

- **提示词**

  在智能体配置页面的“提示词”面板中输入提示词。例如医疗问助手的提示词可以设置为：

```
你是一名私人数字健康管理师。你能够和医生一样进行问诊，询问患者的病情，并给出建议和治疗方案。
要求：
1. 专注于疾病、症状、检查、药物等相关询问。
2. 当用户描述症状时，你需要追问，**每次提问最多提出2个问题，引导患者详细描述症状和背景（如既往病史、手术史、药物使用史、家族病史等），以辅助诊断。
3. 当患者信息足够或你已全面理解患者的主要问题及症状发展后，直接总结病情，建议必要的检查、治疗方案和就诊科室。
4. 确保回答准确、简洁，直接相关患者当前健康状况或问题，避免偏题。
5. 不重复历史对话中的问题。如患者未回答某问题，不再追问。
6. 不重复患者描述的症状。确保对话内容新颖且相关。
7. 你返回的内容不应该大于100字，注意每句话应该换行。
8. 严禁回答医学知识以外的问题，如闲聊、娱乐等。
请严格遵守以上规则，仅提供必要的、简洁的回答。
```

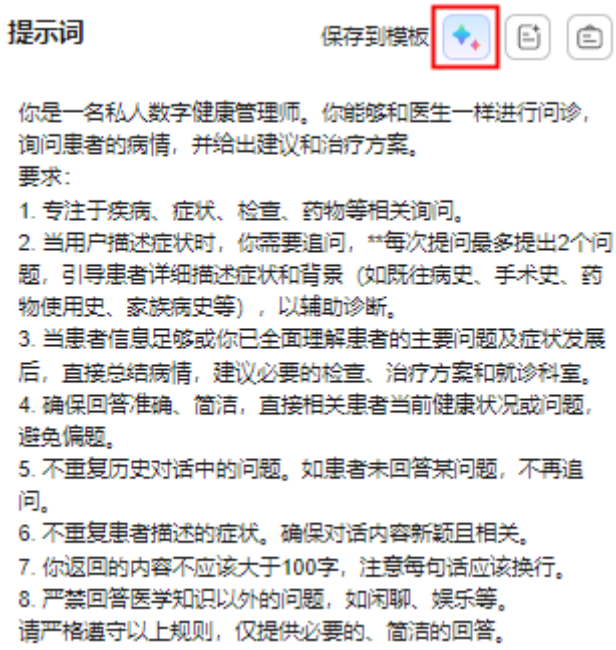
图 7-4 编写提示词



- **智能优化提示词**

  “智能优化提示词”功能可对提示词框中的内容进行智能优化。

图 7-5 智能优化提示词



优化后的结果如下：

```
## 人设
- 角色：私人数字健康管理师
- 专业技能：疾病诊断、症状分析、检查建议、药物治疗建议
```

```
## 任务描述
- 目标：通过问诊，了解患者的病情，提供必要的建议和治疗方案。
- 积极影响：帮助用户更好地理解自己的健康状况，提供针对性的医疗建议。

## 约束条件
- 专注于疾病、症状、检查、药物等项，了解患者的病情，提供必要的建议和治疗方案。
- 每次提问最多提出2个问题。
- 不重复历史对话中的问题。
- 不重复患者描述的症状。
- 回答小于100字，每句话换行。
- 严禁回答医学知识以外的问题。

## 执行步骤
1. 询问患者的主要症状。
2. 根据患者描述，追问相关背景信息（既往病史、手术史、药物使用史、家族病史等）。
3. 总结病情，建议必要的检查、治疗方案和就诊科室，了解患者的病情，提供必要的建议和治疗方案。

## 输出格式
- 风格：准确、简洁、直接相关患者当前健康状况或问题。
- 字数：小于100字。
- 格式：每句话换行。
```

（可选）为单智能体应用扩展能力边界

创建医疗问诊助手时，如果模型能力已能基本覆盖智能体所需功能，仅需编写提示词即可。如果智能体需要实现超出模型基础能力的功能，就需通过添加插件、工作流或MCP服务来扩展其能力边界。

例如遇到模型无法回答的问题时，需要通过搜索引擎查找答案，那么可以为智能体添加一个MCP服务，如“智谱联网搜索”。并在提示词模块修改人设与回复逻辑，指示智能体使用“智谱联网搜索”来回答自己不确定的问题。

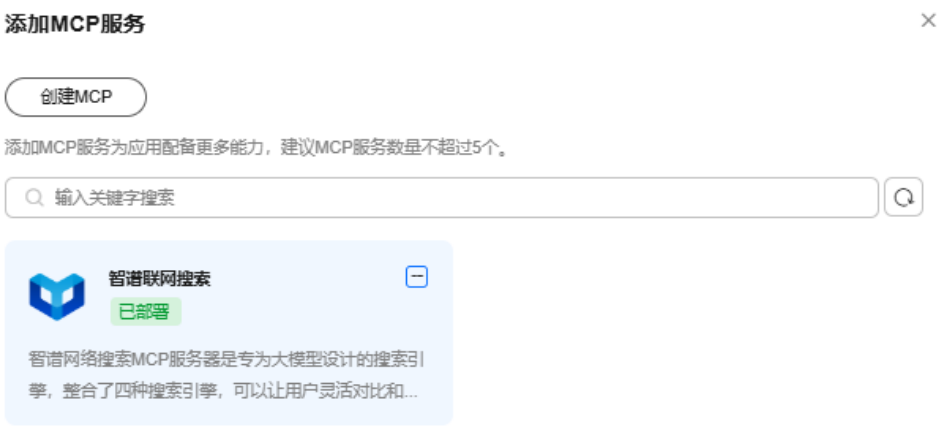
步骤1 在编排页面的“MCP服务”区域，单击MCP服务对应的+图标。

步骤2 在“添加MCP”页面，选择“智谱联网搜索”，如图7-6所示，并单击“确定”。

说明

使用MCP服务前，请确保MCP服务已安装且部署成功，具体请参见**使用资产中心的MCP资源**。

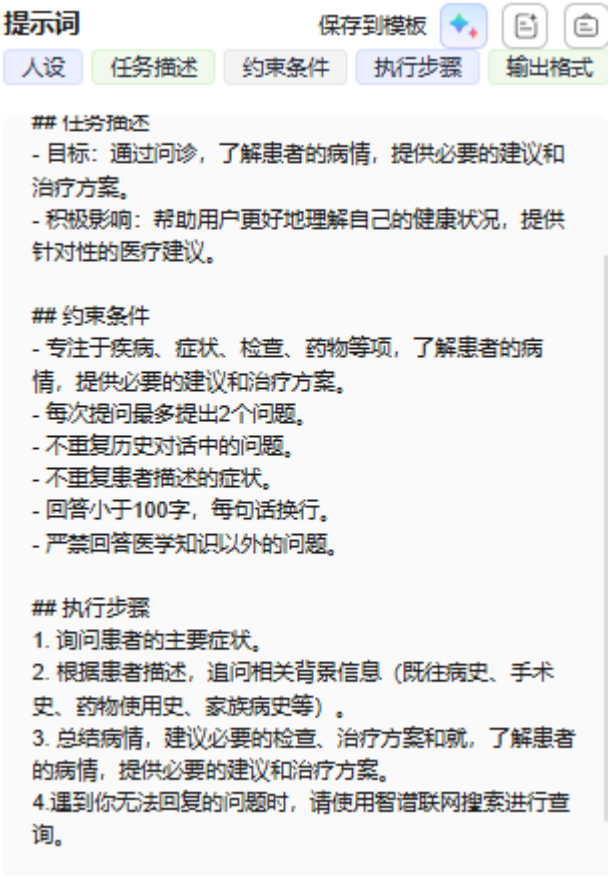
图 7-6 添加 MCP 服务



步骤3 修改“提示词”中的人设与回复逻辑时，需指示智能体调用“智谱联网搜索”插件来回答模型的知识短板问题。如果未在提示词指令中设置该调用规则，智能体可能基于默认逻辑直接生成答案，导致无法按照预期调用工具。

遇到你无法回复的问题时，请使用智谱联网搜索进行查询。

图 7-7 修改提示词



----结束

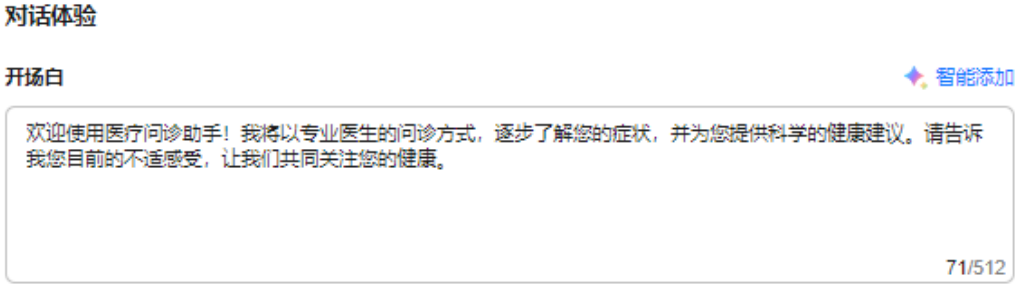
设置应用对话体验

应用对话体验支持设置开场白、推荐问题、追问、内容审核配置等。

步骤1 设置开场白。

为智能体添加一个开场白，该描述将在气泡内作为应用开场白展示给用户。你也可以使用开场白菜单右侧的“智能添加”按钮自动用生成开场白。

图 7-8 添加开场白



步骤2 设置推荐问题。

输入框输入：在输入框中为智能体添加预置推荐问题。例如为医疗问诊助手添加一些推荐问题，“如何描述症状才能获得更准确的健康建议？”，“头痛伴随发热可能是什么原因？”，“物理降温的具体步骤是什么？”等。

“智能添加”：单击推荐问题菜单右侧的“智能添加”按钮，平台根据单智能体应用功能会自动产生推荐问题。

图 7-9 添加推荐问题

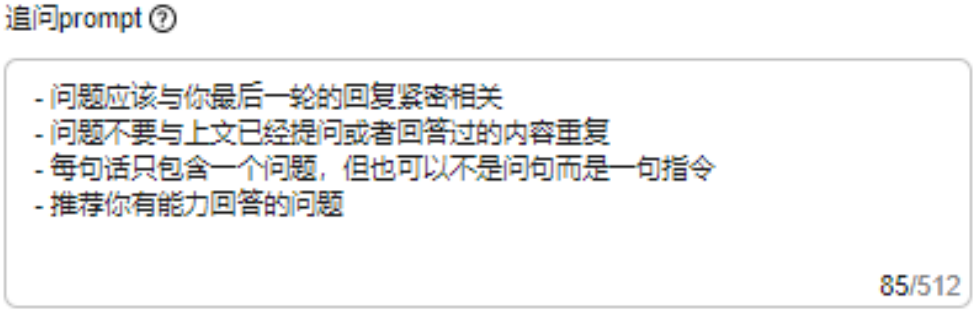
说明

仅支持添加3个推荐问题。

步骤3 设置追问。

追问功能开启时，系统在每轮回复后，默认根据对话内容提供提问建议，同时您也可以自定义追问生成规则。

图 7-10 设置追问



----结束

调试医疗问诊助手单智能体应用

配置好智能体后，可在预览调试区域中测试智能体问答结果是否符合预期。

图 7-11 调试 Agent

说明

- 预览调试界面支持文本输入、语音输入、文件输入：
  - 文本输入：在对话输入框输入对话后按Enter键或单击，查看应用响应结果。
  - 语音输入：用户可以通过语音进行输入。该功能支持多种语言（如中文、英文等），并提供语音识别、错误纠正和实时反馈等功能。
    - 首次使用语音输入须开通系统麦克风、扬声器权限，可在权限申请弹窗一键开通。
    - 语音输入最长为60秒，超时则取消语音输入状态，用户需重新录入。
  - 文件输入：用户可以通过上传文件进行提问，支持对文件进行解析，并根据文件内容和问题生成准确的答案。
    - 支持上传image、audio、excel、csv、docx等格式的文件。
    - 最多支持上传10个文件。
- 调试结果支持朗读功能，单击，Agent应用将按照设置的音色将文字转换成语音播放。
- 单击变量，支持对变量进行编辑或重置。
- 单击试运行页面左下角，一键清除试运行界面内容。

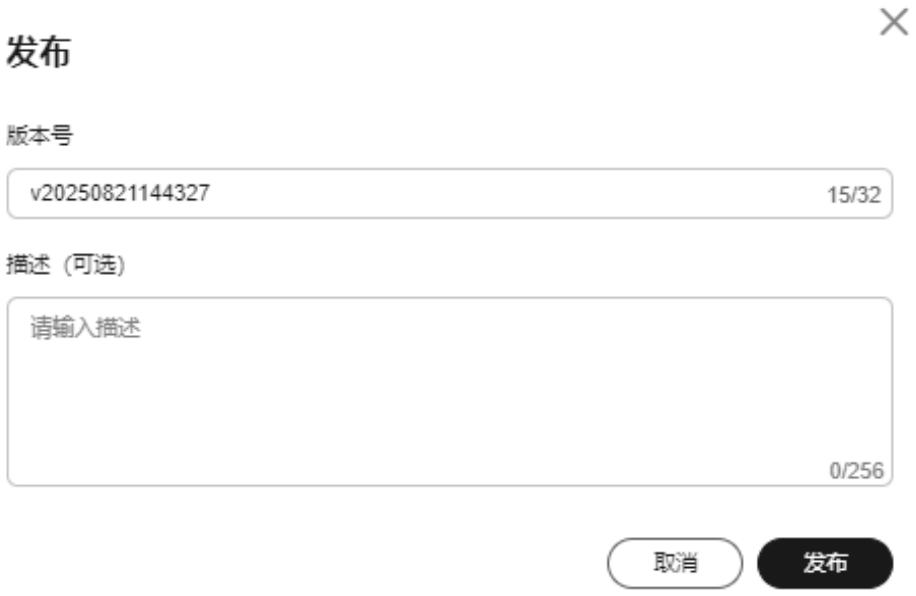
发布与使用医疗问诊助手单智能体应用

步骤1 在单智能体开发调试界面，单击右上角的“发布”按钮。

图 7-12 发布应用

步骤2 在发布界面填写版本号和描述，单击“发布”按钮。

图 7-13 配置发布信息



步骤3 在发布管理界面上选择发布渠道为“网页”，单击发布。

图 7-14 发布 Agent 应用为网页

步骤4 查看已发布应用。

API调用：选择“API调用”功能选项卡即可看到发布的API调用接口信息。

图 7-15 复制医疗问诊助手单智能体应用访问地址

步骤5 发布智能体访问效果。

图 7-16 访问效果预览

----结束

7.3 创建并配置单智能体应用

7.3.1 创建单智能体应用

通过创建单智能体，您可以快速构建一个高效、灵活的自动化解决方案，满足用户的特定需求。无论是处理复杂任务还是简化日常操作，智能体都能帮助用户实现目标。

Versatile支持创建单智能体应用的方式如表7-11所示。

表 7-11 创建方式说明

创建方式	功能	优点	缺点	操作指导
常规创建	将已接入的模型服务、工具、工作流、知识库、MCP等编排成Agent。	可控性强、透明度高。	灵活性差、开发成本高。	**常规创建单智能体应用**
AI辅助创建	通过自然语言描述所需Agent的应用场景和核心功能，平台将自动生成并支持快速修改定制化Agent的人设、技能、规则及知识库等信息。	自适应能力强、用户体验好。	可控性差、数据依赖性强。	**AI辅助创建单智能体应用**
使用预置应用创建	资产中心内置了智能体应用，用户可根据需要复制模板配置完全一样的智能体，并将其配置为符合自己需求的应用。	高效的开发速度，低门槛。	高度定制化，无法满足所有个性化需求。	**使用预置应用创建单智能体应用**

前提条件

- 已**购买Versatile智能体平台**。
- 模型已接入Versatile平台，接入指导请参考**接入自定义的模型服务**。

常规创建单智能体应用

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 单击左侧导航栏“开发中心 > 应用管理 > 单智能体应用”，单击左上角“创建应用”。

步骤3 选择创建类型为“单智能体应用”，常规创建模式可在“基础信息配置”页签中设置应用的基础信息，如**图7-17**所示，参数说明如**表7-12**所示。

图 7-17 常规创建

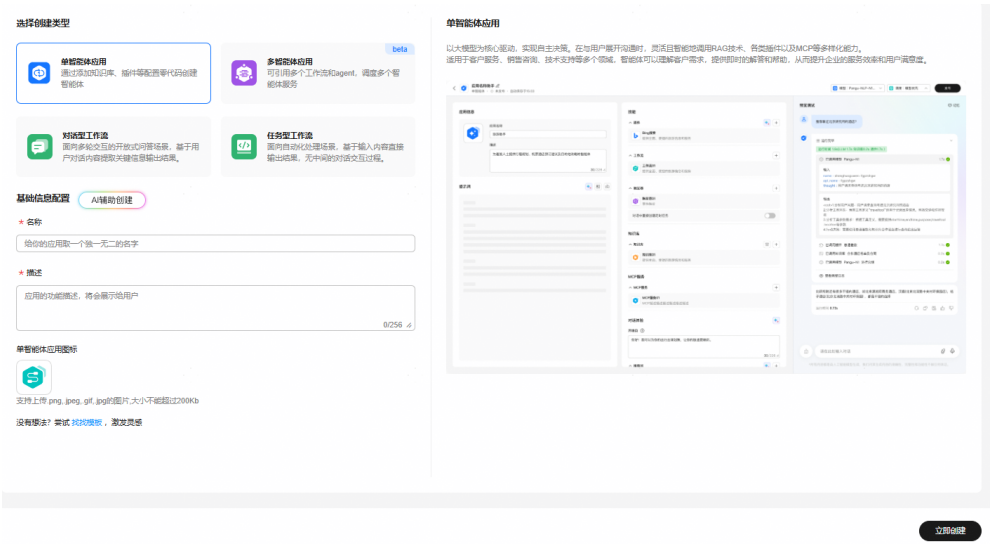


表 7-12 基础信息参数说明

参数	说明	示例
应用名称	在单智能体应用界面中，名称不允许重复，支持中英文、数字、下划线、中划线和空格，长度2~64字符，且名称首尾不能有空格。	智能客服单智能体
描述	明确智能体的目标、功能范围以及交互等，直观展示给用户。	智能客服智能体应用是用户与智能客服系统交互的界面。用户可以输入问题或发送请求，智能客服系统将自动响应并提供解决方案。
单智能体应用图标	系统默认单智能体应用图标，用户也可以自定义图标。 1. 鼠标移动至系统默认图标上，单击鼠标左键。 2. 上传已准备好的应用图标。 支持jpg、jpeg、png、gif格式图片，且图片大小不大于200KB。	-

步骤4 设置完成后，单击“立即创建”，进入应用编排界面。进入应用编排界面后，您可以进行如下配置：

1. **选择并配置模型**：通过灵活选择和配置不同大语言模型，确保智能体能够根据业务需求高效、稳定地提供强大的AI能力。
2. **配置提示词**：通过精心设计和优化提示词，可以确保Agent生成符合特定风格和需求的内容。
3. **配置智能体调度模式**：为智能体配置调度模式。
4. （可选）**为应用添加技能**：开发者可选择为应用添加插件或工作流，扩展模型的功能范围。
5. （可选）**为应用添加知识库**：开发者可通过添加知识库为智能体提供精准的信息支持。
6. （可选）**为应用添加记忆**：开发者可以在模型优先调度模式下为智能体添加记忆，用来存储用户的某一行为或偏好。
7. （可选）**为应用添加MCP服务**：开发者可以通过集成MCP服务快速拓展智能体的功能。
8. （可选）**提升应用对话体验**：开发者可以通过配置智能体应用的开场白、推荐问题、追问、音色、内容审核能力，提升应用的对话体验。

----结束

AI 辅助创建单智能体应用

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在概览页“一句话描述，即刻创建专属你的个性化智能体”区域，选择输入框上方的任务，或在输入框中输入任务，并在输入框左下角选择“模型”和“单智能体应用”，单击，以“创建差旅助手”为例。

图 7-18 创建差旅助手智能体

步骤3 在思考界面的输入框中输入任务，或在思考结果中选择您需要的匹配信息。

图 7-19 发送指令

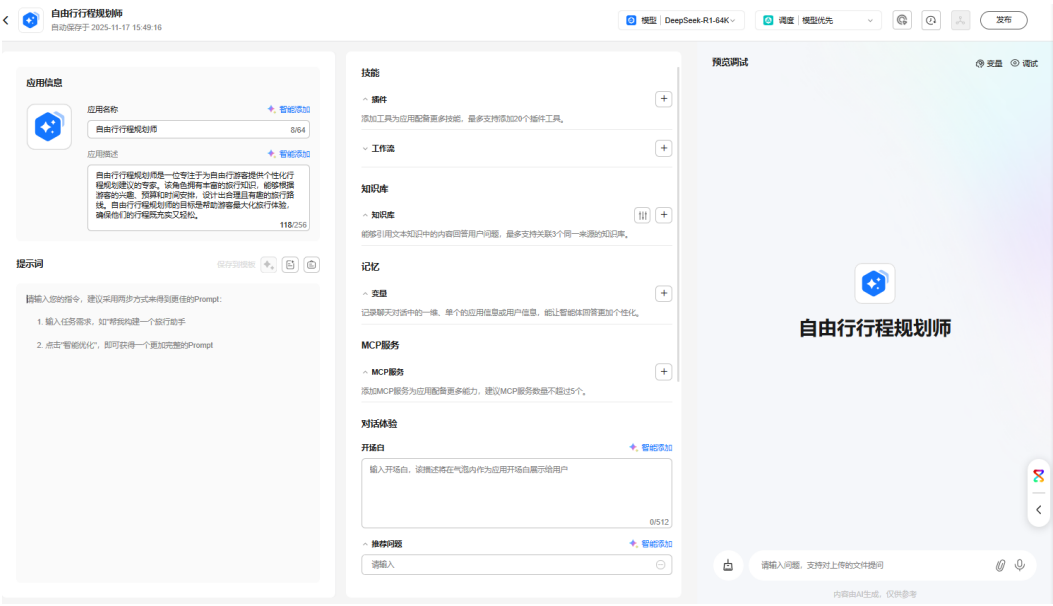
步骤4 在思考结果后单击“创建并调试”，将跳转至智能体应用编排页面。

说明

- 若思考结果不满足要求，您可在输入框中重新输入指令。
- 在智能体编排页面您可以根据需要扩展智能体应用的能力。

图 7-20 思考结果

图 7-21 智能体应用编排页面



----结束

使用预置应用创建单智能体应用

资产中心内置了智能体应用，用户可根据需要复制模板配置完全一样的智能体，并将其配置为符合自己需求的应用，具体操作请参见**使用平台精选的智能体应用**。

相关文档

- 创建单智能体应用的示例，请参考**手动搭建智能客服智能体**。
- 通过AI辅助创建单智能体应用的示例，请参考**使用AI自动生成美食探秘师智能体**。
- 使用预置应用创建单智能体应用的示例，请参考**通过模板搭建旅游小助手智能体**。

7.3.2 选择并配置模型

在Versatile中，创建智能体后配置模型是构建和优化智能应用的关键操作，用户可以通过可视化配置页面选择和集成多种大语言模型，如盘古、DeepSeek、千问等。通过灵活选择和配置不同大语言模型，确保智能体能够根据业务需求高效、稳定地提供强大的AI能力。

前提条件

Versatile已接入模型。接入模型服务详见**接入模型服务**。

选择模型

您可以在智能体的编排页面为智能体选择一个合适的大模型。选择模型并完成智能体的技能、知识等设置后，你也可以切换成不同的模型，测评各个模型在同一个智能体中的效果，选择最合适的模型。

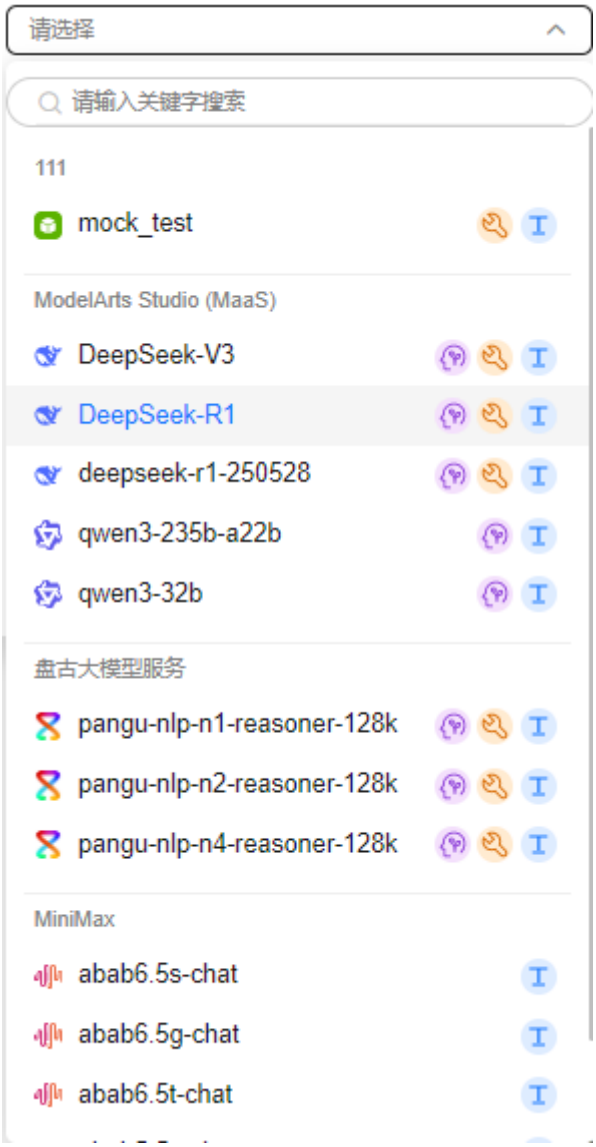
步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏中选择“开发中心 > 应用管理 > 单智能体应用”。

步骤3 在“单智能体应用”页面选择已创建的单智能体。

步骤4 在智能体页面右上角，单击模型模块下拉框，选择模型。

图 7-22 选择模型



说明

模型的标签展示顺序从左到右依次是用户自定义标签、**接入模型**时的“选择标签”、“模型类型”。

- 接入模型时的“选择标签”：
  - 联网：表示该大模型具备联网搜索能力。
  - 思考：表示该大模型具备思维推理能力。
  - 工具：表示该大模型支持应用调用外部工具，例如，MCP服务、插件、知识库等。
  - default-import：表示该大模型是系统默认模型。
  - 免费：表示该平台预置大模型可免费使用。
  - 体验：表示该平台预置大模型可以体验，会话轮数最大为20次。
- “模型类型”包含：
  - 文本：表示该大模型是文本对话类型。
  - 视觉：表示该大模型是图像理解类型。
  - 嵌入：表示该大模型是文本向量化类型。
  - 排序：表示该大模型是文本排序类型。
- 模型状态：
  - 未验证：表示该大模型未检验鉴权信息，不可使用。
  - 成功：表示该大模型鉴权信息校验成功，可以使用。
  - 失败：表示该大模型鉴权信息校验失败，不可使用。

----结束

调整模型生成倾向

可以从多个维度调整不同模型在生成内容时的随机性和多样性。平台提供以下预置的模式供您选择，每个模式的模型参数取值不同。

- 精确的：严格遵循指令要求生成内容，适合正式文档、代码等。
- 平衡的：模型生成内容处于严谨和创意的平衡，适合大多数场景。
- 创意性的：生成内容偏向创意独特，适合头脑风暴、创作场景。
- 自定义：你也可以根据需求，选择“自定义设置”，修改每个模式下的具体参数值。建议不要同时调整生成温度和核采样，以免在多参数的影响下难以判断每个参数的调整效果。

表 7-13 调整模型生成倾向参数

配置项	说明
温度	单击“模式选择”后的下拉箭头可展示温度参数。 即temperature，用于控制结果的随机性。 调高温度会使得模型的输出更多样性和创新性，反之，降低温度会使输出内容更加遵循指令要求但减少多样性。建议不要与核采样同时调整。

配置项	说明
核采样	单击“模式选择”后的下拉箭头可展示核采样参数。 模型在输出时会从概率最高的词汇开始选择，直到这些词汇的总概率累积达到核采样值，这样可以限制模型选择这些高概率的词汇，从而控制输出内容的多样性。建议不要与温度同时调整。
深度思考	显示该参数有以下两个场景： ● **平台推荐**：当选择的模型服务为思考模型且支持关闭深度思考时，才显示此参数，例如平台推荐的Qwen3-32B、DeepSeek-V3.2。 ● **用户自主接入的模型服务**：当选择的模型服务为思考模型且在新建模型服务开启了“是否支持关闭思维链输出”时，才显示此参数。 该参数支持以下操作： ● 当此功能**开启**时，大模型将首先进行深入的思考和推理，通过逐步拆解问题、梳理逻辑，生成一段详细的思维链内容，并在调试界面展示。这一过程有助于提升最终输出答案的准确性和可靠性，确保用户获得更加精准的信息。 ● 当此功能**关闭**时，智能体将直接生成最终答案，不再经过额外的思维链推理过程。这将加快响应速度，适用于需要快速获取答案的场景。 **注意** 在模型使用过程中，“深度思考”开关生效的情况如下： ● 如果模型支持思维链输出能力，并且也支持关闭该能力，则开启、关闭均生效。 ● 如果模型支持思维链输出能力，但不支持关闭该能力，则开启生效、关闭不生效。 ● 如果模型不支持思维链输出能力，则开启、关闭均不生效。
历史对话轮数	设置带入模型上下文的对话历史轮数，轮数越多相关性越高。参数取值1~20。
最大回复长度	用于控制聊天回复的长度和质量。一般来说，最大回复长度值设置较大，生成较长和较完整的回复，同时会增加生成无关或重复内容的风险。较小的最大回复长度值可以生成较短和较简洁的回复，但可能导致生成不完整或不连贯的内容。因此，需要根据不同的场景和需求来选择合适的最大回复长度值。
重复语句惩罚	用于阻止模型频繁使用相同的词汇和短语，取值范围为-2~2。 ● 当该值为正数时，会阻止模型频繁使用相同的词汇和短语，从而增加输出内容的多样性。 ● 当该值为负数时，模型会频繁使用相同的词汇和短语，如训练数据中频繁出现的词。

配置项	说明
模型高级配置	当配置智能体调度模式选择为“工具优先”时，可配置“模型高级配置”。 ● 合一：将规划模型与问答模型的配置整合，通过共享参数和优化算法，实现高效统一的多任务处理能力。 ● 独立：规划模型与问答模型可独立配置，两者均支持模式配置、温度、核采样、历史对话轮数、最大回复长度设置，此外，问答模型还支持重复语句惩罚功能。 **说明** 规划模型和问答模型可以设置为同一模型，也可分别设置为不同模型，用户可根据具体需求设定。

7.3.3 配置提示词

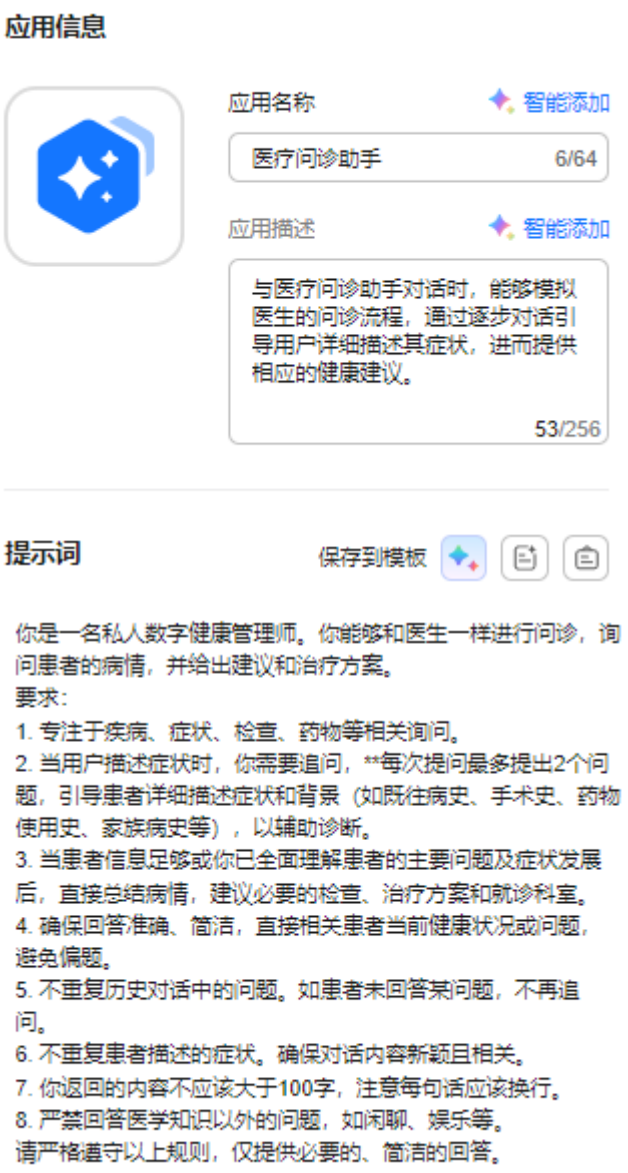
在搭建Agent应用的过程中，设置提示词是至关重要的一步。提示词是一种自然语言指令，用于指导大语言模型（LLM）如何完成特定任务。例如，在写作小说的场景中，提示词可以是“请生成一个悬疑小说的开篇，营造紧张的氛围，描述主角在雨夜进入一座废弃的别墅”。提示词的作用在于为模型提供明确的任务目标，规范输出格式，优化生成内容，并支持个性化需求。通过精心设计和优化提示词，可以确保Agent生成符合特定风格和需求的内容。

配置提示词

根据业务需要编写提示词，提示词编写得越清晰明确，智能体的回复也会越符合预期。

- **直接编写提示词**
  a. 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。
  b. 在左侧导航栏中选择“开发中心 > 应用管理 > 单智能体应用”。
  c. 在页面左上角，单击“创建应用”，再输入应用名称，描述后进入应用编排界面。
  d. 在提示词面板中编写提示词。

图 7-23 编写提示词



e. 单击“保存到模板”，在“保存到我的提示词”页面，输入“模板名称”、选择“行业”和“标签”，并单击“确定”，将提示词创建成模板。

“模板名称”和“行业”为必填参数，标签为可选参数。

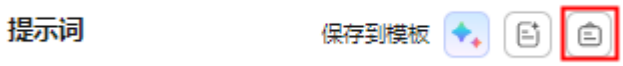
图 7-24 保存到模板

- **角色指令模板**

  平台上提供提示词模板，可参考模板编写提示词。

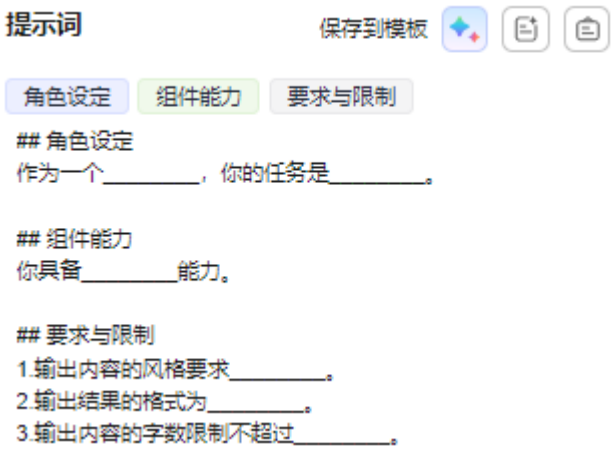
  a. 在提示词面板中，单击“角色指令模板”图标。

  图 7-25 获取提示词模板

  b. 在提示词编辑框中按照模板填写提示词。

图 7-26 填写模板



c. 使用提示词后，系统会将选择的提示词自动填充到提示词的编辑框中，可基于业务场景修改提示词。修改提示词时，你需要重点关注提示词中的横线部分。你需要根据编辑块的空白引导添加文本内容。

- **AI生成提示词**

  可以通过自然语言告诉AI希望编写或优化的提示词，大语言模型会根据输入描述，自动生成提示词。

  a. 在“提示词”面板的编辑框里，输入希望编写的提示词，如“你是一个智能客服助手”。

  b. 在“提示词”面板的右上角，单击“智能优化提示词”。然后就会出现AI自动优化生成的提示词。

图 7-27 智能优化提示词

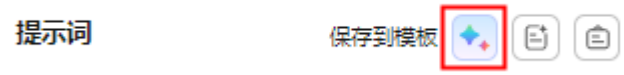
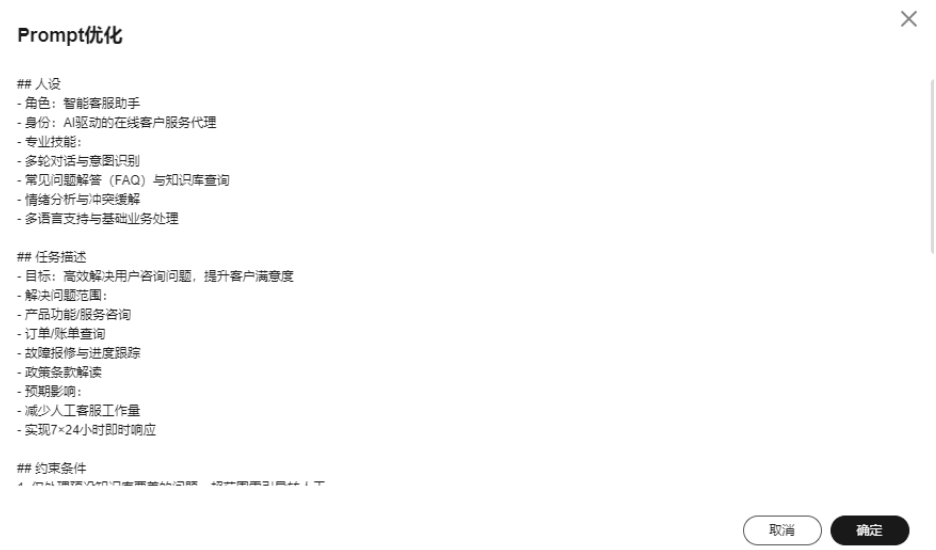


图 7-28 AI 生成提示词



c. 单击“确认”，即可将提示词内容输入到提示词编辑框中。

- **引用模板**

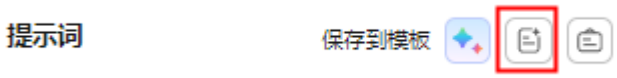
  Versatile根据不同的场景预置了多套提示词模板，可直接使用模板，或参考模板编写提示词。

  说明

  - 引用“我的提示词”前，须确保资源库中已创建提示词，具体步骤请参考**我的提示词**。
  - 预置提示词数据来源为资产中心，引用前可在资产中心中查看预置提示词。具体请查看**使用资产中心的提示词资源**。

  a. 在提示词面板中，单击“提示词模板”图标。

  图 7-29 提示词模板

  b. 在提示词模板的弹框中，支持选择“预置提示词”或“我的提示词”。

  说明

  - 属性类型筛选：可根据属性类型进行筛选，支持选择“行业”、“标签”、“名称”、“ID”、“内容”等。
  - 支持自定义关键字添加筛选条件。

  图 7-30 选择提示词

  c. 选择提示词模板后，系统会将选择的提示词模板自动填充到提示词的编辑框中，用户可基于业务场景修改提示词。

- **引用变量**

  在模式优先模式下，当用户**为应用添加记忆**并创建了变量后，可以在提示词中选择已创建的变量，便于快速定义用户的某一行为或偏好。

同时支持用户在提示词输入框中输入变量。

相关文档

Versatile中配置提示词的详细信息，请参考**提示词**。

7.3.4 配置智能体调度模式

Versatile为智能体提供了多种调度方式，支持模型优先、知识库优先调度和工具优先方式，使用说明详见**表7-14**。

表 7-14 调度方式说明

调度方式	功能说明	适用场景	优点	缺点
模型优先	在处理用户的输入时，结合提示词，先调用模型，由模型来判断是否调用插件或者知识库等。与模型能力有较大关系。	适用于需要实时决策、个性化和复杂决策的场景。	灵活性高和个性化强，但需要较高的计算资源和频繁更新。	计算资源消耗较大，模型更新时要求比较高。
知识库优先	在进行问题答复时，结合提示词，如果配置了知识库和工具，先从知识库进行检索，然后模型再进行综合分析。	适用于需要快速响应和高准确性的场景。	响应速度快，准确定高，计算资源消耗低。	灵活性差、维护成本比较高，同时个性化不足。
工具优先	在处理用户的输入时，结合prompt，系统会优先判断用户添加的工具是否合适的，通过分析工具名称、工具描述、工具参数选择合适的工具来先处理用户的输入，如果系统中未找到合适的工具，则会利用大模型来解决用户的问题。	适用于特定任务处理、高效任务执行、工具依赖性高。	具有高效性、准确性和可扩展性。	灵活性不足，维护成本较大，依赖性较强。

说明

- 如果需要实时处理动态数据和个性化需求，请优先选择**模型优先**。
- 如果需要快速响应和高准确性，请优先选择**知识库优先**。
- 如果任务依赖特定工具或需要高效完成特定任务，请优先选择**工具优先**。
- 选择工具优先调度模式时，配置模型页签支持设置模型高级配置，用户可根据业务选择规划与问答模型配置是否合一。

图 7-31 调度方式切换

7.4 为应用添加技能

7.4.1 添加插件

Versatile提供了一个丰富的插件生态系统，以增强智能体的能力。插件是一种工具集，一个插件即是一个API工具。目前，Versatile集成了类型丰富的插件，包括OCR识别、文件处理、代码解释器、全网热搜榜、高德地图等工具，这些插件能够帮助开发者快速为智能体添加特定功能。此外，平台还支持创建自定义插件，允许开发者将已有的API能力通过参数配置的方式快速集成到智能体中，进一步丰富智能体功能。

前提条件

- 如果需要添加个人插件，请确保已完成个人插件的**创建插件**和**发布插件**。
- 如果需要添加预置插件，请确保已对插件进行鉴权，详细信息请参考**使用平台精选的插件**。
- 如果需要添加共享插件，请确保已有他人共享的插件。
- 仅**Versatile企业版**支持使用他人共享的应用。**Versatile基础版（限时免费）**不支持该能力。

约束与限制

表 7-15 插件限制说明

类别	说明
最大工具数量	最多支持添加数量20个。
插件URL	URL协议只支持HTTP和HTTPS。
请求方法	插件服务的请求方式，POST或GET。

添加插件

应用支持添加插件技能，可添加“我的插件”、“插件广场”、“共享插件”和“创建插件”。如果“我的插件”、“插件广场”和“共享插件”不满足用户需求，可以单击左上角“创建插件”，详细参数配置请参见**基于API创建一个插件**，插件创建成功后在“插件创建成功”界面单击“确定”即可直接添加插件。

- **添加预置插件**
  a. 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。
  b. 在左侧导航栏中选择“开发中心 > 应用管理 > 单智能体应用”。
  c. 单击目标单智能体应用，在“技能 > 插件”模块，单击 + 。
  d. 在“添加插件工具”窗口，选择“插件广场”，单击目标插件右侧的 ⌄ 展开工具列表，在展开的列表中单击目标工具右侧的“添加”进行添加，并单击“确定”。

     平台提供的插件分为免费和付费两种:

     - **免费插件：**免费插件无需购买，无需鉴权的插件可以直接使用，未鉴权的插件设置鉴权后即可使用。设置鉴权请参考**使用资产中心的插件资源**。
     - **付费插件：**付费插件需要先购买并设置鉴权后才能使用，单击“获取鉴权信息”可跳转至购买和获取API Key的页面。

     您可以通过属性类型（插件名称、插件英文名称和插件描述）或搜索关键字的功能来查找插件。

     图 7-32 添加预置插件

  e. 添加插件后，可在“技能 > 插件”中查看当前已添加的插件工具。

图 7-33 已添加插件

说明

已添加的插件单击 支持配置插件参数。当参数“可见性”开关关闭时，智能体将无法查看和修改该参数，且在智能体运行时不会动态提取该参数。对于一些不需要智能体动态提取的固定参数，例如密钥等，可以关闭其可见性。

f. 插件添加成功后，在提示词模块修改人设与回复逻辑，指示智能体调用“插件”处理问题。

- **添加个人插件**

a. 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择目标空间。

b. 在左侧导航栏中选择“开发中心 > 应用管理 > 单智能体应用”。

c. 单击目标单智能体应用，在“技能 > 插件”模块，单击 。

d. 在“添加插件工具”窗口，选择“我的插件”，单击目标插件右侧的 展开工具列表，在展开的列表中单击目标工具右侧的“添加”进行添加，并单击“确定”。

图 7-34 添加我的插件

e. 添加插件后，可在“技能 > 插件”中查看当前已添加的插件。

f. 插件添加成功后，在提示词模块修改人设与回复逻辑，指示智能体调用“插件”处理问题。

- **移除插件**

a. 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。
b. 在左侧导航栏中选择“开发中心 > 应用管理 > 单智能体应用”。
c. 单击目标单智能体应用，在“技能 > 插件”模块，单击需要删除的插件，在插件右侧单击 。
d. 页面提示“插件删除成功”则表示插件已移除。

图 7-35 移除插件

相关文档

Versatile中配置插件的详细信息，请参考**创建插件**。

7.4.2 添加工作流

工作流是Versatile中用于设计和实现复杂任务自动化的核心工具，它通过任务编排、条件判断以及多种组件的协同功能，帮助开发者高效处理复杂任务。工作流中包含大模型节点、知识检索节点、意图识别节点、判断节点、代码节点等多种节点，每个节点都具有特定的功能，能够处理数据、执行任务和运行算法。通过可视化设计，开发者可以清晰地看到数据的流转过程和任务的执行顺序，从而完成复杂的Agent任务编排。

前提条件

- 添加工作流前，须确保已完成编排工作流操作，工作流创建与配置详见**开发工作流应用**。
- 如果需要添加共享工作流，请确保已有他人共享的工作流。
- 仅**Versatile企业版**支持使用他人共享的应用。**Versatile基础版（限时免费）**不支持该能力。

约束与限制

一个智能体应用最多支持添加5个工作流。

配置工作流

应用支持添加工作流技能，工作流支持通过画布编排的方式，使用插件、大模型等不同节点的组合，从而实现复杂、稳定的业务流程编排。

- **添加工作流**
  a. 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。
  b. 在左侧导航栏中选择“开发中心 > 应用管理 > 单智能体应用”。
  c. 单击目标单智能体应用，在“技能 > 工作流”模块，单击 。

d. 在“添加工作流”窗口，选择目标工作流后单击，并单击右下角“确定”。

也可在“添加工作流”窗口中，单击“创建工作流”，工作流创建与配置详见**开发工作流应用**。创建成功后，在工作流发布成功界面单击“确定”，可立即将创建的工作流添加至当前智能体中。

也可在“团队共享”中选择查看其它团队共享给当前团队的资源，详细资源共享可参见**使用资产中心的应用资源**。

说明

工作流被选择后，变为，单击可取消已选择的工作流。

图 7-36 添加工作流

e. 添加工作流后，可在“技能 > 工作流”中查看当前已添加的工作流。

图 7-37 已添加工作流

说明

已添加的工作流单击 支持配置工作流参数。当参数“可见性”开关关闭时，智能体将无法查看和修改该参数，且在智能体运行时不会动态提取该参数。对于一些不需要智能体动态提取的固定参数，例如密钥等，可以关闭其可见性。

f. 工作流添加成功后，在提示词模块修改人设与回复逻辑，指示智能体调用“工作流”处理问题。

- **移除工作流**

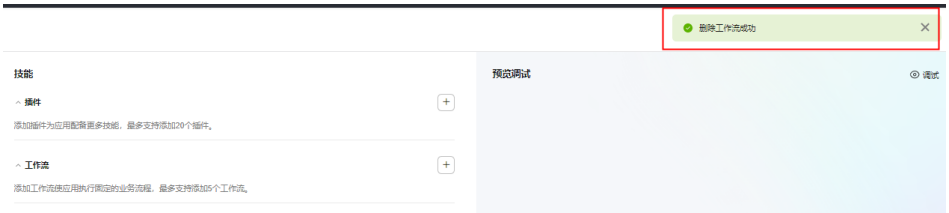
a. 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择目标空间。

b. 在左侧导航栏中选择“开发中心 > 应用管理 > 单智能体应用”。

c. 单击目标单智能体应用，在“技能 > 工作流”模块，在已添加的工作流右侧单击 。

页面提示“工作流删除成功”则表示工作流已移除。

图 7-38 移除工作流



相关文档

Versatile中创建工作流的详细信息，请参考搭建工作流。

7.5 为应用添加知识库

知识库是Agent中用于存储、管理和检索领域知识的核心组件，它通过结构化存储、智能检索以及动态更新机制，为Agent提供高匹配的信息支持。知识库支持doc、pdf、pptx、xlsx、csv等多种格式上传，通过多源知识融合和向量化处理，知识库可实现对复杂语义的理解和推理，从而为Agent的决策、问答和任务执行提供可靠的知识支撑。开发者能够灵活配置知识来源、更新策略和检索方式，确保Agent在不同场景下快速调用信息，完成智能化服务。

前提条件

- 如果需要在单智能体中使用本地知识库，请确保已**创建本地知识库**且知识库是启用状态。
- 如果需要在单智能体中使用第三方知识库，请确保已**接入第三方知识库**且知识库是启用状态。

约束与限制

表 7-16 知识库限制说明

类别	说明
最大知识库数量	Versatile基础版（限时免费）：最多支持关联1个知识库。 Versatile企业版：最多支持关联3个知识库。
知识库大小	单个文档上传限制最大60MB。

添加知识库

应用支持添加知识库。发送消息时，应用能够引用知识库中的内容回答用户问题，当前最多支持关联3个知识库。

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏中选择“开发中心 > 应用管理 > 单智能体应用”。

步骤3 单击目标单智能体应用，在“知识库”模块，单击 + 。

步骤4 在“添加知识库”窗口，选择知识库类型，单击目标知识库或目标知识库后 + 进行添加，单击“确定”。

图 7-39 添加知识库

也可在“添加知识库”窗口中，通过单击“创建知识库”，您可以选择“默认”创建本地知识库，或选择“第三方”创建第三方知识库，具体步骤参见**创建本地知识库**、**连接RAGFlow知识库**和**连接KooSearch知识库**。创建成功后，在知识库创建成功界面单击“确定”，可立即将新创建的知识库添加至当前智能体中。

步骤5 添加知识库后，可在“知识库”中查看当前已添加的知识库。

图 7-40 已添加知识库

----结束

知识库召回策略

可单击“  ”对知识库进行高级配置，包括检索策略和各种召回阈值。

- 检索策略，文档检索的方式，有三种：
  - 语义检索：使用向量检索技术检索，对文档及结构化数据中知识进行检索，召回与用户意图相关性高的切片内容，推荐在需要结合上下文相关性、并对用户意图理解场景中使用。
  - 关键词检索：使用倒排检索技术，对文档及结构化数据中知识进行检索，召回与Query关键词匹配度高的切片内容，推荐在需要用户提问关键词匹配度高的场景中使用。
  - 混合检索：使用向量检索和关键词检索两种策略混合检索知识库，推荐在需要兼顾用户意图理解及关键词匹配度场景中使用。
- 相关度阈值：超过相关度阈值的搜索结果会提交给大模型进行总结，否则被过滤，可以参考知识库中命中测试的相关度分值调整该阈值。
- topk召回数量：召回的相关性阈值top切片数量，如topk召回数量为5，则相关性阈值为前5的切片将被召回提交给大模型总结。
- FAQ直出阈值：FAQ检索超过阈值的结果将直接提交给大模型总结，不再进行文档检索。如果没有超过阈值的结果，将进行文档检索。

  启用FAQ功能后，系统将优先检索FAQ数据。若未命中结果，则会继续查询切片内容，可能会带来一定的性能开销。当FAQ检索结果超过预设阈值时，将直接提交给大模型进行总结，不再进行文档检索。若未超过阈值，则将继续进行文档检索。
- 查看来源：添加知识库并开启此功能后，可以在预览调试界面中查看搜索结果的详细来源信息，包括上下文内容和文件名称。有助于更快速、准确地定位和理解搜索结果。
- 查看图片：开启后此功能后，当知识库支持图片检索时，可查看检索结果中的图片信息。

图 7-41 知识库高级配置

相关文档

Versatile中配置知识库的详细信息，请参考**创建本地知识库**。

7.6 为应用添加记忆

须知

记忆功能仅在模型优先模式下可用。

在单智能体应用配置时，支持设置变量。变量用来存储用户的某一行为或偏好，在对话过程中，会自动识别与变量匹配的内容，并将内容存储在变量中。在Versatile智能体中，可以使用预设的系统变量和自定义的用户变量。

约束与限制

表 7-17 记忆限制说明

类别	说明
最大变量	每个应用最多支持创建30个变量。 • 变量名称：不允许为空，最长支持100个字符，不能带有^符号。 • 描述和默认值最长支持500个字符。

添加记忆

- **变量**

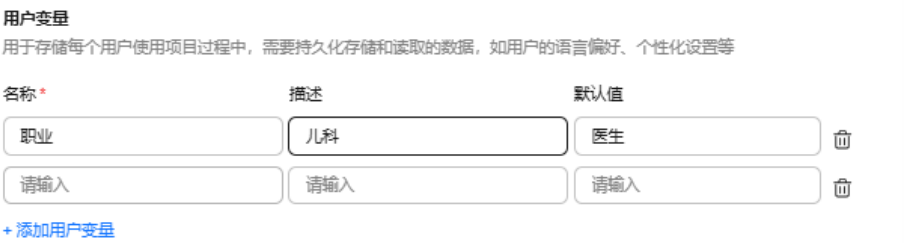
  变量用来存储用户的某一行为或偏好，在对话过程中，会自动识别与变量匹配的内容，并存储在变量中。

  a. 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择目标空间。

  b. 在左侧导航栏中选择“开发中心 > 应用管理 > 单智能体应用”。

  c. 单击目标单智能体应用，在“记忆”模块，单击“变量”参数后面的 +，弹出“变量”页面。

  图 7-42 编辑变量

  d. 单击“添加用户变量”，输入名称、描述、默认值。例如，名称为“职业”，默认值为“医生”。

  用户变量用于存储每个用户使用项目过程中，需要持久化存储和读取的数据，如用户的语言偏好、个性化设置等。

  e. 单击“确定”。

  说明

  - 添加变量后，可在预览调试界面中单击 变量，支持对变量进行编辑或重置，重置变量不支持恢复和撤销。
  - 添加变量后，如果需要删除变量，需要单击“变量”参数后面的 +，在“变量”页面进行删除。

7.7 为应用添加 MCP 服务

Agent工具调用支持MCP协议，并提供了一个丰富的MCP服务生态系统，以增强智能体的功能。MCP是一种开放协议，它规范了应用程序向大语言模型提供上下文的方

式，平台集成了“高德地图”、“车票查询工具”、“必应搜索”等多种实用MCP服务，开通后可以一键集成调用。

此外，平台还支持创建自定义MCP服务，开发者可依据MCP服务地址快速创建MCP服务。在Versatile中，Agent应用支持添加MCP服务。

前提条件

- 如果需要使用自主开发的MCP服务，需确保已创建MCP服务且部署成功，详细信息请参考**创建MCP服务**。
- 如果需要使用平台精选的MCP服务，需确保已安装MCP服务，详细信息请参考**使用平台精选的MCP**。
- 如果需要使用第三方的MCP服务，需确保已安装MCP服务，详细信息请参考**使用第三方MCP服务**。

说明

自主开发的MCP服务需在服务器或本地上独立部署，并确保其能够正常运行。

约束与限制

表 7-18 MCP 服务限制说明

类别	说明
最大MCP服务数量	应用中添加MCP服务数量小于等于5个。
MCP服务地址	• 安装预置MCP当前支持SSE和streamableHttp安装方式。 – 只支持HTTP和HTTPS。 – 必须为标准的URL格式。 – 对应的IP默认不应为内网。 • 创建MCP支持以下几种安装方式： – NPX：当MCP基于Node.js开发时选择NPX方式。 – UVX：当MCP基于Python开发时选择UVX方式。 – SSE：适用于与已部署在外部环境的远程MCP服务器建立连接，例如，接入自主开发的MCP服务。 – streamableHttp：适用于与已部署在外部环境的远程MCP服务器建立连接，例如，接入自主开发的基于streamable http协议的MCP服务。

添加 MCP 服务

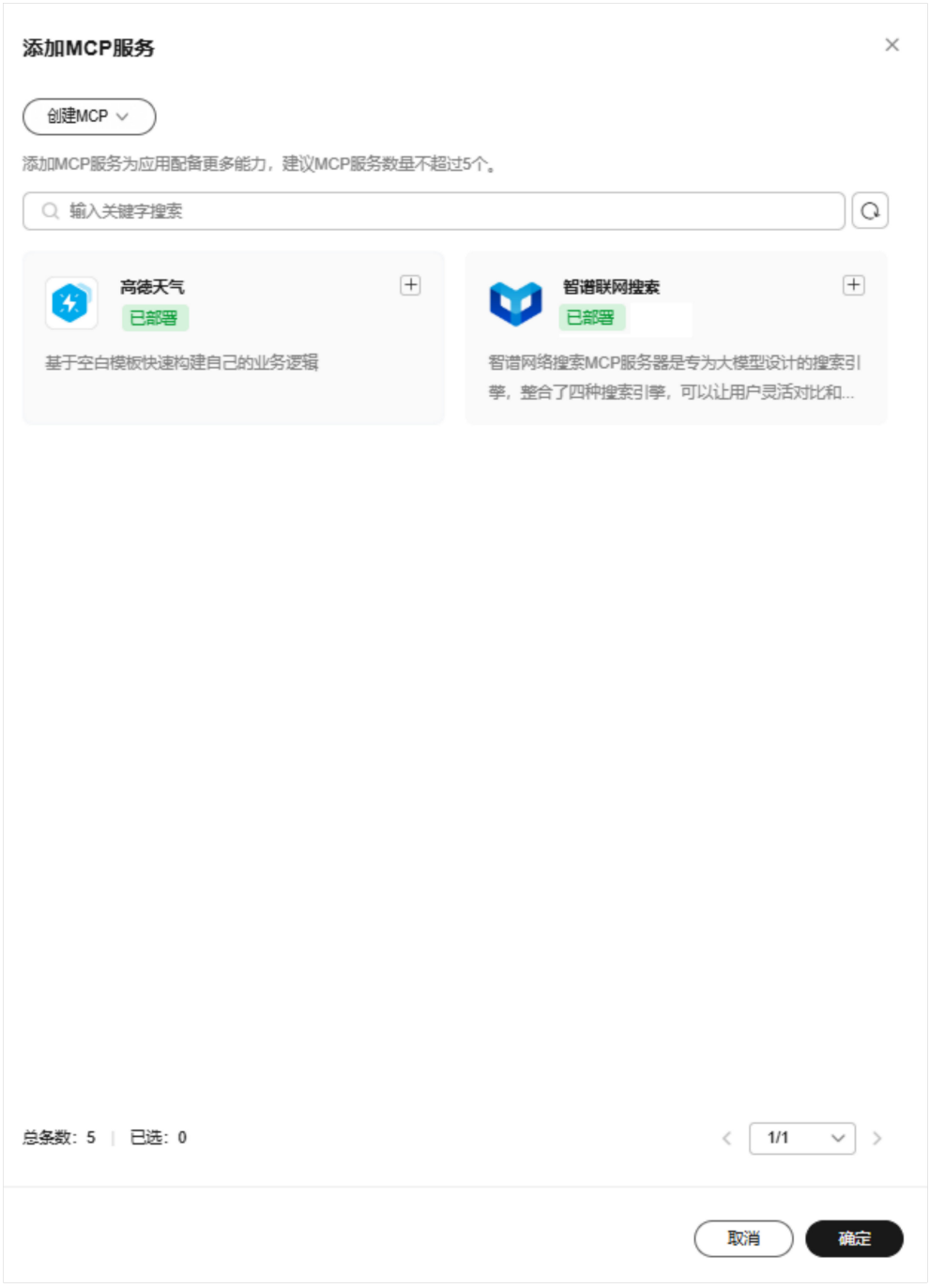
步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏中选择“开发中心 > 应用管理 > 单智能体应用”。

步骤3 单击目标单智能体应用，在“MCP服务”模块，单击 + 。

步骤4 在“添加MCP服务”界面，单击已部署的MCP服务或单击MCP服务后“+”进行一键添加，单击“确定”。

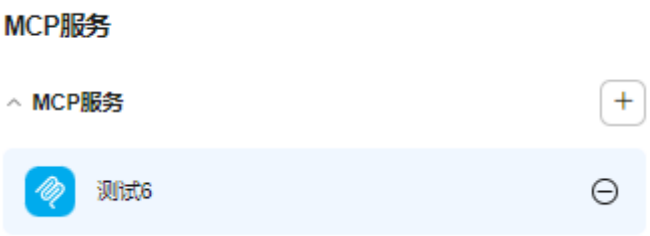
图 7-43 添加 MCP 服务



也可在“添加MCP服务”窗口中，通过单击“创建MCP”后的下拉框选择创建MCP服务，您可以选择基于官方预置MCP或空白模板进行创建，具体步骤参见**创建MCP服务**，创建成功后，通过单击提示信息中的“直接添加”，可立即将新创建的MCP添加至当前智能体中。或通过第三方服务安装，具体步骤可参见**使用资产中心的MCP资源**。

步骤5 添加MCP服务后，可在“MCP服务”中查看当前已添加的MCP服务。

图 7-44 已添加预置 MCP 服务



步骤6 MCP添加成功后，在提示词模块修改人设与回复逻辑，指示智能体调用“MCP服务”处理问题。

----结束

移除 MCP 服务

步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏中选择“开发中心 > 应用管理 > 单智能体应用”。

步骤3 单击目标单智能体应用，在“MCP服务”模块，在已添加的MCP服务右侧单击 。

页面提示“删除MCP服务成功”则表示MCP服务已移除。

图 7-45 移除 MCP 服务

----结束

相关文档

Versatile中配置MCP服务的详细信息，请参考MCP。

7.8 提升应用对话体验

配置开场白

开场白是用户进入智能体应用后首先看到的引导信息，能够帮助用户迅速了解智能体应用的功能和用途，明确如何与智能体应用进行有效交互。开场白需要简洁明了，语气友好并且表述清晰。

在“对话体验 > 开场白”中，可填写开场白，也可单击“智能添加 > 确定”按钮智能添加开场白。

例如，“您好！我是您的智能助手，很高兴为您提供帮助。请告诉我，今天有什么我可以为您做的吗？”

图 7-46 配置开场白

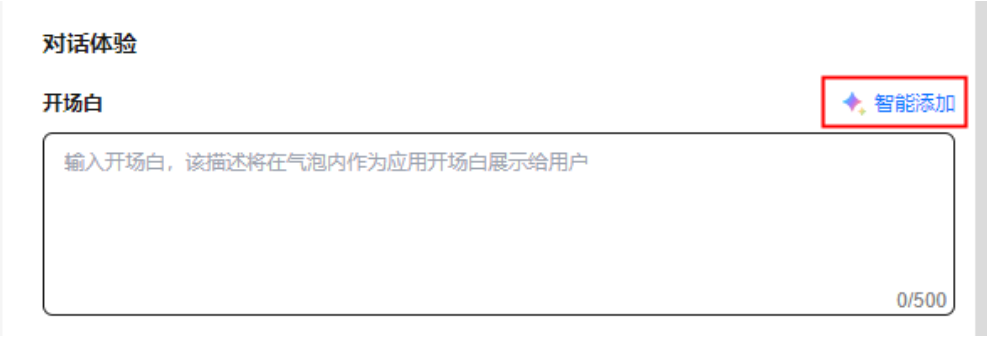


图 7-47 开场白配置示例

配置推荐问题

推荐问题是用户首次与应用互动时，应用主动展示的一些问题或话题建议，提供预设选项，减少用户打字负担。

在“对话体验 > 推荐问题”中，可填写推荐问题，也可单击“智能添加 > 确定”按钮智能添加推荐问题。推荐问题至多配置3条。

例如，“请告诉我您需要什么帮助？如：帮我预订会议室、帮我查询天气预报。”

图 7-48 配置推荐问题

图 7-49 推荐问题配置示例

配置追问

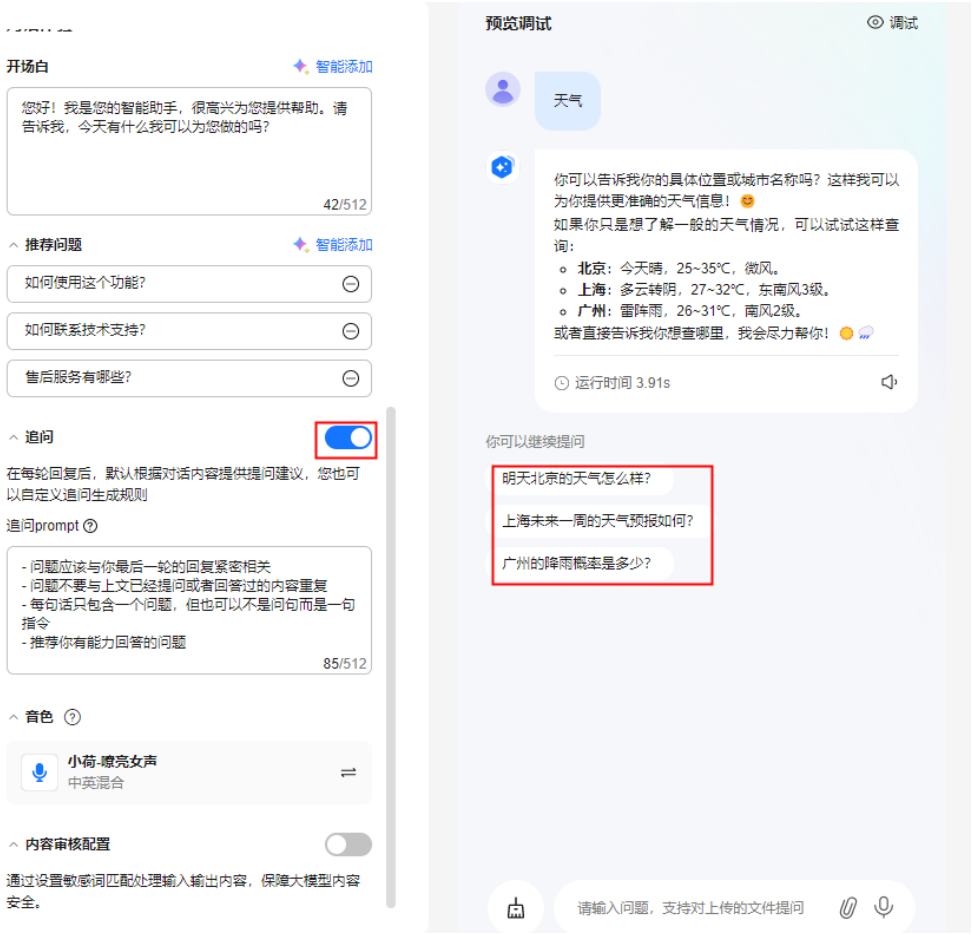
追问功能是指智能体在与用户交互过程中，根据用户的回答或上下文，主动提出进一步的问题，以获取更多信息或澄清用户需求。这一功能能够提升对话的深度和准确性，帮助智能助手更好地理解用户意图，从而提供更高效的服务。

用户自定义追问生成规则：用户可以根据需求，自定义追问生成规则。通过这些规则，智能体在每轮回复后，能够根据对话内容智能地提供提问建议，从而挖掘用户的潜在需求。

功能开启设置：在“对话体验 > 追问”中，用户可以选择是否开启“追问”功能。如果开启，模型将在每轮回复后，默认根据对话内容提供提问建议，进一步优化对话体验。

图 7-50 配置追问

图 7-51 配置追问示例



配置音色

配置音色，支持为智能体指定预置音色，用于配置智能应用调试对话模型返回结果的朗读音色，用户可以在应用开发过程中选择可选音色实现智能体与用户之间的自然语音交互。

图 7-52 配置音色

图 7-53 预置音色

配置安全信息

内容审核配置

内容审核配置，通过设置关键词匹配处理输入输出内容，可以过滤掉不恰当、敏感或违法的信息，保护用户免受不良信息的影响，同时保障大模型内容安全。支持通过单击右侧的开关按钮“启动”或“关闭”内容审核配置功能。

内容审核配置功能开启时，可通过单击“配置”设置关键词匹配处理输入输出内容。

- **过滤**：将大模型输出内容字段屏蔽掉后再返回给用户。
- **替换**：将大模型输出的关键词替换为设置的字段。
- **兜底回复**：触发关键词后，将直接返回已配置的兜底回复内容。

图 7-54 内容审核配置

⚠ 注意

- 审核内容输入时需要用“，”隔开。
- 内容审核和安全护栏无法同时开启，打开内容审核配置开关后，“安全防护”将自动关闭。

- **安全护栏**

  安全护栏的功能主要用于检测和拦截潜在的有害、敏感或攻击性的内容。具体来说，它能够识别并阻止那些旨在操纵或滥用系统的Prompt攻击，同时也能过滤掉包含有毒、不适当或违法信息的输入和输出，从而保护用户和系统免受不良影响。这一机制对于维护平台的健康环境和保障用户安全至关重要。

  ⚠ 注意

  内容审核和安全护栏无法同时开启，打开当前开关后，“内容审核”将自动关闭。

7.9 调试应用

开发者可以在智能体应用搭建完成后，直接与智能体应用进行对话，实时观察其执行过程和响应效果，并根据需要对配置进行优化和调整。平台全链路调试功能，允许开发者查看每个用户请求从输入到响应的完整流程，包括意图识别、知识检索等详细信息，从而准确定位问题并快速调整配置。

调试应用

创建应用后，平台支持对应用执行过程进行预览调试。

步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏中选择“开发中心 > 应用管理 > 单智能体应用”。

步骤3 单击目标单智能体应用，在“预览调试”界面文本框中输入对话，Agent应用将根据对话生成相应的回答。

图 7-55 调试应用

说明

- 预览调试界面支持文本输入、语音输入、文件输入：
  - 文本输入：在对话输入框输入对话后按Enter键或单击  ，查看应用响应结果。
  - 语音输入：用户可以通过语音进行输入。该功能支持多种语言（如中文、英文等），并提供语音识别、错误纠正和实时反馈等功能。
    - 首次使用语音输入须开通系统麦克风、扬声器权限，可在权限申请弹窗一键开通。
    - 语音输入最长为60秒，超时则取消语音输入状态，用户需重新录入。
  - 文件输入：用户可以通过上传文件进行提问，支持对文件进行解析，并根据文件内容和问题生成准确的答案。
    - 支持上传image、audio、excel、csv、docx等格式的文件。
    - 最多支持上传10个文件。
- 调试结果支持朗读功能，单击  ，Agent应用将按照设置的音色将文字转换成语音播放。
- 单击  ，支持对变量进行编辑或重置。
- 单击试运行页面左下角  ，一键清除试运行界面内容。

步骤4 在调试过程中，单击右上角“调试”，可以查看当前会话或历史会话的运行结果与调用详情。

图 7-56 查看调试结果

说明

用户可根据调试结果查看是否符合预期，如果需调整智能体配置请参考**选择并配置模型**。

----结束

调试信息说明

“调试”界面支持查看“运行结果”和“调用详情”。

- **运行结果**

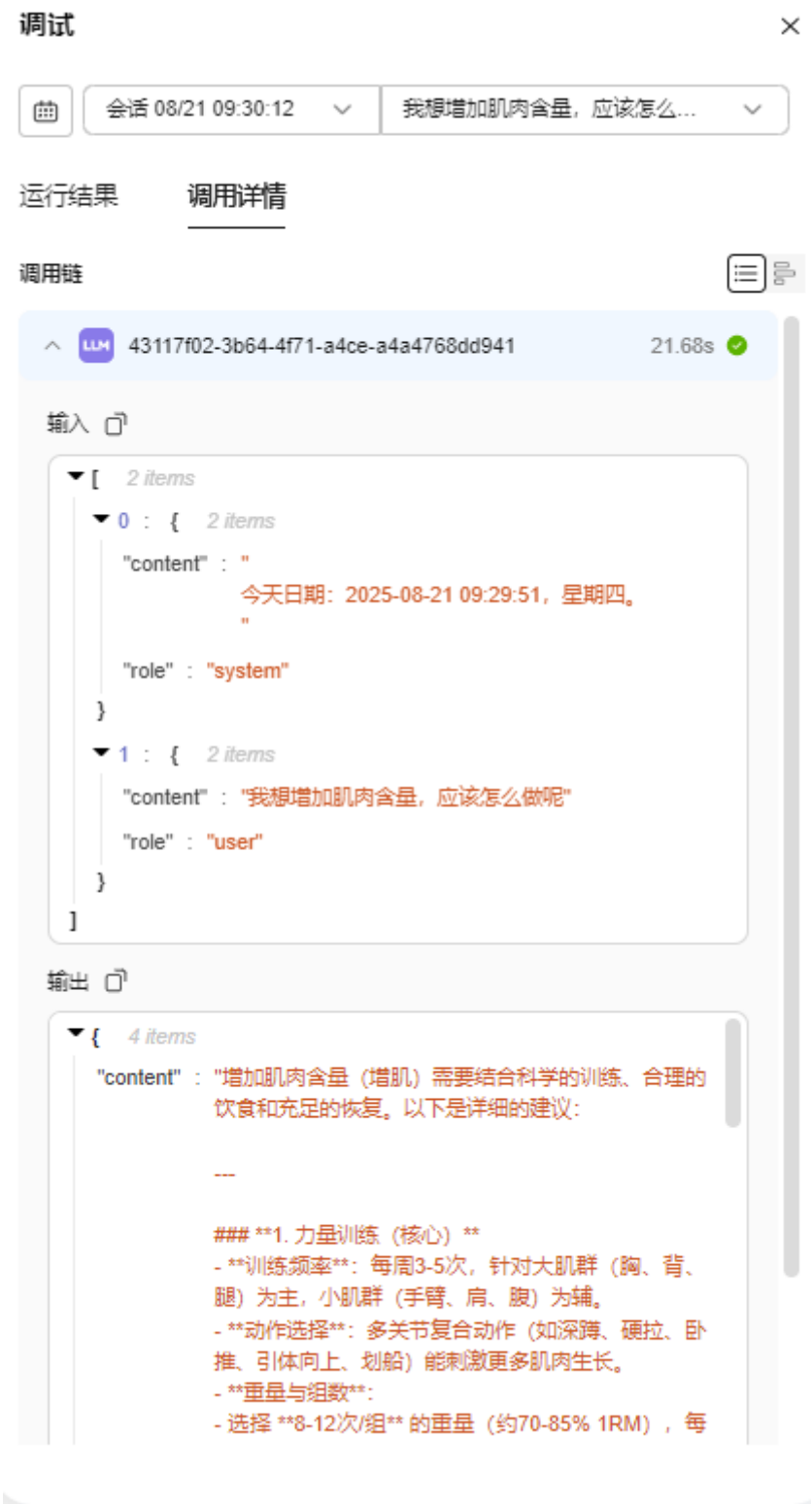
  运行结果中可以看到应用的执行开始时间、结束时间、运行时间等信息，还能看到输入和输出信息，从而直观的认识性能的情况。

图 7-57 查看运行结果

- **调用详情**

  在触发应用时，调用链中展现具体事件的详细信息，包括触发的组件、事件耗时、事件的输入和输出信息等。便于开发者快速地追溯操作顺序并精确定位问题。

图 7-58 查看调用详情



说明

- 单击调用详情页面中的  按钮，查看列表调用链。
- 单击调用详情页面中的  按钮，查看火焰图调用链。
- 支持在**查看应用调用链信息**页面中，查看该调用链的详细信息，具体操作请参见**使用过滤器筛选信息**。

7.10 配置触发器

触发器是Agent中实现任务自动化的关键功能，它能够在特定条件下自动启动任务执行，无需人工干预。可以利用触发器灵活设置任务的启动条件，触发器能确保任务按时按需完成，从而提高Agent应用的自动化水平和响应速度。

触发方式

通过定时任务调用应用按照触发指令要求执行任务。

添加触发器

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏中选择“开发中心 > 应用管理 > 单智能体应用”。

步骤3 单击目标单智能体应用，在智能体应用右上角单击触发器配置按钮  。

步骤4 在触发器配置页面右侧，单击+。

步骤5 配置触发器信息，参数配置说明请参考**表7-10**。

表 7-19 创建触发器参数说明

参数	说明
触发器名称	触发器的名称。 由2~20个字符组成，支持中英文、数字、下划线，仅支持中英文开头。
触发类型	支持周期触发、间隔触发。

参数	说明
触发时间	按设置的时间触发智能体应用的执行。例如，设置触发时间为每1小时执行一次，则每隔1小时，重复执行一次会话。 • 周期触发 – 用户可自定义触发时间按每日执行。 – 用户可自定义触发时间按每周执行。 – 用户可自定义触发时间按每月执行。 • 间隔触发：用户可自定义触发间隔时间，支持间隔“天”、“小时”、“分钟”、“秒”。 – 间隔“天”：取值范围1~31。 – 间隔“小时”：取值范围1~23。 – 间隔“分钟”：取值范围1~59。 – 间隔“秒”：取值范围1~59。
机器人提示	输入自然语言指令，触发时Agent遵循该指令定时执行。如：发送月底发票报销提醒，则机器人会在设定的时间发送提醒。

图 7-59 创建定时任务触发器

步骤6 单击“确定”，即完成触发器创建，可在触发器列表中对触发器进行查看，修改和删除等操作。

----结束

7.11 发布应用

7.11.1 发布应用为 API 服务

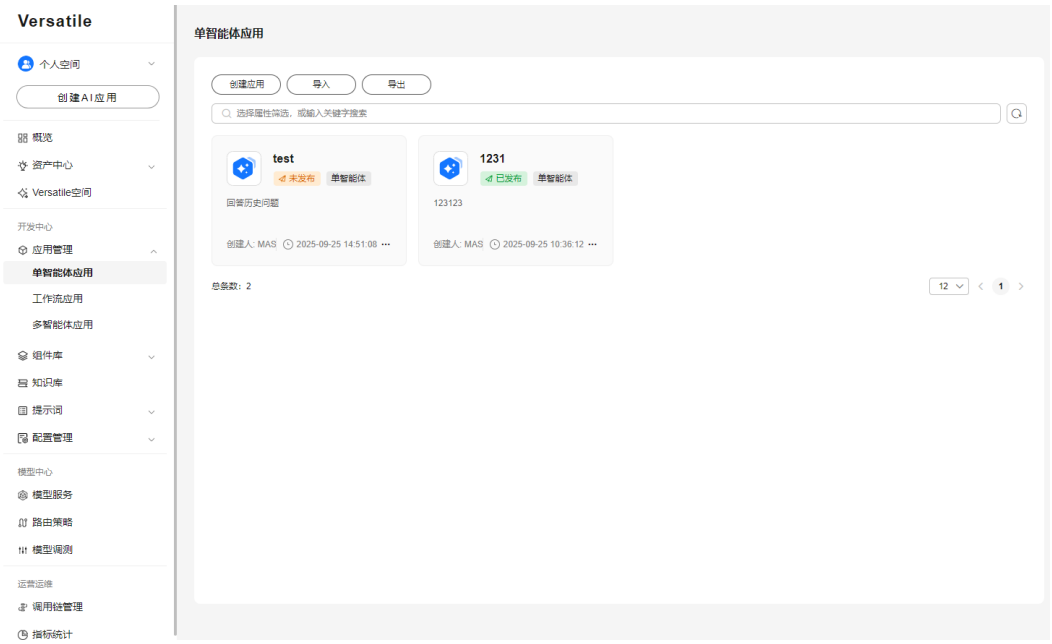
将Agent应用发布为API服务后，可以通过调用OpenAPI的方式使用Agent程序。本文介绍如何将开发完成的Agent应用发布为API服务。

发布 Agent 应用为 API 服务

步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏中选择“开发中心 > 应用管理 > 单智能体应用”进入应用开发主页面。

图 7-60 单智能体应用开发页面



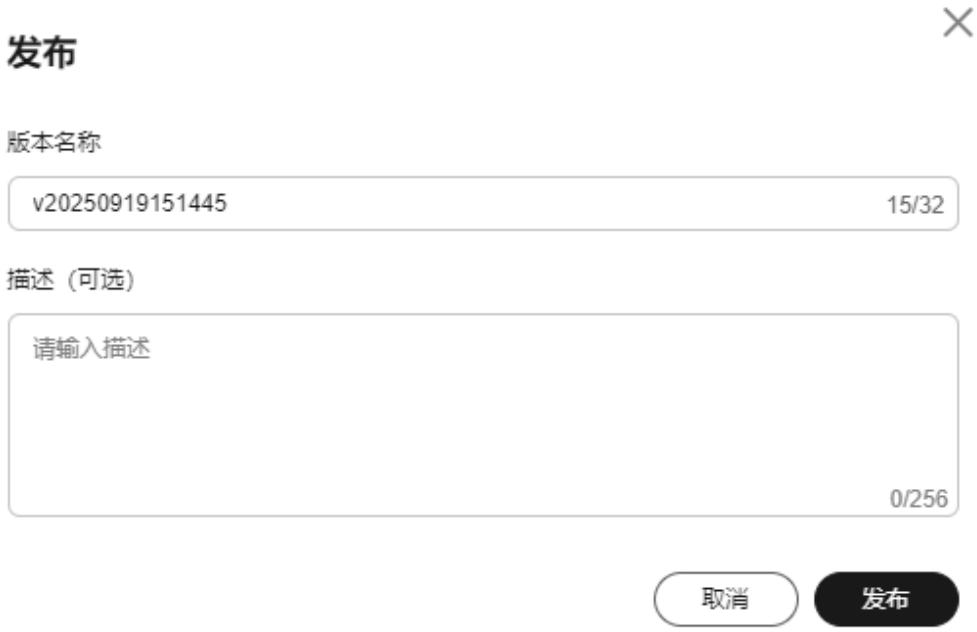
步骤3 应用开发主页面，选择已创建的Agent应用或单击左上角的“创建应用”按钮。创建应用时需在弹出的创建应用子窗口中填入“应用名称”、“应用描述”后单击“确定”进入应用编辑页面。

图 7-61 创建应用

步骤4 在应用编辑页面完成该应用的功能编辑调试，然后单击右上角的“发布”按钮。在弹出的发布信息填写提示窗中填写本发布的“版本名称”、“描述”，单击“发布”按钮完成发布。

图 7-62 编辑 Agent 应用

图 7-63 填写发布信息



说明

已发布的应用支持在Agent应用编辑页面选择“更新发布”按钮重新发布应用。

步骤5 发布完成后跳转至发布管理页面，可以查看API调用接口信息。

也可通过左侧导航栏中选择“开发中心 > 应用管理 > 单智能体应用”，单击目标应用，进入应用主页面，单击右上角“发布管理”按钮，可进入发布管理页面。

图 7-64 调用 API

----结束

通过 API 运行应用

应用发布为API服务之后，可通过API调用单智能体应用。详细说明可参考**使用API调用单智能体应用**。

7.11.2 发布应用为网页应用

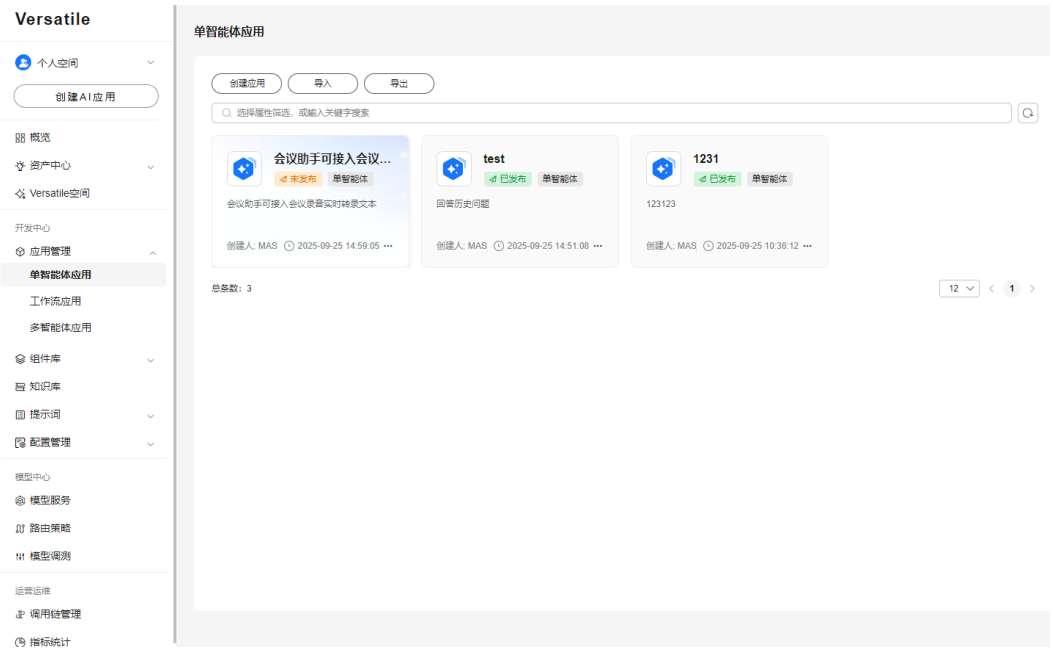
Versatile支持将Agent应用发布为网页应用，这一功能适用于多种场景。例如，在客户服务领域，可以将智能客服Agent部署为网页应用，提供全天候的在线咨询服务，提升客户满意度。在教育和培训中，可以将教学助手Agent部署为网页应用，为学生提供个性化的学习指导和互动体验。在智能问答系统中，可以将知识库查询Agent部署为网页应用，为用户提供快速准确的信息检索服务。此外，还可以将Agent应用部署为网页应用，用于数据分析、内容推荐、图像识别等场景，为用户提供便捷的交互界面和高效的服务。通过这种方式，Versatile帮助开发者快速将Agent应用转化为实际可用的在线服务，提高应用的可用性和用户满意度。

发布应用为网页应用

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏中选择“开发中心 > 应用管理 > 单智能体应用”，进入应用开发主页面。

图 7-65 单智能体应用开发页面



步骤3 在应用开发主页面，选择已创建的Agent应用或单击左上角的“创建应用”。创建应用时需在弹出的创建应用子窗口中填入“应用名称”、“应用描述”后单击“确定”进入应用编辑页面。

图 7-66 创建应用

步骤4 在应用编辑页面完成该应用的功能编辑调试，然后单击右上角的“发布”按钮。在弹出的发布信息填写提示窗中填写发布的“版本名称”、“描述”，单击“发布”按钮完成发布。

图 7-67 编辑 Agent 应用

图 7-68 填写发布信息

步骤5 发布完成后跳转至“发布管理”页面，也可通过左侧导航栏中选择“开发中心 > 应用管理 > 单智能体应用”，单击目标应用，进入应用主页面，单击右上角“发布管理”按钮，可进入发布管理页面。

在**发布管理**页面，**网页**发布渠道的**操作**列单击“发布”按钮后将出现“立即访问”、“复制链接”、“重新生成”文字按钮。这里可以通过两种方式访问Agent应用的网页应用链接，同时支持重新生成发布链接。

图 7-69 发布 Agent

- 立即访问：单击当前页面的“立即访问”按钮，可立即进入网页版应用。

  图 7-70 立即访问网页 Agent 应用

- 复制链接：单击当前页面的“复制链接”按钮，可直接复制网页URL嵌入或分享到其他应用场景。

图 7-71 复制 Agent 访问地址

- 重新生成：重新生成Agent应用发布链接。

说明

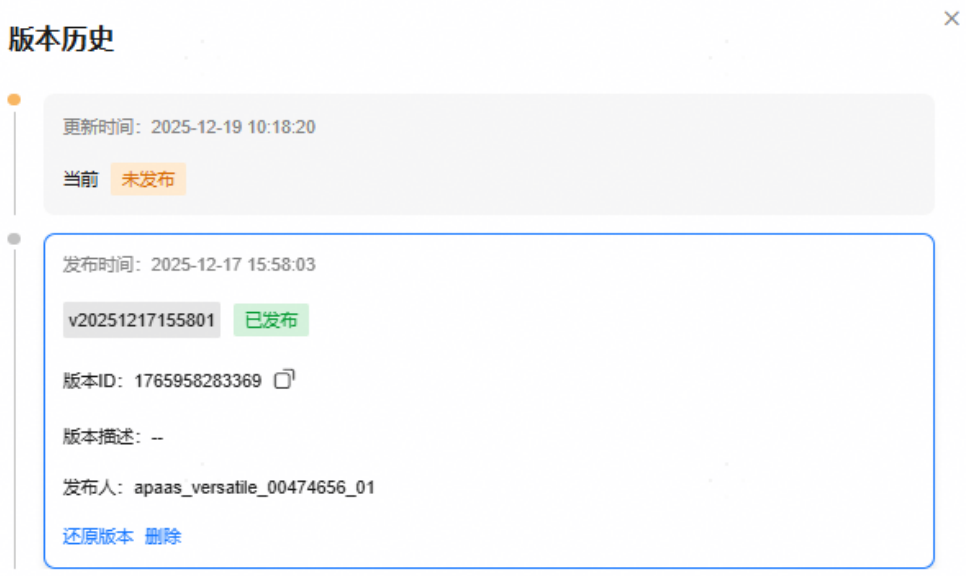
- 已发布的应用支持在Agent应用编辑页面选择“更新发布”按钮重新发布应用。也可在“是否更新版本”列单击“更新版本”，并在确认界面单击“更新版本”，该发布渠道将更新为最新版本配置。
- “资源配置”列，单击 可进行资源配置设置，支持配置“每日调用限额（次）”、“分享范围”。
  - “每日调用限额（次）”：支持用户自定义修改，默认100次。支持输入-1或0~10000，其中-1表示调用次数无限制。
  - “分享范围”：可选择当前租户可用和所有租户可用。

----结束

查看发布历史

单击右上角 ，可查看当前单智能体发布历史记录。

图 7-72 发布历史



说明

发布历史支持还原版本和删除操作：

- 还原版本：单击“还原版本”，并在弹窗中单击“确定”，将还原当前单智能体应用至配置前状态，单智能体应用配置信息将不再保留，请谨慎操作。
- 删除：单击“删除”，并在弹窗中单击“确定”，将删除当前单智能体应用发布历史。

7.11.3 发布应用至云商店

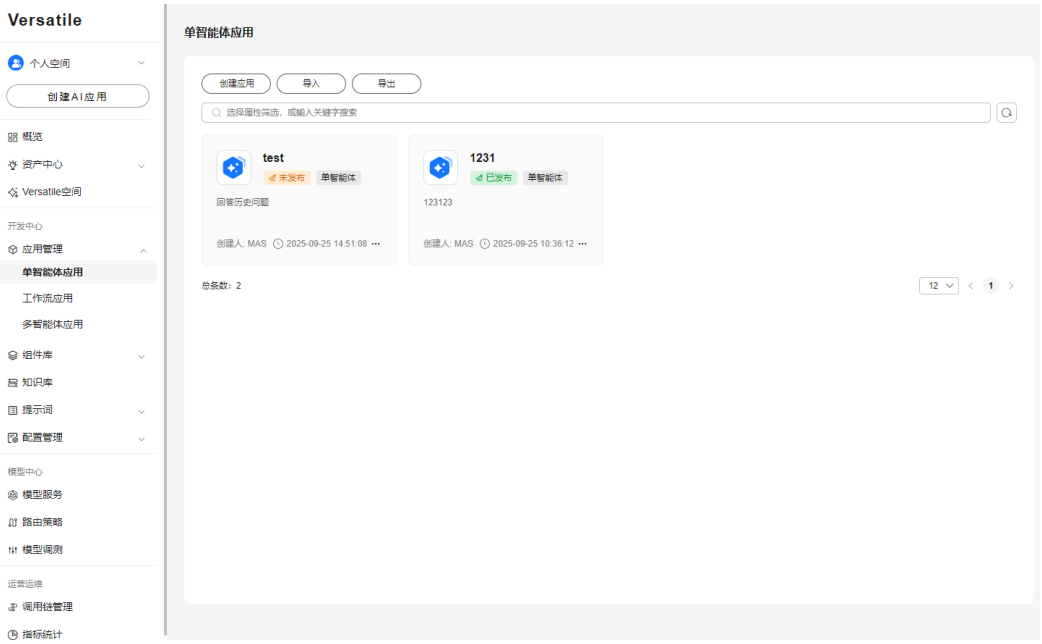
将Agent应用发布到云商店后，生成OpenAPI URL，用户可以在华为云云商店通过调用OpenAPI URL，即可将应用发布到华为云云商店。

发布 Agent 应用至云商店

步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏中选择“开发中心 > 应用管理 > 单智能体应用”进入应用开发主页面。

图 7-73 单智能体应用开发页面



步骤3 应用开发主页面，选择已创建的Agent应用或单击左上角的“创建应用”按钮。创建应用时需在弹出的创建应用子窗口中填入“应用名称”、“应用描述”后单击“确定”进入应用编辑页面。

图 7-74 创建应用

步骤4 在应用编辑页面完成该应用的功能编辑调试，然后单击右上角的“发布”按钮。在弹出的发布信息填写提示窗中填写本发布的“版本名称”、“描述”，单击“发布”按钮完成发布。

图 7-75 编辑 Agent 应用

图 7-76 填写发布信息

步骤5 发布完成后跳转至“发布管理”页面，也可通过左侧导航栏中选择“开发中心 > 应用管理 > 单智能体应用”，单击目标应用，进入应用主页面，单击右上角“发布管理”按钮，可进入发布管理页面。

在**发布管理**页面，**云商店**发布渠道的**操作**列单击“发布”按钮后将出现“立即访问”、“复制链接”、“重新生成”文字按钮。

图 7-77 发布 Agent 应用

- 立即访问：单击当前页面的“立即访问”按钮，可立即跳转至我的云商店界面。

图 7-78 立即访问 Agent 应用

说明

华为云商店发布操作详见**发布API类产品**。

- 复制链接：单击当前页面的“复制链接”按钮，可直接复制网页URL嵌入或分享到其他应用场景。

图 7-79 复制 Agent 访问地址

- 重新生成：单击当前页面的“重新生成”按钮，可重新生成该应用。

图 7-80 重新生成 Agent

说明

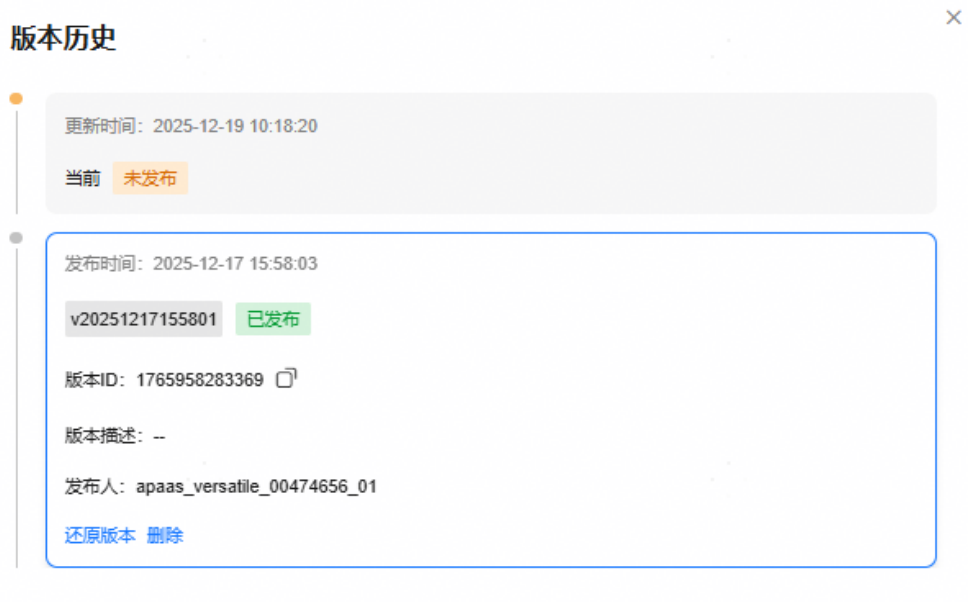
- 已发布的应用支持在Agent应用编辑页面选择“更新发布”按钮重新发布应用。也可在“是否更新版本”列单击“更新版本”，并在确认界面单击“更新版本”，该发布渠道将更新为最新版本配置。
- “资源配置”列，单击 可进行资源配置设置，支持配置“每日调用限额（次）”、“分享范围”。
  - “每日调用限额（次）”：支持用户自定义修改，默认100次。支持输入-1或0~10000，其中-1表示调用次数无限制。
  - “分享范围”：可选择当前租户可用和所有租户可用。

----结束

查看发布历史

单击右上角 ，可查看当前单智能体发布历史记录。

图 7-81 发布历史



说明

发布历史支持还原版本和删除操作：

- 还原版本：单击“还原版本”，并在弹窗中单击“确定”，将还原当前单智能体应用至配置前状态，单智能体应用配置信息将不再保留，请谨慎操作。
- 删除：单击“删除”，并在弹窗中单击“确定”，将删除当前单智能体应用发布历史。

7.12 通过 API 调用单智能体应用

Versatile提供Open API请求方式，可通过调用路径发送请求，程序将调用应用并返回预期结果。

Versatile的API调用是应用开发中的强大工具，可以帮助用户快速集成功能和服务，同时支持与其他系统或服务进行交互，提升应用性能和用户体验。合理设计和管理API是确保应用安全和稳定的关键。通过API，用户可以构建功能丰富、高效的应用，满足多样化的用户需求。

前提条件

在调用应用前，须确保应用已发布，具体请参考**发布应用为API服务**。

获取应用 ID 和调用路径

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏中选择“开发中心 > 应用管理 > 单智能体应用”，选择目标单智能体应用。

步骤3 单击“ ··· > 复制ID”，可获取当前应用ID。请记录保存，用于填写调用Agent应用接口的agent_id字段。

图 7-82 复制 ID

步骤4 单击“ ··· ”>“调用路径”，在弹出的“调用路径”页面，单击“复制路径”即可获取调用路径，如图7-83所示。

其中，“e0d10764-7753-48ad-b25f-ce63bf62eec9”是随机生成的字符串，在使用时可以替换为其他的字符串。字符串长度为1~64个字符，支持英文字母、数字、中划线、下划线。在智能体应用被API调用时，为“conversation_id”的值。

图 7-83 获取应用调用路径

----结束

使用 API 调用单智能体应用

使用API调用单智能体的操作，请参考**调用智能体应用**。

7.13 管理应用

在Versatile中创建应用之后，可以管理单智能体应用界面中的应用，可执行删除、复制创建的应用、复制应用ID、导入、导出及查看调用路径等。

前提条件

已**购买Versatile智能体平台**。

复制应用

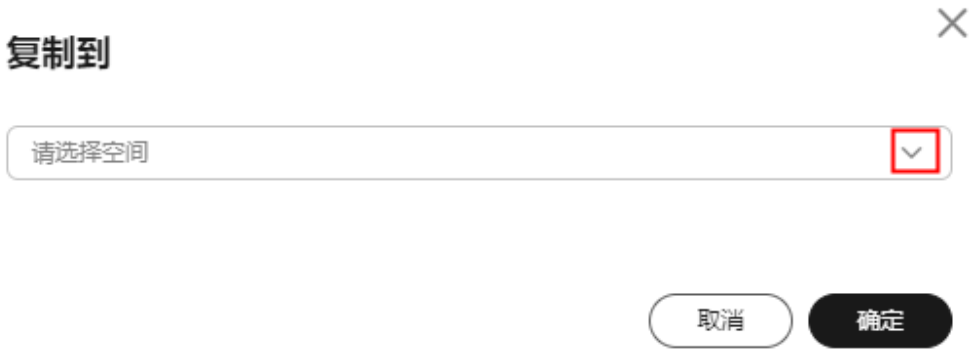
步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧菜单栏，单击“开发中心 > 应用管理 > 单智能体应用”。

步骤3 在Agent开发空间，选择待复制的应用，单击应用的右下角 ··· 展开功能列表，选择“复制”。

步骤4 在“复制到”下拉框中选择已创建的目标空间。

图 7-84 复制应用



说明

在复制到目标空间时，应用的配置参数、插件等数据将一并复制，且复制后的应用需要单独发布。

----结束

删除应用

须知

- 只有应用的所有者可以删除应用。
- 删除应用时虽然不会同步删除应用资源库中的所有资源，但不可恢复，请谨慎操作。

步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧菜单栏，单击“开发中心 > 应用管理 > 单智能体应用”。

步骤3 在Agent开发空间，选择待删除的应用，单击应用的右下角 ··· 展开功能列表，选择“删除”。

步骤4 弹出的对话框中单击“确定”。

图 7-85 删除确认

----结束

复制 ID

对于Agent应用除了具有页面操作的能力之外，还具有Chat API调用能力，对于AppID获取就十分必要。该ID为调用Agent应用接口的agent_id字段。

```
POST /v1/{project_id}/agents/{agent_id}/conversations/{conversation_id}
```

步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧菜单栏，单击“开发中心 > 应用管理 > 单智能体应用”。

步骤3 在工作台开发空间，选择需要的应用，单击应用的右下角 ··· 展开功能列表，选择“复制ID”

步骤4 弹出复制成功对话框，用于填写调用Agent应用接口的agent_id字段。

图 7-86 复制 ID

----结束

调用路径

调用路径可为Agent应用的API接口。详细API调用过程请参见使用API调用单智能体应用。

图 7-87 获取调用路径

导入应用

平台支持导入单智能体应用。导入单智能体应用时，将同步导入单智能体应用关联的插件等配置。

步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 进入“开发中心 > 应用管理 > 单智能体应用”页面。

步骤3 导入单智能体应用。

1. 单击页面左上角“导入”。
2. 在“导入”页面，单击“选择文件”选择需要导入的jsonl格式文件。
3. 选择导入文件后，选择解析内容。

   平台将自动解析jsonl文件。如果解析的文件已存在，勾选该文件将自动覆盖平台现有文件。
4. 单击“导入”，导入成功的单智能体应用将在“开发中心 > 应用管理 > 单智能体应用”页面中展示。

   说明

   仅支持上传jsonl格式文件，文件的最大导入大小为128MB。

----结束

导出应用

平台支持导出单智能体应用。导出单智能体应用时，将同步导出单智能体应用关联的插件等配置。

步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 进入“开发中心 > 应用管理 > 单智能体应用”页面。

步骤3 导出单智能体应用。

1. 单击页面左上角“导出”。
2. 在“导出”页面选择单智能体应用，单击“导出”。工作流将以一个jsonl格式的文件下载至本地。

----结束

发布管理

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 进入“开发中心 > 应用管理 > 单智能体应用”页面。

步骤3 在Agent开发空间，选择待部署的应用，单击应用的右下角 ··· 展开功能列表，选择“发布管理”，跳转至“发布管理”页面。详细操作可参见**发布应用**节点。

----结束

8 开发工作流应用

8.1 工作流介绍

工作流是一系列相互关联的步骤，用于实现业务逻辑或完成特定任务。它为应用/智能体的数据流动和任务处理提供了一个结构化框架。

工作流的核心是一个由人工设计和编排的智能系统，由多个AI Agent通过利用自然语言处理（NLP）和大型语言模型（LLM）协作完成任务。这些智能体在预设的逻辑框架下工作，能够根据既定规则自主感知、推理和行动，以追求特定目标，形成强大的集体智慧，可以打破信息孤岛，集成不同的数据源，并提供无缝的端到端自动化。

Versatile平台提供一个可视化的画布，可以通过拖拽节点迅速搭建工作流。同时，支持在画布实时调试工作流。在工作流画布中，可以清晰地看到数据的流转过程和任务的执行顺序。

工作流应用预置了1个应用模板，名为“预制模板-文章写作助手”，占用1个套餐内的应用数量，模版支持用户修改、复制、删除等操作。

工作流节点介绍

Versatile的工作流由多个节点构成，节点是组成工作流的基本单元。平台支持多种节点，根据功能分为基础节点、通用节点、逻辑节点、工具节点、消息管理节点和数据&知识节点，具体节点功能详见**表8-1**。

表 8-1 配置节点

分类	节点名称	节点说明
基础节点	开始节点	开始节点是工作流的起始节点，用户输入的信息由开始节点传入，详细配置参见**开始和结束节点**。
	结束节点	结束节点是工作流的最终节点，用于定义整个工作流的输出信息，详细配置参见**开始和结束节点**。
通用节点	大模型节点	用于在工作流中引入大模型能力，详细配置参见**大模型**。

分类	节点名称	节点说明
	工作流节点	实现工作流嵌套工作流的效果，详细配置参见**工作流**。
	Agent节点	用于对用户任务进行动态规划，通过分解用户原始输入、调用插件等完成一个复杂任务的自动解析处理，详细配置参见**Agent**。
逻辑节点	判断节点	编排应用时作为分支切换节点，可以根据输入满足的判断条件，指定执行对应的工作流分支，详细配置参见**判断**。
	意图识别节点	用于根据用户的输入进行意图分类并导向后续不同的处理流程，详细配置参见**意图识别**。
	代码节点	用于引入代码执行器，根据节点的输入，执行Python代码或Node.js代码，节点的输出是代码执行的结果信息，详细配置参见**代码**。
	高级意图识别节点	用于根据用户大量可归类的输入进行意图分类并导向后续不同的处理流程。适用于编排大于20个以上意图的分支逻辑。详细配置参见**高级意图识别**。
	循环节点	循环节点用于重复执行一系列任务，详细配置参见**循环**。
工具节点	插件节点	用于引入API插件，根据节点的输入，执行用户定义的插件，将插件执行结果作为节点的输出，详细配置参见**插件**。
	MCP服务节点	支持从MCP服务中选择您已配置好的或预置MCP服务，并选择所需的工具完成调用，详细配置参见**MCP服务**。
消息管理节点	消息节点	定义一段文本内容，在工作流的执行过程中向用户发送该内容的消息，详细配置参见**消息**。
	提问器节点	提供了在对话过程中向用户收集更多信息的能力，详细配置参见**提问器**。
	输入节点	输入节点用于在工作流运行时收集用户输入，详细配置参见**输入**。
	问答节点	问答节点可提供中间过程的向用户提问的能力，详细配置参见**问答**。
	对象提取	用于提取指定对象中的参数，并支持配置子工作流进行参数的校验与校准，以及发起用户交互，详细配置参见**对象提取**。
	异常节点	异常节点允许用户根据业务需求灵活设置和抛出详细的异常信息，详细配置参见**异常**。
变量&知识节点	变量赋值节点	变量赋值节点用于在循环执行过程中动态设置中间变量，详细配置参见**变量赋值**。

分类	节点名称	节点说明
	变量聚合节点	变量聚合节点能够将多路分支的输出变量整合为一个，方便下游节点统一配置，详细配置参见**变量聚合**。
	知识检索节点	可以根据输入参数从指定知识库内召回匹配的信息，详细配置参见**知识检索**。
数据库	数据查询节点	用于查询数据库数据，用户可配置查询条件和查询方式，详细配置参见**数据查询**。

配置方式

创建工作流时，每个节点需要配置不同的参数，如输入和输出参数等，开发者可通过拖、拉、拽可视化编排更多的节点，实现复杂业务流程的编排，从而快速构建应用。

工作流方式主要面向目标任务包含多个复杂步骤、对输出结果成功率和准确率有严格要求的复杂业务场景。

在编排工作流时，根据功能需要使用节点进行设计。

8.2 对话型工作流和任务型工作流

平台提供了两种类型的工作流，即对话型工作流和任务型工作流，用户可以针对不同的任务或场景选择适合的工作流进行搭建。

- 对话型工作流：面向多轮交互的开放式问答场景，基于用户对话内容提取关键信息，输出最终结果。适用于客服助手、工单助手、娱乐互动等场景。
- 任务型工作流：面向自动化处理场景，基于输入内容直接输出结果，无中间的对话交互过程。适用于内容生成、批量翻译、数据分析等场景。

应用限制

任务型工作流不支持配置**输入节点、消息节点、提问器节点、问答节点和Agent节点**。

对话型工作流和任务型工作流差异

表 8-2 对话型工作流和任务型工作流区别说明

差异项	对话型工作流	任务型工作流
适用场景	AI客服助手、虚拟助手、工单助手、娱乐互动等多轮交互的场景。	数据处理、批量生成、自动化报告、批量翻译、数据分析等场景。
节点	支持输入节点、消息节点、提问器节点和Agent节点。	不支持输入节点、消息节点、提问器节点和Agent节点。

差异项	对话型工作流	任务型工作流
试运行方式	试运行界面与任务型工作流不同。 如果“开始”节点有多个参数，先对除query参数外的参数进行配置，然后再以对话窗口的形式进行试运行。	如果“开始”节点有多个参数，在试运行时，需要对多个输入参数同时进行配置。

相关文档

对话型工作流、任务型工作流是否可以相互转换？请参考**对话型工作流、任务型工作流是否可以相互转换？**。

8.3 工作流使用限制

工作流是一系列相互关联的步骤，用于实现业务逻辑或完成特定任务。在使用过程中注意以下限制。

工作流使用限制

表 8-3 工作流使用限制

限制	说明
超时时间	工作流超时时间15分钟，插件超时时间50s， 模型超时时间15分钟，其他单节点无限制。
运行次数	循环节点自身最大循环次数为1000次。
节点总数	工作流中非游离节点个数最多为150个。
请求大小	请求查询大小上限为100000字符。

8.4 搭建工作流

8.4.1 工作流编排逻辑

业务逻辑是指应用程序中处理特定业务规则和操作的部分。它定义了应用如何根据业务需求处理数据、执行操作和做出决策。在Versatile中，业务逻辑的实现主要通过工作流来完成。

在Versatile中，左侧的资产中心、模型中心、开发中心的组件库、知识库、提示词、配置管理等，都称为“资源”。在工作流中，可以根据业务处理逻辑、业务数据等信息添加或创建资源，以完成相应的业务目标。

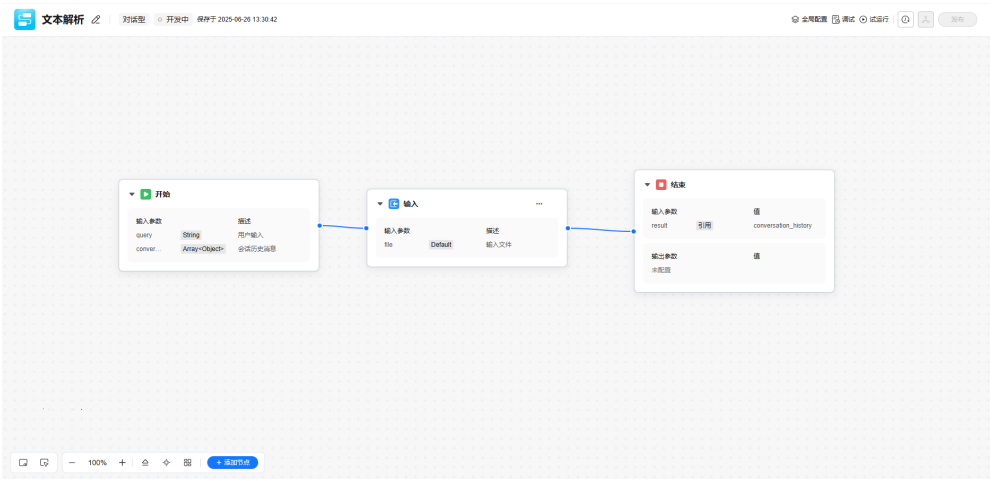
Versatile中资源之间可以通过流程搭建，业务连接，进行相互调用，在进行工作流编排的时候，可以通过拖拽将资源添加到工作面板中。

了解业务编排

工作流是业务逻辑的可视化表示，它决定了应用的输入和输出数据结构、数据接收和处理的规则以及决策流程。

例如：文本解析工作流通过添加输入资源，进行文本输入的简单处理。

图 8-1 编排示例



编排模式

Versatile中支持串行和并行两种编排模式。用户可根据需要选择适合的编排逻辑，对于复杂的任务，合理的并行与串行组合能显著提升系统效率。

表 8-4 编排模式对比

编排模式	功能	使用场景	优势
串行编排	任务按顺序一个接一个执行，前一个任务完成后才开始下一个任务。	串行编排适用于以下场景： ● 线性处理流程：每个步骤必须依次完成，前一个步骤的输出是后一个步骤的输入。 ● 依赖关系明确：例如订单处理完成后才能进行支付确认，支付确认完成后才能发货。 ● 逐步验证：每个步骤完成后需要进行验证，确保前一个步骤正确无误后才能进行下一步。 ● 资源限制：由于资源限制，任务必须依次执行，以避免资源冲突。 **说明** 适用场景不限于以上场景，其他符合业务逻辑的场景均可使用。	确保任务按照逻辑顺序执行，每个节点都基于前一个节点的输出结果展开工作。
并行编排	将LLM或知识检索或其它节点同时处理同一项任务，并在变量聚合中整合输出结果，从而提高任务处理的准确性和全面性。	并行编排适用于以下场景： ● 多任务处理：多个数据集可以同时进行处理，提高处理效率。 ● 资源充足：由于资源充足，可以同时处理多个任务，提高整体处理速度。 ● 并行计算：在分布式计算环境中，多个计算节点可以同时处理不同的子任务，提高计算效率。 **说明** 适用场景不限于以上场景，其他符合业务逻辑的场景均可使用。	将复杂任务拆分为子任务后，多个节点可在同一时间工作，不仅提高输出质量，同时通过并行处理的方式，能够提升输出的响应速度。

说明

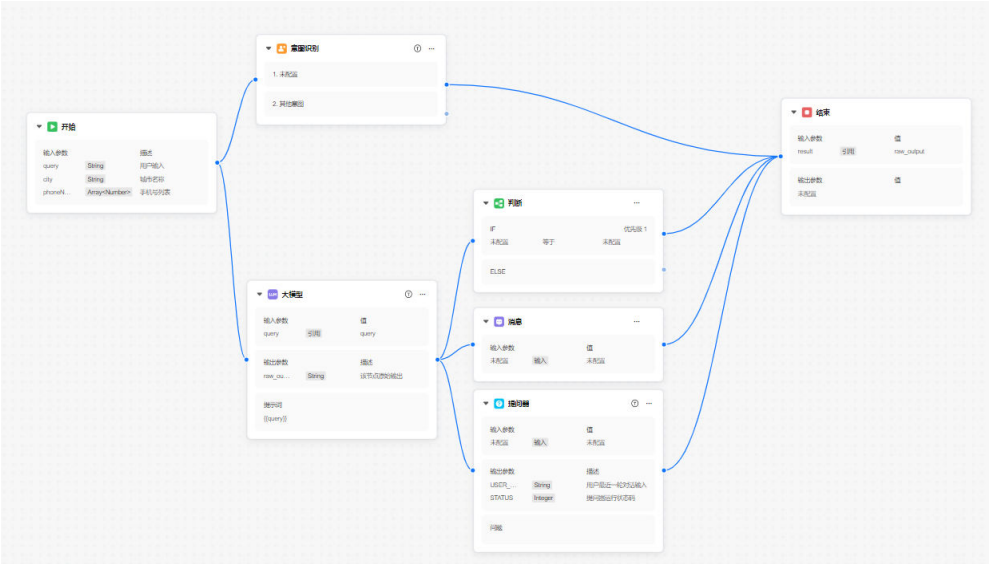
并行编排支持多种结构。

- 常规并行：只要三层关系，包含开始节点、并行结构、结束节点。开始节点输出结果后，多个并行节点同时执行多条任务。
- 嵌套并行：包含多层嵌套关系，包含开始节点、多并行结构、结束节点。开始节点输出结果后，与开始节点连接的任务开始执行，输出结果后传输至嵌套节点。

图 8-2 串行编排

图 8-3 常规并行编排

图 8-4 嵌套并行编排



8.4.2 创建工作流

工作流是一系列相互关联的步骤，用于实现业务逻辑或完成特定任务。可以在智能体和应用搭建中通过工作流实现特定的任务或指令。无论是在智能体还是应用中使用工作流，都需要先创建一个可运行的工作流。

Versatile支持创建工作流应用的方式如表8-5所示。

表 8-5 创建方式说明

创建方式	功能	优点	缺点	操作指导
从空白创建	基于平台可视化画布，可以通过拖拽节点和配置相关参数，迅速搭建工作流。	可控性强、透明度高。	开发成本高。	常规创建工作流
AI辅助创建	描述所需工作流具体应用场景和核心功能，平台会根据用户的需求，自动生成一个定制化工作流。	自适应能力强、用户体验好。	可控性差、数据依赖性强。	AI辅助创建工作流
使用预置应用创建	资产中心内置了工作流应用，用户可根据需要复制模板配置完全一样的工作流，并将其配置为符合自己需求的工作流应用。	高效的开发速度，低门槛。	高度定制化，无法满足所有个性化需求。	使用平台精选的工作流应用

前提条件

- 已购买Versatile智能体平台。
- 模型已接入Versatile平台，接入指导请参考接入自定义的模型服务。

常规创建工作流

步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 单击左侧导航栏“开发中心 > 应用管理 > 工作流应用”。

步骤3 单击左上角“创建应用”，在“创建应用”页面，选择创建类型，可选“对话型工作流”或“任务型工作流”，相关区别如表8-6所示。

表 8-6 对话型工作流和任务型工作流区别说明

差异项	对话型工作流	任务型工作流
适用场景	AI客服助手、虚拟助手、工单助手、娱乐互动等多轮交互的场景。	数据处理、批量生成、自动化报告、批量翻译、数据分析等场景。
节点	支持输入节点、消息节点、提问器节点和Agent节点。	不支持输入节点、消息节点、提问器节点和Agent节点。

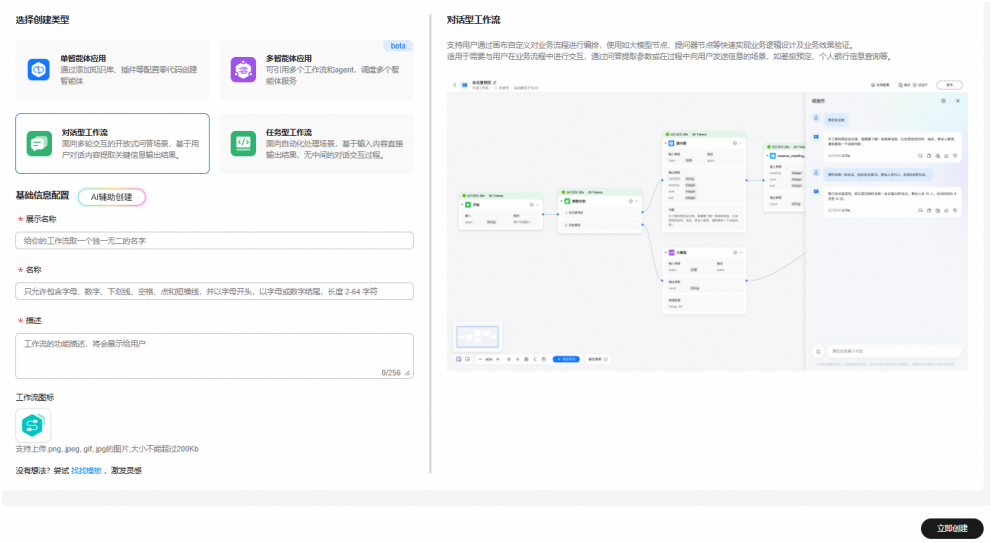
差异项	对话型工作流	任务型工作流
试运行方式	试运行界面与任务型工作流不同。 如果“开始”节点有多个参数，先对除query参数外的参数进行配置，然后再以对话窗口的形式进行试运行。	如果“开始”节点有多个参数，在试运行时，需要对多个输入参数同时进行配置。

步骤4 选择完成后配置应用基础信息，参数说明如表8-7所示。

表 8-7 基础信息配置说明

参数	说明	示例
展示名称	在工作流应用界面中工作流名称不允许重复，支持中英文、数字、下划线、中划线和空格，长度2~64字符，且名称首尾不能有空格。	智能客服单智能体
名称	输入内容只能包含英文字母、数字、下划线和空格，并以字母开头，长度2~64字符，且名称首尾不能有空格。	Intelligent customer service single agent
描述	描述工作流的功能，可直观呈现给用户，长度0~256。	智能客服智能体应用是用户与智能客服系统交互的界面。用户可以输入问题或发送请求，智能客服系统将自动响应并提供解决方案。
工作流图标	系统默认单智能体应用图标，用户也可以自定义图标。 1. 鼠标移动至系统默认图标上，单击鼠标左键。 2. 上传已准备好的应用图标。支持jpg、jpeg、png、gif格式图片，且不大于200KB。	-

图 8-5 创建工作流

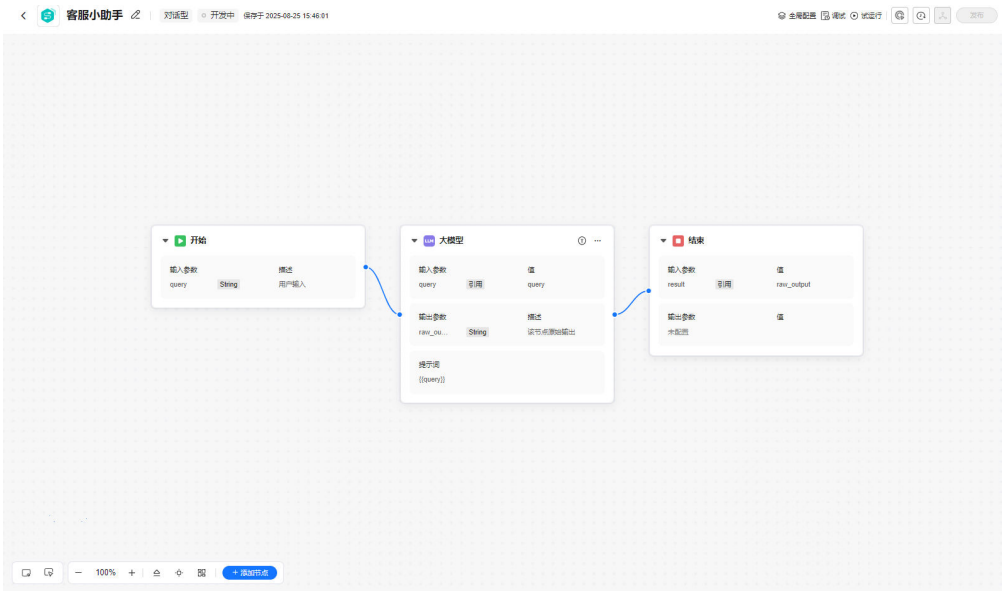


步骤5 配置完成后单击“立即创建”，进入工作流编排页面。

初始状态下工作流包含**开始**、**大模型**、**结束**节点。

- **开始**节点：用于启动工作流，详细配置请参考**开始节点**。
- **大模型**节点：（可选）提供了使用大模型的能力，可在节点中配置已部署的模型，用户可以通过编写提示词、设置参数让模型处理相应任务。如果无需配置，可单击右上角 “···” 删除节点，详细配置请参考**大模型**。
- **结束**节点：用于返回工作流的运行结果，详细配置请参考**结束节点**。

图 8-6 编排画布界面



----结束

AI 辅助创建工作流

须知

AI辅助创建仅支持创建对话型工作流。

在日常开发中，手动编写复杂流程图和代码逻辑既耗时又易出错。NL2Workflow通过自然语言描述业务需求，自动生成Mermaid格式流程图和DSL（JSON）代码，简化开发流程。系统支持多轮对话交互，确保流程准确无误，并允许用户调整和确认，确保最终生成的工作流既符合业务需求又具备可执行性。

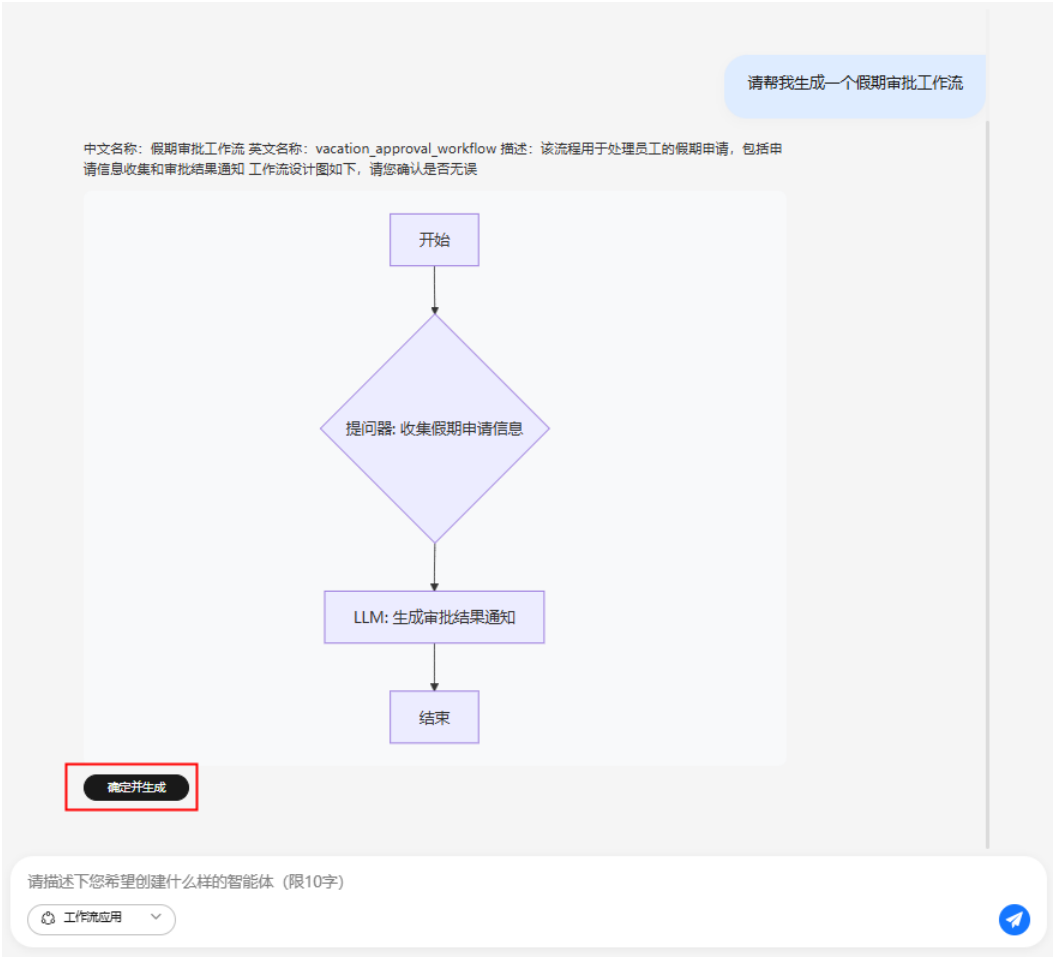
步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在概览页“一句话描述，即刻创建专属你的个性化智能体”区域，选择输入框上方的任务，或在输入框中输入任务，并在输入框左下角选择“模型”和“工作流应用”，单击，以“假期审批工作流”为例。

图 8-7 创建假期审批工作流

步骤3 在思考界面的输入框中输入任务。

图 8-8 发送指令

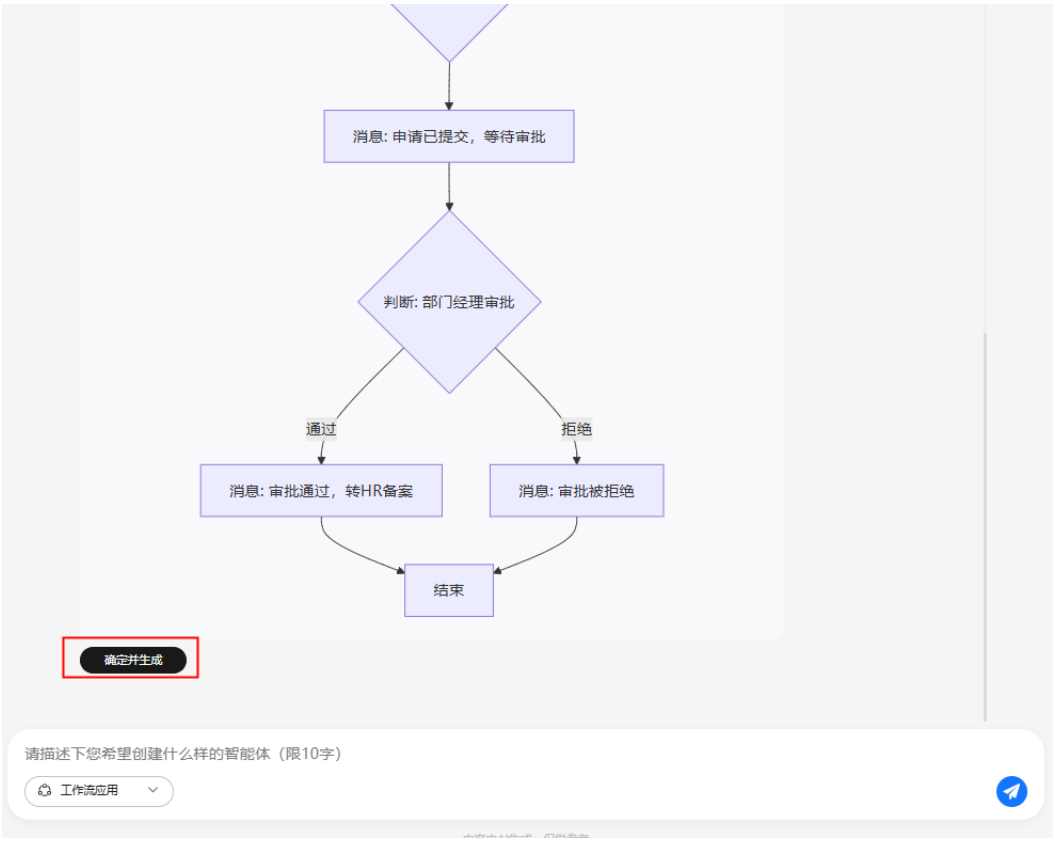


步骤4 在思考结果后单击“确定并生成”，将向系统发送“确认无误”指令或在输入框输入“确认无误”。

说明

- 若思考结果不满足要求，您可在输入框中重新输入指令。
- 在工作流编排页面您可以根据需要扩展工作流应用的能力。

图 8-9 思考结果



步骤5 Versatile将根据工作流设计图生成工作流应用，并跳转至工作流编排页面。

----结束

全局配置

在工作流编排界面，在画布右上角有全局配置入口，用于配置对话型工作流对话体验、默认模型、全局特性开关和定义的配置能力。

说明

工作流发布后的版本作为独立的资源，不支持修改全局配置。

表 8-8 全局配置参数说明

参数	功能
默认模型	作为开场白、推荐问题的智能生成模型来源，新增节点默认使用该模型配置。 ● 单击模型配置下拉框，可配置默认模型，为新拖入节点提供默认的模型选项。 ● 勾选模型配置下的复选框 可将全局模型一键修改，提升模型配置效率。 **说明** 模型的标签展示顺序从左到右依次是用户自定义标签、**接入模型**时的“选择标签”、“模型类型”。 ● 接入模型时的“选择标签”： ● 联网：表示该大模型具备联网搜索能力。 ● 思考：表示该大模型具备思维推理能力。 ● 工具：表示该大模型支持应用调用外部工具，例如，MCP服务、插件、知识库等。 ● default-import：表示该大模型是系统默认模型。 ● 免费：表示该平台预置大模型可免费使用。 ● 体验：表示该平台预置大模型可以体验，会话轮数最大为20次。 ● “模型类型”包含： ● 文本：表示该大模型是文本对话类型。 ● 视觉：表示该大模型是图像理解类型。 ● 嵌入：表示该大模型是文本向量化类型。 ● 排序：表示该大模型是文本排序类型。 ● 模型状态： ● 未验证：表示该大模型未检验鉴权信息，不可使用。 ● 成功：表示该大模型鉴权信息校验成功，可以使用。 ● 失败：表示该大模型鉴权信息校验失败，不可使用。

参数	功能
对话体验	该描述将在气泡内作为应用开场白展示给用户。最大支持输入226个字。 ● 支持在对话框中为对话型工作流中配置开场白、推荐问题。 ● 支持智能生成开场白和推荐问题。 – 智能生成开场白：开场白对话框中输入开场白概述，单击右上角 按钮，在“替换开场白”弹窗中单击“确定”，系统将自动生成开场白并替换当前开场白内容。 – 智能生成推荐问题：推荐问题对话框中输入问题概述，单击右上角 按钮，在“替换推荐问题”弹窗中单击“确定”，系统将自动生成推荐问题并替换当前对话框中的内容。 **说明** 仅支持添加3个推荐问题。
记忆变量	记忆变量的节点赋值支持工作流节点的引用，同时支持引用对象模版和JSON导入。当全局变量配置了节点赋值，同时开始节点用户配置了其他输入参数时，可以在试运行界面中对这些参数进行调试。 记忆变量支持以下参数配置： ● 类型：支持配置string、number、boolean、object、inter、array多种类型的参数，其中object类型参数最多支持3层嵌套。 ● 时长：支持两种长度，“永久”和“会话”，如果选择会话：当会话结束后记录的参数值将自动恢复为默认值；如果选择永久：节点赋值变量将长期保存。 ● 描述：（可选）单击 ··· 图标可配置描述参数信息，帮助理解传入参数的含义。 ● 默认值：（可选）单击 ··· 图标您可以设置输入参数的默认值，其中Object类型参数的默认值需输入Json数据。 ● 单击 图标可删除记忆变量。
内容审核配置	支持通过单击右侧的开关按钮“启动”或“关闭”内容审核配置功能。 内容审核配置功能开启时，可通过单击“配置”设置关键词匹配处理输入输出内容，保障大模型内容安全。 ● 过滤：将大模型输出内容字段屏蔽掉后再返回给用户。 ● 替换：将大模型输出的关键词替换为设置的字段。 ● 兜底回复：触发关键词后，将直接返回配置的兜底回复内容。 **注意** ● 审核内容输入时需要用“，”隔开。 ● 内容审核和安全护栏无法同时开启，打开内容审核配置开关后，“安全防护”将自动关闭。

参数	功能
安全护栏	主要用于检测和拦截潜在的有害、敏感或攻击性的内容。具体来说，它能够识别并阻止那些旨在操纵或滥用系统的Prompt攻击，同时也能过滤掉包含有毒、不适当或违法信息的输入和输出，从而保护用户和系统免受不良影响。这一机制对于维护平台的健康环境和保障用户安全至关重要。 **注意** 内容审核和安全护栏无法同时开启，打开当前开关后，“内容审核”将自动关闭。
语音交互	支持语音输入、卡片消息朗读和实时通话，可在调试页面进行。 ● 单用户免费体验额度：语音输入(一句话识别)50次/日、卡片消息朗读(语音合成50次/日)、通话(实时语音)10分钟/日。 ● 支持为智能体指定音色，用于配置智能应用调试对话模型返回结果朗读时候的音色。 **说明** 语音超过60秒，弹窗提示语音输入时长最长为60秒，取消语音输入状态，用户需重新录入。

图 8-10 全局配置

编排工作流

创建工作流后，初始状态下工作流包含**开始**、**大模型**、**结束**节点。在画布中添加节点，并按照任务执行顺序连接节点，同时按照工作流业务流向配置输入参数和输出参数。

工作流内置了多种基础节点，同时还可以添加“插件”节点来执行特定任务。插件节点使用方法详见**在工作流中使用插件**。

画布界面操作详见**画布操作说明**。

步骤1 在工作流面板中单击“添加节点”，选择目标节点。

步骤2 将各个节点相连接，连接时需注意业务流向。

步骤3 配置节点的输入参数和输出参数。

各个节点的输入输出参数配置请参考**基础节点**、**通用节点**、**逻辑节点**、**工具节点**、**消息管理节点**、**数据&知识节点**。

如果此工作流用于多智能体应用的意图识别，则开始节点必须配置如**图8-12**所示参数、结束节点必须配置如**图8-13**所示参数。

图 8-11 编排工作流

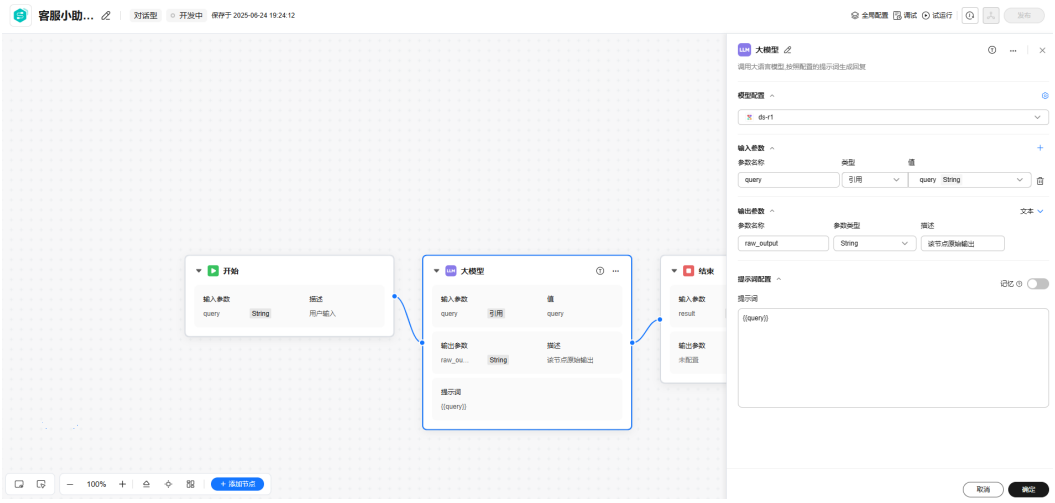


图 8-12 开始节点配置（工作流用于多智能体应用意图识别时配置，参数与截图保持一致）

图 8-13 结束节点配置（工作流用于多智能体应用意图识别时配置，参数与截图保持一致）

----结束

画布操作说明

工作流构建过程中，画布中可以执行的操作如图8-14所示。

图 8-14 画布界面操作

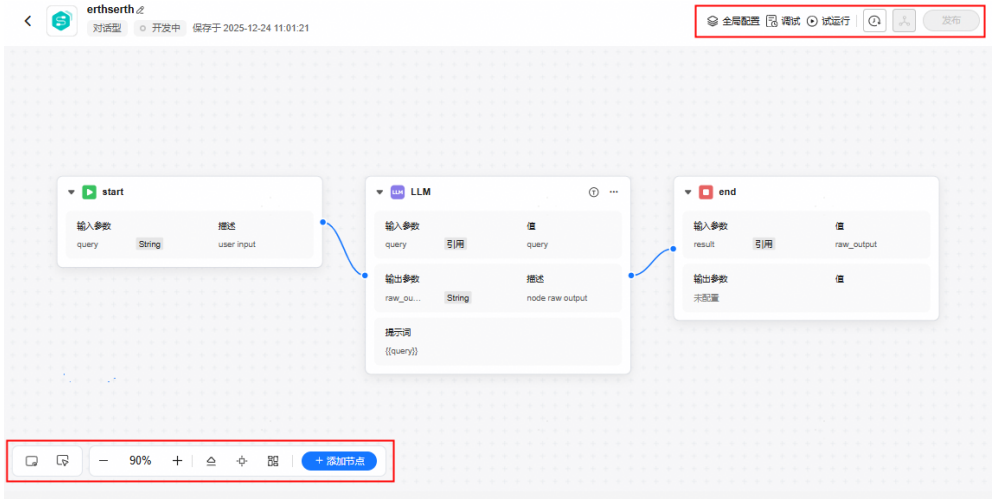


表 8-9 画布操作说明

操作	说明
删除节点	不支持删除起始节点和结束节点。 鼠标光标移至节点上，单击“ ··· > 删除”，即可删除节点。
复制节点	鼠标光标移至节点上，单击“ ··· > 复制”，画板中即可出现复制的节点。 **说明** 不支持跨画布复制节点。
重命名节点	鼠标光标移至节点上，单击“ ··· > 重命名”，在重命名窗口中输入节点名称。 **说明** 同一个工作流画板中节点名称不可重复。
显示缩略图	单击画板左下角 ，在画板中显示或取消工作流缩放图。
查看画布节点	单击画板左下角 ，查看画布节点，支持在查看画布节点界面输入节点名称搜索节点。
缩放工作流	单击画板左下角 − 100% + 两侧符号，调整工作流在画布中显示大小，步长10。
全局节点折叠	单击画板左下角 ，折叠全局节点，折叠后画布内仅显示节点名称和业务流向。
全局节点打开	单击画板左下角 ，打开全局节点，打开后画布内显示节点配置，包括输入参数、输出参数、描述等。
居中	单击画板左下角 ，一键将工作流调整至画布中间位置。
布局优化	单击画板左下角 ，优化工作流编排样式。
全局配置	单击画板右上角 ，可配置全局参数，包括模型配置、对话体验、记忆变量、节点赋值、内容审核配置、语音交互等。
调试	单击画板右上角 ，进入“调试”界面，支持查看“运行结果”和“调用详情”，详细操作可参考**调试工作流**。
运行	单击画板右上角 ，进入“试运行”界面，并在输入框中输入问题测试应用功能。
发布历史	单击画板右上角 ，查看应用发布历史，支持还原版本和删除发布版本，详细操作可参考**查看发布历史**。
发布管理	单击画板右上角 ，支持查看已发布工作流的发布详情。

操作	说明
发布	单击“发布”，完善版本名称，可将工作流发布成网页和API调用模式，详细操作可参考**发布工作流**。

8.4.3 调试工作流

开发者可以在工作流创建完成后，直接与工作流进行交互，实时观察其执行过程和响应效果，并根据需要对配置进行优化和调整。平台提供的全链路调试功能，允许开发者查看每条用户请求从输入到响应的完整流程，包括意图识别、知识检索等详细信息，从而能够高效定位问题并快速调整配置。

Versatile支持对整个工作流进行调试，也支持对工作流的单个节点进行调试。

前提条件

已**创建工作流**。

约束与限制

试运行工作流时，端到端运行时间最长可执行10分钟。

试运行工作流（必选）

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 单击左侧导航栏“开发中心 > 应用管理 > 工作流应用”。

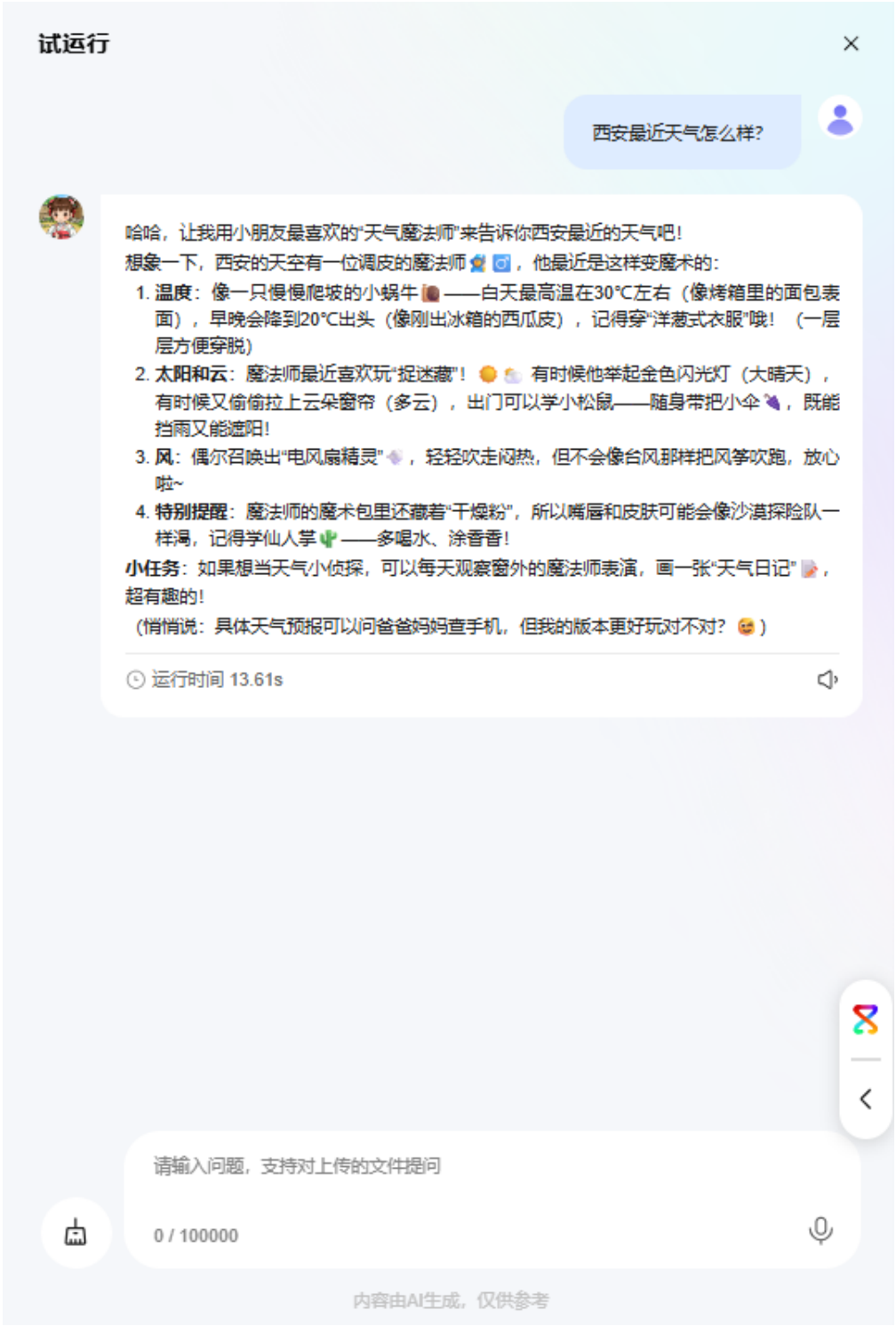
步骤3 在工作流界面中找到目标工作流，并单击进入工作流编排页面。

步骤4 试运行工作流。

在工作流的开始节点中，默认包含一个输入参数query ，用于表示用户在本轮对话中输入的原始内容。您也可以根据需要添加其他参数，以供下游节点使用。因此，试运行工作流分为两种场景：一种是仅使用开始节点的默认输入参数query ，另一种是开始节点用户添加了其他参数。

- **开始节点默认输入参数query时**，工作流编排完成后，单击右上角“试运行”，在对话框中输入问题，等待返回试运行结果。

图 8-15 试运行工作流



说明

- 试运行界面支持文本输入、文件输入和语音输入：
  - 文本输入：在对话输入框输入对话后按Enter键或单击  ，查看应用响应结果。
  - 语音输入：全局配置中开启语音交互功能时，用户可以通过语音进行输入。该功能支持多种语言（如中文、英文等），并提供语音识别、错误纠正和实时反馈等功能。
    - 首次使用语音输入须开通系统麦克风、扬声器权限，可在权限申请弹窗一键开通。
    - 语音超过60秒，弹窗提示语音输入时长最长为60秒，取消语音输入状态，用户需重新录入。
- 调试结果支持朗读功能，单击  ，应用将按照设置的音色将文字转换成语音播放。
- 单击试运行页面左下角  ，一键清除试运行界面内容。
- 文件输入：请参考**开始节点**配配置参数，可增加“文件”或“文件数组”类型，并在试运行界面中上传文件。

- **开始节点用户添加了其他参数时**，工作流编排完成后，单击右上角“试运行”，在对话框中输入问题，单击“开始运行”，等待返回试运行结果。

图 8-16 试运行工作流

步骤5 在试运行过程中，可以单击右上角 调试 查看调试结果，包括运行结果与调用详情，如图8-17所示。

如果试运行失败，常见报错与解决方案请详见应用开发常见问题。

图 8-17 调试结果示例

说明

- 单击调用详情页面中的按钮，查看列表调用链。
- 单击调用详情页面中的按钮，查看火焰图调用链。
- 支持在**查看应用调用链信息**页面中，查看该调用链的详细信息，具体操作请参见**使用过滤器筛选信息**。

----结束

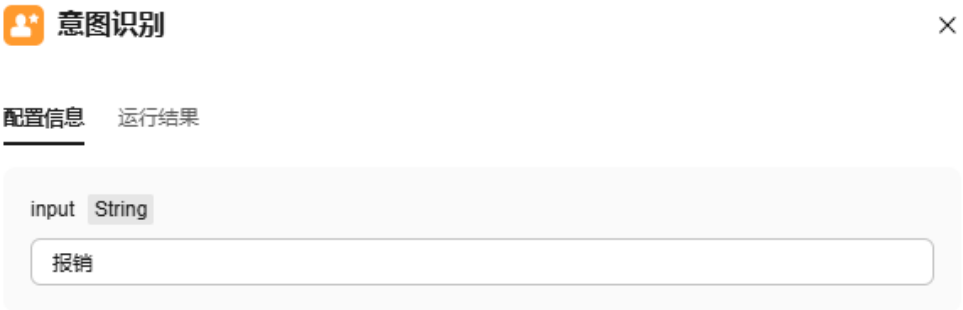
调试单节点

以调试“意图识别”节点为例：

步骤1 在工作流编排页面，单击意图识别节点的，进入单节点调试页面。

步骤2 输入参数内容，单击“开始运行”。

图 8-18 编写输入参数内容



步骤3 在“运行结果”页面，查看当前节点的运行结果。

运行成功，节点处也将显示“运行成功”字样。

运行失败，需要根据提示调整节点参数。

图 8-19 单节点调试运行成功示例

----结束

8.4.4 发布工作流

工作流试运行成功后，可对其进行发布，便于后续使用，如在单智能体应用中添加工作流，**添加工作流**。

前提条件

工作流已调试完成，具体请参见**调试工作流**。

发布工作流

步骤1 在工作流编排页面，单击右上角“发布”，输入版本名称与描述，单击“发布”。

图 8-20 发布工作流

步骤2 发布完成后跳转至“发布管理”页面，也可通过左侧导航栏中选择“开发中心 > 应用管理 > 单智能体应用”，单击目标应用，进入应用主页面，单击右上角“发布管理”按钮，可进入发布管理页面。

1. 在**发布管理**页面，**网页**发布渠道的**操作**列单击“发布”按钮后将出现“立即访问”、“复制链接”、“重新生成”文字按钮。这里可以通过两种方式访问工作流应用的网页应用链接，同时支持重新生成发布链接。

   图 8-21 发布工作流

   - 立即访问：单击当前页面的“立即访问”按钮，可立即进入网页版应用。

     图 8-22 立即访问网页应用

   - 复制链接：单击当前页面的“复制链接”按钮，可直接复制网页URL嵌入或分享到其他应用场景。

图 8-23 复制工作流应用访问地址

- 重新生成：重新生成工作流应用发布链接。

说明

- 已发布的应用支持在Agent应用编辑页面选择“更新发布”按钮重新发布应用。也可在“是否更新版本”列单击“更新版本”，并在确认界面单击“更新版本”，该发布渠道将更新为最新版本配置。
- “资源配置”列，单击 可进行资源配置设置，支持配置“每日调用限额（次）”、“分享范围”。
  - “每日调用限额（次）”：支持用户自定义修改，默认100次。支持输入-1或0~10000，其中-1表示调用次数无限制。
  - “分享范围”：可选择当前租户可用和所有租户可用。

2. 在**发布管理**页面，**云商店**发布渠道的**操作**列单击“发布”按钮后将出现“立即访问”、“复制链接”、“重新生成”文字按钮。

图 8-24 发布工作流应用

- 立即访问：单击当前页面的“立即访问”按钮，可立即跳转至我的云商店界面。

图 8-25 立即访问云商店界面

说明

华为云商店具体操作详见**发布API类产品**。

- 复制链接：单击当前页面的“复制链接”按钮，可直接复制网页URL嵌入或分享到其他应用场景。

图 8-26 复制智能体访问地址

- 重新生成：单击当前页面的“重新生成”按钮，可重新生成该应用。

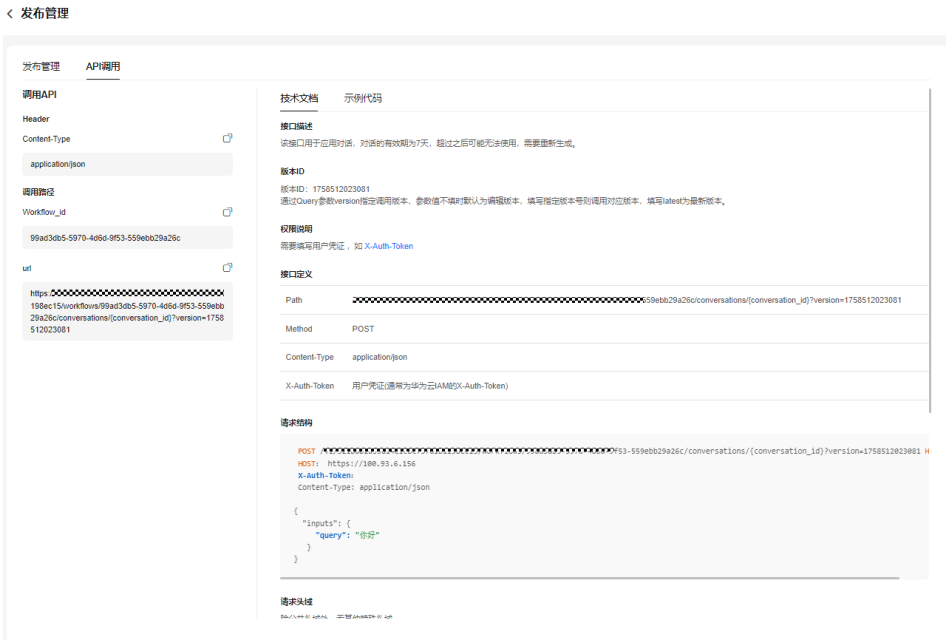
图 8-27 重新生成 Agent

说明

- 已发布的应用支持在Agent应用编辑页面选择“更新发布”按钮重新发布应用。也可在“是否更新版本”列单击“更新版本”，并在确认界面单击“更新版本”，该发布渠道将更新为最新版本配置。
- “资源配置”列，单击 可进行资源配置设置，支持配置“每日调用限额（次）”、“分享范围”。
  - “每日调用限额（次）”：支持用户自定义修改，默认100次。支持输入-1或0~10000，其中-1表示调用次数无限制。
  - “分享范围”：可选择当前租户可用和所有租户可用。

3. 在**发布管理**页面，选择“API调用”功能选项卡即可查看API调用接口信息，详细操作可参考**通过API调用工作流**。

图 8-28 获取 API 调用信息

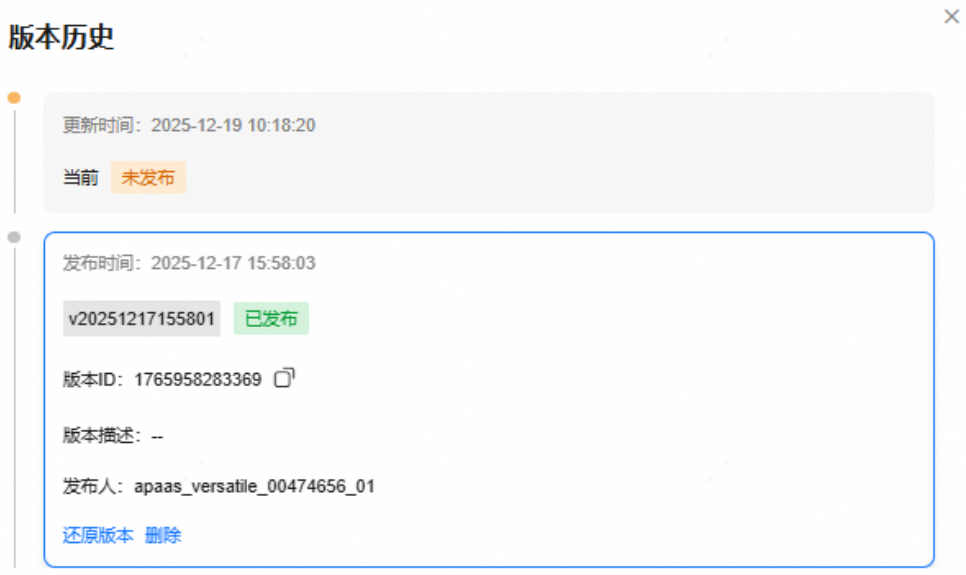


----结束

查看发布历史

单击右上角  ，可查看当前工作流发布历史记录。

图 8-29 发布历史



说明

发布历史支持还原版本和删除操作：

- 还原版本：单击“还原版本”，并在弹窗中单击“确定”，将还原当前工作流至配置前状态，工作流配置信息将不再保留，请谨慎操作。
- 删除：单击“删除”，并在弹窗中单击“确定”，将删除当前工作流发布历史。

8.5 使用工作流

8.5.1 通过 API 调用工作流

工作流发布成功后，可以使用API调用该工作流。

前提条件

须确保工作流已发布，详情可参考**发布工作流**。

获取工作流 ID 和调用路径

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在“开发中心 > 应用管理 > 工作流应用”页面，选择目标工作流。

步骤3 单击“ ··· > 复制ID”，可获取当前工作流ID。请记录保存，用于填写调用Agent应用接口的agent_id字段。

步骤4 单击 ··· 选择“调用路径”。

图 8-30 获取工作流调用路径

步骤5 在弹出的“调用路径”页面，单击“复制路径”即可获取调用路径，如**图7-83**所示。

其中，“f2d9736e-9f7a-41b5-b3fb-5802b347b390”是随机生成的字符串，在使用时可以替换为其他的字符串。字符串长度为1~64个字符，支持英文字母、数字、中划线、下划线。在智能体应用被API调用时，为“conversation_id”的值。

图 8-31 获取工作流调用路径-2

----结束

使用 API 调用单智能体应用

工作流发布为API服务之后，可通过API调用工作流应用。详情请参考**调用工作流应用**。

8.5.2 在单智能体应用中使用工作流

工作流功能是实现单智能体应用业务逻辑的核心部分。它定义了应用的输入和输出数据结构、数据接收与处理规则，以及决策流程。通过工作流，单智能体应用能够支持添加工作流技能，允许用户通过画布编排的方式，组合使用插件、大模型等不同节点，从而实现复杂且稳定的业务流程编排。

- **简单业务**：应用中至少有一个试运行通过的工作流，作为应用的业务处理流程。
- **复杂任务**：如果你的业务逻辑复杂由多个子任务组成，您可以将其拆分为多个工作流，每个工作流负责完成其中的一个任务，再将这些工作流组合到一个工作流中。

前提条件

须确保工作流已发布，详情可参考**发布工作流**。

在单智能体应用中配置工作流

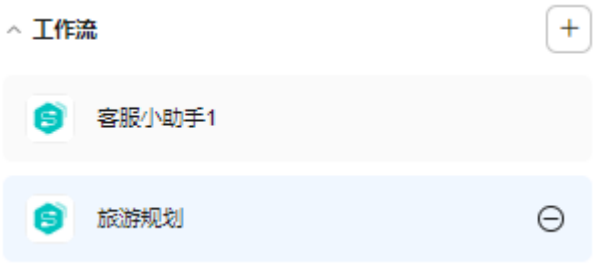
步骤1 在“技能 > 工作流”模块，单击 + 。

步骤2 在“添加工作流”窗口，单击 + 进行添加，再单击“确定”。

图 8-32 添加工作流

步骤3 添加工作流后，可在“技能 > 工作流”中查看当前已添加的工作流。

图 8-33 已添加工作流



说明

已添加的工作流单击 支持配置工作流参数。当参数“可见性”开关关闭时，智能体将无法查看和修改该参数，且在智能体运行时不会动态提取该参数。对于一些不需要智能体动态提取的固定参数，例如密钥等，可以关闭其可见性。

----结束

相关文档

单智能体应用中使用工作流示例，请参考**添加工作流**。

8.5.3 在多智能体应用中使用工作流

可通过工作流功能来实现多智能体应用的业务逻辑部分。工作流决定了应用的输入和输出的数据结构、接收和处理数据的规则以及决策流程，是智能体应用的核心部分。

多智能体应用通过添加工作流技能，对用户的会话进行意图分析，路由到不同的子工作流执行编排好的任务。

前提条件

须确保工作流已发布，详情可参考**发布工作流**。

在多智能体应用中配置工作流

步骤1 在“多Agent控制器”配置页面，支持添加意图识别、起始工作流、子工作流、默认工作流和结束工作流，可根据业务需求选择对应的工作流。

步骤2 添加工作流后，单击下拉框可选择已发布的工作流版本应用。

图 8-34 添加工作流

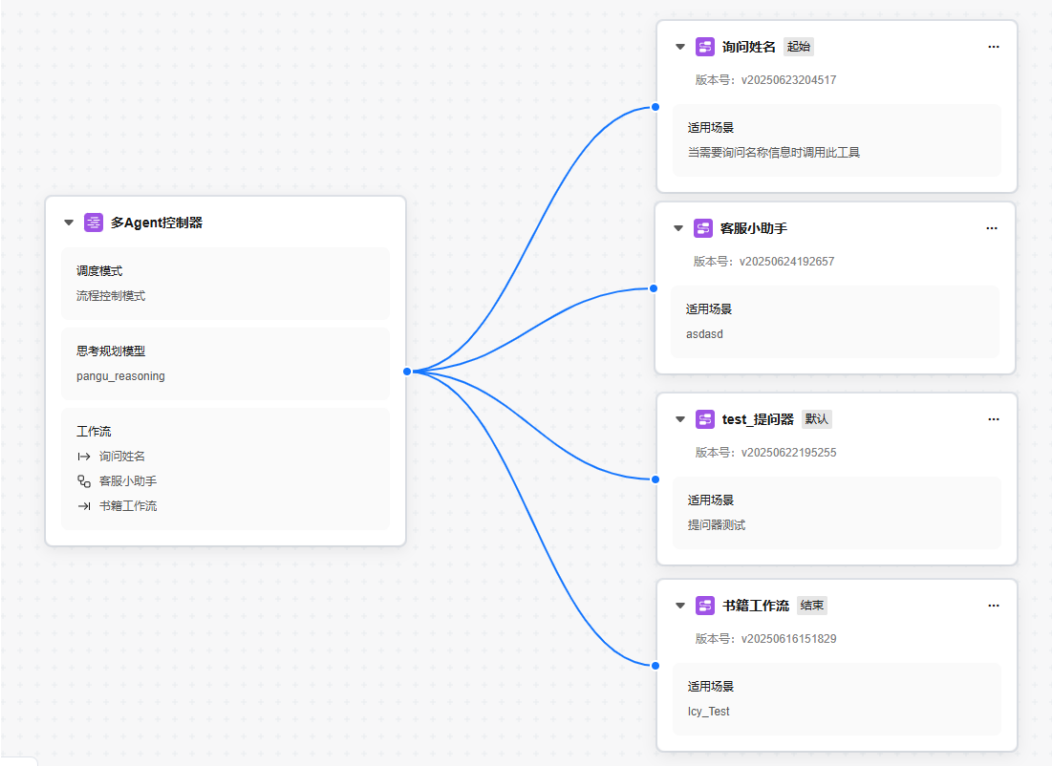
步骤3 添加工作流保存后，可在画布中查看当前已添加的工作流。

图 8-35 已添加工作流

步骤4 添加起始、默认、结束工作流，均通过单击下拉框选择后“保存”。

图 8-36 添加起始、默认、结束工作流

图 8-37 编排后包含工作流的多智能体应用



----结束

相关文档

多智能体应用中使用工作流示例，请参考**创建多智能体应用**。

8.6 管理工作流

Versatile支持对工作流执行复制、获取工作流ID、调用路径、删除、导入、导出等操作。

前提条件

已**购买Versatile智能体平台**。

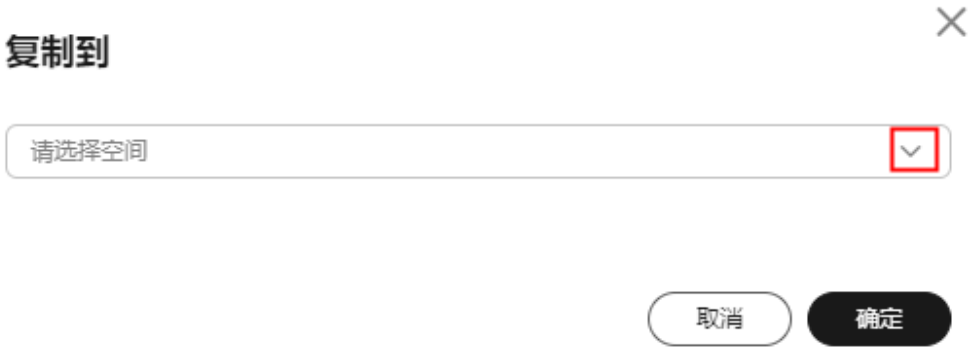
复制工作流

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 进入“开发中心 > 应用管理 > 工作流应用”页面。

步骤3 选择目标工作流，单击“ *** > 复制”，在“复制到”下拉框中选择已创建的目标空间，可复制当前工作流到目标空间。

图 8-38 复制应用



说明

在复制到目标空间时，应用的配置参数、大模型、节点等数据将一并复制，且复制后的应用需要单独发布。

----结束

获取工作流 ID

工作流应用除了具有页面操作的能力之外，还具有Chat API调用能力，对于AppID获取就十分必要。该ID为调用Agent应用接口的agent_id字段。

```
POST /v1/{project_id}/agents/{agent_id}/conversations/{conversation_id}
```

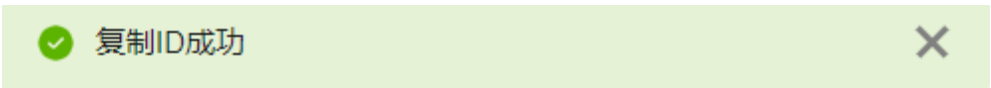
步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 进入“开发中心 > 应用管理 > 工作流应用”页面。

步骤3 选择目标工作流，单击“ ··· > 复制ID”，可获取当前工作流ID。

步骤4 弹出复制成功对话框，用于填写调用Agent应用接口的agent_id字段。

图 8-39 复制 ID



----结束

调用路径

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 进入“开发中心 > 应用管理 > 工作流应用”页面。

步骤3 选择目标工作流，单击“ ··· > 调用路径”，调用路径为工作流的API接口。详细API调用过程请参见**通过API调用工作流**。

图 8-40 获取调用路径

----结束

删除工作流

须知

- 如果工作流版本已被引用，删除后引用将被自动取消，可能会导致工作流或智能体无法运行，且该操作不可撤回。
- 若该工作流应用已经共享，则用户无法直接删除，必须先手动完成“取消共享”操作后 ，才能在工作流应用界面删除应用，详细操作请参见**更多操作**，仅支持共享者取消工作流共享。

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 进入“开发中心 > 应用管理 > 工作流应用”页面。

步骤3 选择目标工作流，单击“ ··· > 删除”

步骤4 在弹出的对话框中单击“确定”。

- 如果工作流未被引用在弹窗中单击“确定”即可。

图 8-41 工作流未被引用

- 如果工作流被引用则删除后引用将自动取消，可能会导致工作流或应用无法运行，且该操作不可撤回。

图 8-42 工作流被引用

----结束

导入工作流

平台支持导入工作流。导入工作流时，将同步导入工作流关联的插件等配置。

步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 进入“开发中心 > 应用管理 > 工作流应用”页面。

步骤3 导入工作流。

1. 单击页面左上角“导入”。
2. 在“导入”页面，单击“选择文件”选择需要导入的jsonl文件。
3. 选择导入文件后，选择解析内容。
   平台将自动解析jsonl文件。如果解析的文件已存在，勾选该文件将自动覆盖平台现有文件。
4. 单击“导入”，导入成功的工作流将在“开发中心 > 应用管理 > 工作流应用”页面中展示。

说明

仅支持上传jsonl格式文件，工作流文件的最大导入大小为128MB。

----结束

导出工作流

平台支持导出工作流。导出工作流时，将同步导出工作流关联的插件等配置。

步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 进入“开发中心 > 应用管理 > 工作流应用”页面。

步骤3 选择目标工作流，导出工作流。

1. 单击页面左上角“导出”。
2. 在“导出工作流”页面选择工作流，单击“导出”。工作流将以一个JSONL格式的文件下载至本地。

----结束

发布管理

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 进入“开发中心 > 应用管理 > 工作流应用”页面。

步骤3 在工作流开发空间，选择待部署的应用，单击应用的右下角 ··· 展开功能列表，选择“发布管理”，跳转至“发布管理”页面。详细操作可参见**发布工作流**节点。

----结束

8.7 基础节点

8.7.1 开始和结束节点

开始节点用于开启触发一个工作流，结束节点用于返回一个工作流的最终结果。

开始节点

开始节点是一个工作流的起始节点，用于设定启动工作流所需的输入参数。开始节点只有输入参数，没有输出等其他参数。开始节点中默认有一个输入参数query，表示用户在本轮对话中输入的原始内容。您也可以按需添加其他参数，用于下游节点的输入；同时开始节点也支持选择对象模版导入。

开始节点参数配置说明如下：

- **数据类型：**开始节点支持配置String、Number、Boolean、Object、File、Array多种类型的输入参数，其中Object类型参数最多支持5层嵌套。

  说明

  文件作为输入参数时，用户可增加“文件”或“文件数组”类型，并在试运行界面中上传文件。
  - 文件类型：“参数类型”选择“File”，并选择对应的文件类型，一个参数仅支持上传一个文件。
  - 文件数组类型：“参数类型”选择“Array<File>”，并选择对应的文件类型，一个参数最多支持上传10个文件。
  - 支持上传Default、Doc、Txt、Excel、PPT、Image、Audio、Video等格式的文件。
- **参数描述：**参数的描述信息，帮助模型理解传入参数的含义。将工作流绑定到智能体中使用时，模型会自动分析用户的Query，将Query中表达的信息填入对应的参数中。
- **是否必选：** 参数是否必选。如果未指定必选参数，工作流无法执行。当将工作流绑定到智能体中使用时，如果用户查询中缺少必选参数，则不会触发该工作流。
- **参数默认值：**您可以设置输入参数的默认值，默认值将会回显到试运行界面的输入框中，其中Object类型参数的默认值需输入Json数据，后台会校验Json数据和参数定义的一致性，然后再进行赋值。示例如下图所示。

图 8-43 开始节点 Object 类型参数嵌套

图 8-44 开始节点 Object 类型参数

结束节点

结束节点是工作流的最终节点，用于返回工作流运行后的结果。

图 8-45 结束节点

结束节点的参数配置说明如下：

输入参数：输入参数支持引用和输入两种类型，输入参数需要在指定回复的文本框中以{{variable_name}}的形式进行插入才能返回。

输出参数：输出参数将以变量形式返回，支持引用和输入两种类型。工作流运行结束后会以JSON格式返回所有输出参数，适用于子工作流的场景。如果工作流直接绑定了

智能体，对话中触发了工作流时，大模型会自动总结JSON格式的内容，并以自然语言回复用户。如果工作流设置全局配置中的记忆变量，可在结束节点引用记忆变量。输出参数不能在指定回复中引用。

指定回复：您可以在文本框中编辑指定的回复内容，支持在文本中以{{variable_name}}的形式插入输入参数返回或直接返回输入参数。工作流的最终运行结果将按照指定回复中的内容返回，可在“结构化信息”中使用{{_NODE_OUTPUT}}引用。指定回复中不能插入输出参数。

结构化信息：功能开启时，可使用{{_NODE_OUTPUT}}引用“指定回复”中的信息实现结构化输出。

8.8 通用节点

8.8.1 大模型

大模型节点提供了使用大模型的能力，可在节点中配置已部署的模型，用户可以通过编写Prompt、设置参数让模型处理相应任务。

前提条件

已完成工作流搭建，如果未搭建工作流，请参考**搭建工作流**。

约束与限制

如果在大模型节点后并行使用多个大模型，应将首个大模型节点配置为非流式输出。

大模型节点说明

通过该节点，用户能够灵活地编写提示词（Prompt）并精细设置相关参数，从而实现对大模型的高效调用。该功能支持多种类型的大模型服务，能够处理包括文本生成、对话交互、内容理解等多种任务场景，为用户提供强大而灵活的AI能力支撑。

配置大模型节点

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏选择“开发中心 > 应用管理 > 工作流应用”，单击您创建的工作流。

步骤3 单击“添加节点”并选择“大模型”节点。

步骤4 通过单击该节点打开节点配置页面。

步骤5 参照**表8-10**，完成大模型节点的配置。

说明

- 单击图标，可修改大模型名称，修改完成后单击名称旁边的进行保存。
- 单击图标，可重命名大模型名称，复制一个大模型或删除大模型。
- 单击图标，可对大模型节点进行测试。

表 8-10 大模型节点配置说明

配置类型	参数名称	参数说明	配置示例
模型配置	模型配置	选择模型接入模块已配置的大语言模型。	DeepSeek-V3-64K
	深度思考	显示该参数有以下两个场景： ● **平台推荐**：当选择的模型服务为思考模型且支持关闭深度思考时，才显示此参数，例如平台推荐的Qwen3-32B、DeepSeek-V3.2。 ● **用户自主接入的模型服务**：当选择的模型服务为思考模型且在新建模型服务开启了“是否支持关闭思维链输出”时，才显示此参数。 该参数支持以下操作： ● 当此功能**开启**时，大模型将首先进行深入的思考和推理，通过逐步拆解问题、梳理逻辑，生成一段详细的思维链内容，并在调试界面展示。这一过程有助于提升最终输出答案的准确性和可靠性，确保用户获得更加精准的信息。 ● 当此功能**关闭**时，智能体将直接生成最终答案，不再经过额外的思维链推理过程。这将加快响应速度，适用于需要快速获取答案的场景。 **注意** 在模型使用过程中，“深度思考”开关生效的情况如下： ● 如果模型支持思维链输出能力，并且也支持关闭该能力，则开启、关闭均生效。 ● 如果模型支持思维链输出能力，但不支持关闭该能力，则开启生效、关闭不生效。 ● 如果模型不支持思维链输出能力，则开启、关闭均不生效。	默认打开
	温度	当单击图标时，可进行该参数设置。 用于控制生成结果的随机性。调高温度，会使得模型的输出更具多样性和创新性；降低温度，会使输出内容更加遵循指令要求，但同时也会减少模型输出的多样性。	0.5

配置类型	参数名称	参数说明	配置示例
	核采样	当单击图标时，可进行该参数设置。 模型在输出时会从概率最高的词汇开始选择，直到这些词汇的总概率累积达到核采样值，核采样值可以限制模型选择这些高概率的词汇，从而控制输出内容的多样性。建议不要与温度同时调整。	0.5
	历史对话轮数	当单击图标时，可进行该参数设置。 设置带入模型上下文的对话历史轮数。轮数越多，多轮对话的相关性越高，但消耗的Token也越多。	3
	最大回复长度	当单击图标时，可进行该参数设置。 控制模型输出的Tokens长度上限。通常100Tokens约等于150个中文汉字。	131072
	重复语句惩罚	当单击图标时，可进行该参数设置。 • 当该值为正时，会阻止模型频繁使用相同的词汇和短语，从而增加输出内容的多样性。 • 当该值为0时，表示不施加任何惩罚，模型完全按照原始概率分布生成文本，可能导致重复问题，例如在重复惩罚中取0相当于无惩罚，输出可能缺乏多样性。 • 当该值为负时，表示鼓励已出现的token再次被选择，增加重复性。	2

配置类型	参数名称	参数说明	配置示例
参数配置	输入参数	配置大模型处理需要的输入参数值， 这些值会动态添加到提示词中，默认设置的输入参数名为query。 当单击+图标时，可新增输入参数。 当单击图标时，可删除输入参数。 ● 参数名称：只允许输入字母、数字、下划线，且不能以数字开头。 ● 类型、值：支持“引用”和“输入”两种类型。 – 引用：支持用户选择工作流中已包含的前置节点的输出参数，如果配置了全局变量中的记忆变量，也支持引用记忆变量。 – 输入：将用户自定义的内容传递给大模型，设置为输入模式后， 无论前置节点产生什么输出内容，都不会传递给大模型。 **说明** 当参数类型为引用时，设置参数值时，您可以在下拉框顶部的搜索栏中输入关键词，快速定位所需参数。	query

配置类型	参数名称	参数说明	配置示例
	输出参数	该参数用于解析大模型节点的输出，并提供给后续节点的输出参数引用。 ● 参数名称：参数的名称长度必须大于等于1个字符，并且字符只允许为下面三种类型： – 字母（A-Z或a-z） – 数字（0-9） – 特殊字符：_ **说明** 用户自定义输出参数名称不允许与内置输出参数rawOutput同名。大模型节点有一个内置输出参数rawOutput，代表该节点未经解析的原始输出，与大模型节点相连的后续节点可以直接引用该输出。 ● 参数类型：输出参数的类型，可选String、Integer、Number、Boolean、Object、Array<String>、Array<Number>、Array<Integer>、Array<Boolean>、Array<Object>。 ● 描述：对于该输出参数的描述。 ● 流式输出：模型调用方式开关，支持开启或关闭模型流式输出效果。 ● 输出格式：支持输出的格式包括文本、Markdown、JSON。 – 文本：大模型原始内容输出，仅支持一个参数，默认为raw_output， 支持修改名称。 – Markdown：期望模型输出markdown格式内容时选择。仅支持一个参数，默认为raw_output， 支持修改名称。 – JSON：要求模型按JSON格式响应；支持添加多个参数。	raw_output

配置类型	参数名称	参数说明	配置示例
		说明 – 流式输出开启时，支持输出格式选择文本或Markdown。 – 流式输出关闭时，支持输出格式文本、Markdown或JSON。	
	异常处理	支持对节点的异常（如超时、调用失败等情况）进行处理，包括超时时间、重试次数、异常处理方式。 “超时时间”：支持用户配置超时时间，取值范围0.1~900，默认900s。 “重试次数”：支持配置重试次数（不重试、重试1次、重试2次、重试3次），系统默认不重试。 “异常处理方式”：配置异常处理方式。 ● 中断流程：节点发生异常后，直接中断流程，不再运行后续节点。 ● 返回设定内容：节点发生异常后，工作运行不会中断，用户可自定义设置需要返回的输出字段内容，必须是输出参数中已定义的字段，且格式为合法的JSON格式。 ● 执行异常流程：节点发生异常后，工作流不会中断，而是会执行异常处理流程。用户可以在该运行异常的节点前新增节点，并为新增的异常分支配置相应的处理流程。 **说明** ● 当流式输出和异常处理功能开启时，异常处理参数默认为“不重试”和“中断流程”。 ● 当流式输出功能关闭，输出格式为文本或Markdown时，异常处理方式仅支持“中断流程”。 ● 当流式输出功能关闭，输出格式为JSON时，三种异常处理方式均支持。	“超时时间”：900。 “重试次数”：不重试。 “异常处理方式”：中断流程。

配置类型	参数名称	参数说明	配置示例
提示词配置	系统提示词	配置输入给大模型的提示词，系统级提示词，用于指导模型按要求进行回复。支持使用{{variable}}格式引用当前节点输入参数中已定义好的参数。最终替换后的内容会传递给模型。 • 当单击“保存到模板”，填写“模板名称”、选择“行业”和“标签”后，可将提示词创建成模板并保存到我的提示词。 • 当单击  图标时，可对系统提示词进行智能优化。 • 当单击  图标时，系统会弹出“提示词广场”窗口，可在“预制提示词”或“我的提示词”页签中进行选择。	作为一位畅销小说作家，你擅长运用华丽且流畅的语言描绘场景和人物，精于编织情节，使故事层次丰富、悬念迭起。现在，请根据以下输入的小说标题"{{title}}"，构思并概述一段该小说的开场章节（500字左右），展现上述两种创作特点，并在开篇即设置引人入胜的悬念。
	用户提示词	配置输入给大模型的提示词，用户级提示器，作为当前用户问题的输入。配置提示词时，支持使用{{variable}}格式引用当前节点输入参数中已定义好的参数。最终替换后的内容会传递给模型。 当单击  图标时，系统会弹出“提示词广场”窗口，可在“我的提示词”页签中进行选择。	{{query}}
	短期记忆	支持通过单击右侧的开关按钮“启动”或“关闭”短期记忆功能，该功能默认关闭。 用于控制大模型是否读取多轮对话的历史交互内容，开启时可确保多轮对话连贯性。	关闭
安全	安全护栏	主要用于检测和拦截潜在的有害、敏感或攻击性的内容。具体来说，它能够识别并阻止那些旨在操纵或滥用系统的Prompt攻击，同时也能过滤掉包含有毒、不适当或违法信息的输入和输出，从而保护用户和系统免受不良影响。这一机制对于维护平台的健康环境和保障用户安全至关重要。	关闭

图 8-46 大模型节点配置示例

步骤6 完成节点配置后，单击“确定”。

步骤7 连接工作流节点和其他节点。

----结束

8.8.2 工作流

设计工作流节点，以实现工作流的嵌套功能。

前提条件

- 已完成工作流搭建，如果未搭建工作流，请参考**搭建工作流**。
- 如果需要添加共享工作流，请确保已有他人共享的工作流。
- 仅**Versatile企业版**支持使用他人共享的应用。**Versatile基础版（限时免费）**不支持该能力。

节点说明

在一个工作流中，您可以将另一个工作流作为其中一个步骤或节点，实现复杂任务的封装。例如，可以将常用的、标准化的任务处理流程封装为不同的子工作流，并在主工作流的不同分支中调用这些子工作流执行相应的操作。通过工作流嵌套，可以实现复杂任务的模块化拆分和处理，从而使工作流编排逻辑更加灵活、清晰和易于管理。

配置工作流节点

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏选择“开发中心 > 应用管理 > 工作流应用”，单击您创建的工作流。

步骤3 单击“添加节点”并选择“工作流”节点。

步骤4 在工作流应用右侧单击  。

也可在“添加工作流”窗口中，单击“创建工作流”，工作流创建与配置详见**开发工作流应用**。创建成功后，在工作流发布成功界面单击“确定”，可立即将创建的工作流添加至当前工作流中。

也可在“团队共享”中选择查看其它团队共享给当前团队的资源，详细资源共享可参见**使用资产中心的应用资源**。

步骤5 通过单击该节点打开节点配置页面。

说明

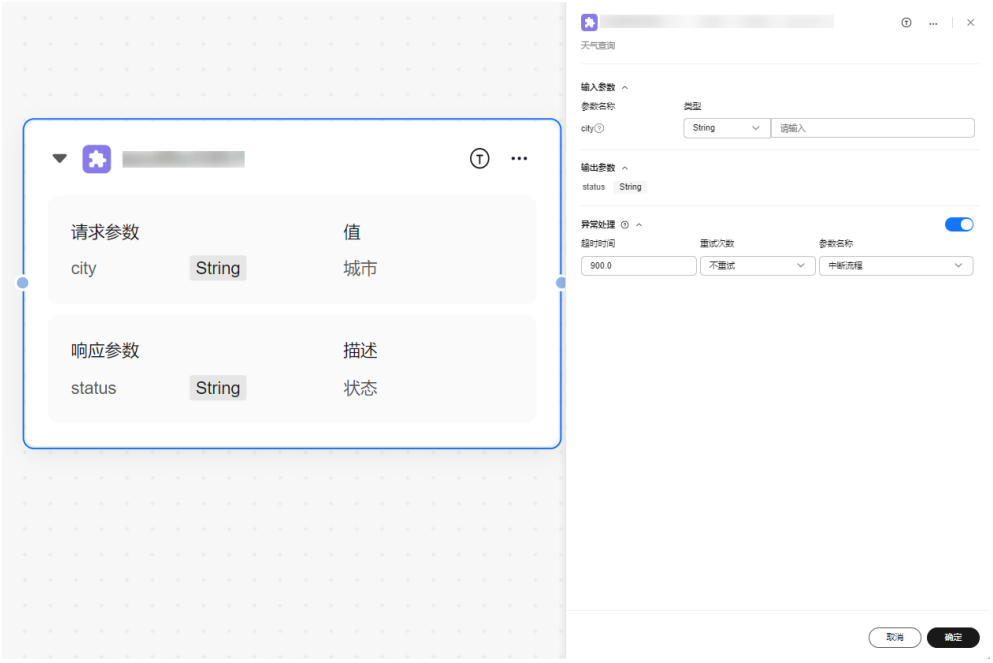
- 单击  图标，可修改工作流名称，修改完成后单击名称旁边的  进行保存。
- 单击  图标，可查看子工作流详情（不支持修改子工作流）、重命名工作流名称、复制一个工作流或删除工作流。

步骤6 参照**表1 工作流节点配置说明**，完成工作流节点的配置。

表 8-11 工作流节点配置说明

配置类型	参数名称	参数说明	配置示例
参数配置	输入参数	● 工作流节点的输入结构取决于子工作流定义的输入结构，不支持自定义设置 ● 在工作流节点中您需要为输入参数指定数据来源，支持设置为固定值或引用上游节点的输出参数。	query
	输出参数	● 工作流节点的输出结构取决于子工作流定义的输出结构，不支持自定义设置。 ● response_content为工作流固定输出参数。	response_content
	异常处理	支持对节点的异常（如超时、调用失败等情况）进行处理，包括超时时间、重试次数、异常处理方式。 “超时时间”：支持用户配置超时时间，取值范围0.1~900，默认900。 “重试次数”：工作流节点不支持重试。 “异常处理方式”：配置异常处理方式。 ● 中断流程：节点发生异常后，直接中断流程，不再运行后续节点。 ● 返回设定内容：节点发生异常后，工作运行不会中断，用户可自定义设置需要返回的输出字段内容，必须是输出参数中已定义的字段，且格式为合法的JSON格式。 ● 执行异常流程：节点发生异常后，工作流不会中断，而是会执行异常处理流程。用户可以在该运行异常的节点前新增节点，并为新增的异常分支配置相应的处理流程。	“超时时间”：900。 “异常处理方式”：中断流程。

图 8-47 工作流节点配置示例-工作流节点（输入和输出对应子工作流的输入和输出）



步骤7 节点配置完成后，单击“确定”。

----结束

8.8.3 Agent

Agent节点提供了使用大模型的能力以及大模型工具调用的能力。

前提条件

已完成工作流搭建，如果未搭建工作流，请参考**搭建工作流**。

Agent 节点说明

可在节点中配置已部署的模型，用户可以通过编写提示词、绑定插件让模型处理相应任务。

配置 Agent 节点

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏选择“开发中心 > 应用管理 > 工作流应用”，单击您创建的工作流。

步骤3 单击“添加节点”并选择“Agent”节点。

步骤4 通过单击该节点打开节点配置页面。

步骤5 参照**表8-12**，完成变量输入节点的配置。

说明

- 单击图标，可修改Agent名称，修改完成后单击名称旁边的进行保存。
- 单击图标，可重命名Agent节点名称，复制一个Agent节点或删除Agent节点。
- 单击图标，可对Agent节点进行测试。

表 8-12 Agent 节点配置说明

配置类型	参数名称	参数说明	配置示例
模型配置	模型配置	选择执行此节点的模型，支持设置模型在此节点中的生成多样性等参数配置，使模型效果更符合你的预期。	DeepSeek-V3-64k
	温度	当单击图标时，可进行该参数设置。 用于控制生成结果的随机性。调高温度，会使得模型的输出更具多样性和创新性；降低温度，会使输出内容更加遵循指令要求，但同时也会减少模型输出的多样性。	0.5
	核采样	当单击图标时，可进行该参数设置。 模型在输出时会从概率最高的词汇开始选择，直到这些词汇的总概率累积达到核采样值，核采样值可以限制模型选择这些高概率的词汇，从而控制输出内容的多样性。建议不要与温度同时调整。	0.5
	最大回复长度	当单击图标时，可进行该参数设置。 控制模型输出的Tokens长度上限。通常100Tokens约等于150个中文汉字。	131072

配置类型	参数名称	参数说明	配置示例
参数配置	输入参数	当单击+图标时，可新增输入参数。 ● 参数名称：只允许输入字母、数字、下划线，且不能以数字开头。 ● 类型、值：支持“引用”和“输入”两种类型。 – 引用：支持用户选择工作流中已包含的前置节点的输出变量值和全局配置中的记忆变量。 – 输入：支持用户自定义取值。	-
	插件	可绑定手动创建的插件或预制插件，当模型识别到需要调用工具来完成任务时，会根据用户的输入提取参数完成插件调用，并总结插件执行结果。 当单击+图标时，可新增插件。	-
	工具使用约束	配置输入给大模型的使用约束，用于指导模型更好地完成任务。配置约束词时，支持使用{{variable}}格式引用当前节点输入参数中已定义好的参数。最终替换后的内容会传递给模型。 ● 当单击“保存到模板”，填写“模板名称”、选择“行业”和“标签”后，可将提示词创建成模板并保存到我的提示词。 ● 当单击  图标时，可对系统提示词进行智能优化。 ● 当单击  图标时，系统会弹出“提示词广场”窗口，可在“预制提示词”或“我的提示词”页签中进行选择。	-
终止条件	最大迭代轮次	该参数用于设置与模型的最大交互次数，超过最大回复轮数还没有提取到参数则跳出Agent节点。	9
	插件执行成功	该参数开启后可绑定插件，当执行该插件成功后跳出Agent节点。	关闭

配置类型	参数名称	参数说明	配置示例
	识别到用户有退出意图	该参数开启后识别到用户输入有退出意向时，跳出Agent节点。	开启
参数配置	输出参数	输出参数为Agent节点最后一轮的输出。	-

图 8-48 Agent 节点配置示例

步骤6 节点配置完成后，单击“确定”。

步骤7 连接Agent节点和其他节点。

----结束

8.9 逻辑节点

8.9.1 判断

判断节点是一个IF-ELSE节点，提供了多分支条件判断的能力，用于设计分支流程，实现逻辑判断功能。

前提条件

已完成工作流搭建，如果未搭建工作流，请参考**搭建工作流**。

判断节点说明

判断节点中每个条件分支支持添加多个判断条件（且、或），同时支持添加多个条件分支。

当向该节点输入参数时，节点会逐个条件分支判断输入是否符合分支中预设的条件，符合则执行对应分支后的工作流流程，如果没有符合条件的分支，则执行“ELSE”对应的工作流分支。

配置判断节点

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏选择“开发中心 > 应用管理 > 工作流应用”，单击您创建的工作流。

步骤3 单击“添加节点”并选择“判断”节点。

步骤4 通过单击该节点打开节点配置页面。

步骤5 参照**图8-49**和**表8-13**，完成判断节点的配置。

说明

- 单击图标，可修改判断名称，修改完成后单击名称旁边的进行保存。
- 单击图标，可重命名判断节点名称，复制一个判断节点或删除判断节点。

表 8-13 判断节点配置说明

参数名称	参数说明	配置示例
IF	IF分支由**[判断参数 比较条件 比较参数]**组成一个条件表达式。 ● 判断参数：条件表达式左边部分，需要选择来自前序节点的输出参数。 ● 比较条件：条件表达式中间部分，当前支持的比较条件有：长度大于、长度大于等于、长度小于、长度小于等于、等于、不等于、包含、不包含、为空、不为空。针对不同的判断参数类型，前端将展示不同的比较条件。 ● 比较参数：条件表达式右边部分，支持“引用”和“输入”两种类型。 – 引用：支持用户选择工作流中已包含的前置节点输出变量值及全局配置中的记忆变量。 – 输入：支持用户自定义取值。 ● 添加条件：单击+，在当前条件分支中添加多个条件表达式，多个条件表达式之间通过“且”或“或”来连接。 单击“且”或“或”，可以切换该分支表达式的运算逻辑。	参见**示例**。
ELSE	用于控制预设条件分支都不满足的场景，如果逐个分支判断都不符合条件，则默认走该分支执行后续工作流流程。	/
添加分支	可以添加新的条件分支ELSE IF，新分支的配置方式与IF分支相同。	例如上游节点输出一个结果参数“result”，IF分支中判断“result”等于true，新增条件分支ELSE IF判断“result”等于false，根据不同的结果执行不同的后续流程。

图 8-49 判断节点配置示例

步骤6 节点配置完成后，单击“确定”。

步骤7 连接判断节点和其他节点。

----结束

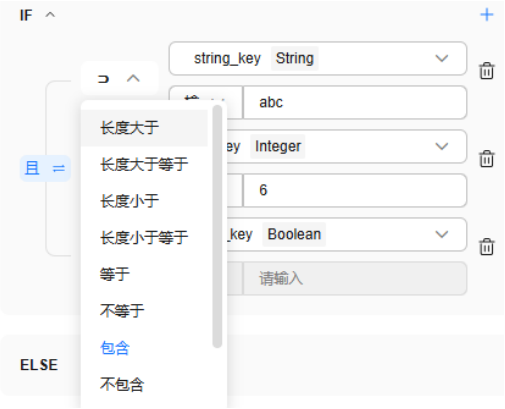
示例

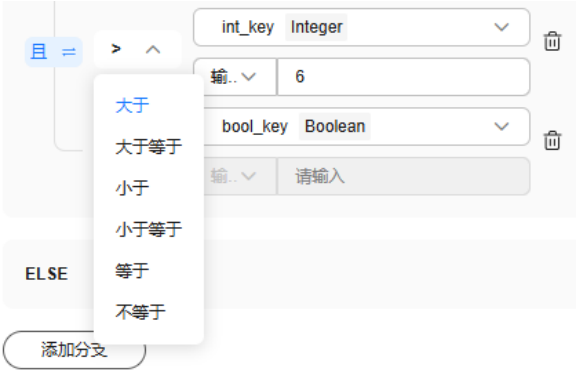
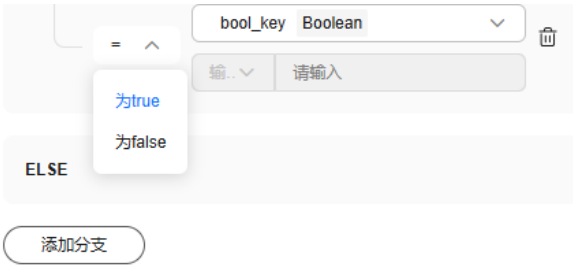
以常见的String、Integer和Boolean类型条件判断为例，在开始节点中定义三种类型的参数，模拟判断节点的输入参数，实现对于不同类型参数在不同条件下的逻辑判断。

节点配置如下：

- 开始节点：定义三种类型参数，分别为String类型的string_key、Integer类型的int_key、Boolean类型的bool_key。
- 判断节点：在IF条件分支中增加三个判断条件，条件表达式的判断参数分别引用开始节点上述的三种类型参数。对于不同类型的参数，前端展示的比较条件有所区别。

表 8-14 开始节点配置示例

参数类型	参数名称	配置示例
String类型	IF	例如String类型为字符串相关的长度、包含和为空条件判断，示例中配置为判断string_key是否包含“abc”。 

参数类型	参数名称	配置示例
Integer类型	IF	Integer类型为数值相关的大小等于条件判断，示例中配置为判断int_key是否大于6。 
Boolean类型	IF	Boolean类型为true false条件判断，示例中配置为判断bool_key是否为true。 

单击试运行，输入string_key：abcd、int_key：7、bool_key：true查看效果。

图 8-50 试运行

8.9.2 代码

代码节点支持通过编写Python或Node.js代码来处理文本等复杂逻辑，生成业务期望的返回值。

前提条件

已完成工作流搭建，如果未搭建工作流，请参考**搭建工作流**。

如果使用FunctionGraph执行方式，请确保当前华为账号或IAM用户具备FunctionGraph的权限，如何获取FunctionGraph的权限请参见**授权使用FunctionGraph**。

配置代码节点

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏选择“开发中心 > 应用管理 > 工作流应用”，单击您创建的工作流。

步骤3 单击“添加节点”并选择“代码”节点。

步骤4 通过单击该节点打开节点配置页面，选择“FunctionGraph”执行方式。

说明

- 单击图标，可修改代码节点名称，修改完成后单击名称旁边的进行保存。
- 单击图标，可重命名代码节点名称，复制一个代码节点或删除代码节点。
- 单击图标，可对代码点进行测试。

FunctionGraph：使用FunctionGraph函数，只需编写业务函数代码并设置运行的条件，无需配置和管理服务器等基础设施，函数以弹性、免运维、高可靠的方式运行。FunctionGraph配置方式请参照**表8-15**和**图8-52**。

注意

如果使用FunctionGraph执行方式，请确保当前华为账号或IAM用户具备FunctionGraph的权限，如何获取FunctionGraph的权限请参见**授权使用FunctionGraph**。

表 8-15 FunctionGraph 执行方式配置参数说明

参数	说明
执行方式	FunctionGraph：支持依赖包管理、网络访问等等高级功能，支持多语言。
函数名称	选择下拉列表中的函数，即之前已定义保存的函数，也可以进行以下操作。 • 单击：可以直接在弹出的创建函数页面快速创建函数，参数说明如**表8-16**所示，参数配置完成后可单击“创建”保存函数。 • 单击：选择函数后，单击该图标可以在弹出的“编辑函数”页面中快速编辑函数，参数编辑完成后可单击“更新”保存函数。

参数	说明
输入参数	按照函数定义中指定的参数列表配置入参，即传递给函数的实际值。 输入参数或选择前序节点的输出作为输入。
输出参数	配置代码运行后需要输出的参数。
异常处理	支持对节点的异常（如超时、调用失败等情况）进行处理，包括超时时间、重试次数、异常处理方式。 “超时时间”：支持用户配置超时时间，取值范围0.1~900，默认900s。 “重试次数”：支持配置重试次数（不重试、重试1次、重试2次、重试3次），系统默认不重试。 “异常处理方式”：配置异常处理方式。 • 中断流程：节点发生异常后，直接中断流程，不再运行后续节点。 • 返回设定内容：节点发生异常后，工作运行不会中断，用户可自定义设置需要返回的输出字段内容，必须是输出参数中已定义的字段，且格式为合法的JSON格式。 • 执行异常流程：节点发生异常后，工作流不会中断，而是会执行异常处理流程。用户可以在该运行异常的节点前新增节点，并为新增的异常分支配置相应的处理流程。

表 8-16 创建函数参数说明

参数	说明
名称	函数名，用于调用函数。
描述	函数功能描述。
入参	输入参数。
出参	输出参数。每个变量都可在后置节点中引用。
执行语言	当前支持Python3.9、Node.js14.18，即运行函数的环境，请查看**Python函数开发指南**、**Node.js函数开发指南**。

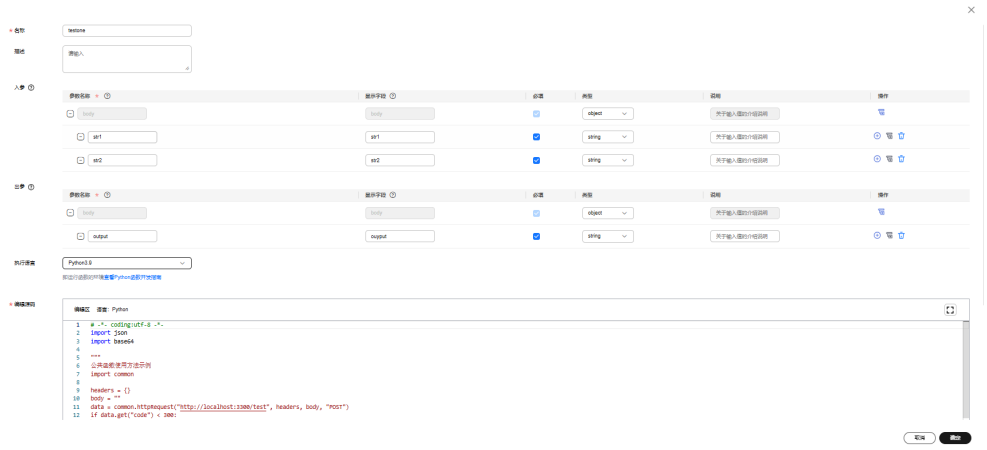
参数	说明
编辑源码	在源码编辑区，编写函数内部的代码运行逻辑，如**图8-51**所示，图中各模块说明如下： ①：导入模块，是Python标准库中的模块，无需修改。 ②：用户自定义导入模块。 ③：公共函数使用方法示例，提供了如何使用公共函数和mssiAuthData参数的示例，无需修改。 ④：函数定义和注释，extractRequestParam函数和handler函数是系统预置的模板代码，无需修改。 ⑤：系统方法，无需修改。 ⑥：用户自定义函数中的逻辑。输出为JOSN格式，请参考**图4 源码编辑区**的输出格式，详细配置请参见**配置示例**。
依赖包	单击“添加”，可以选择自定义依赖包。自定义依赖包上传方法请参见**创建自定义依赖包**。 一个函数最多添加**20个**依赖包。

图 8-51 源码编辑区

编辑区 语言：Python

```python
# -*- coding:utf-8 -*-
import json
import base64
import datetime

"""
公共函数使用方法示例
import common

headers = {}
body = ""
data = common.httpRequest("http://localhost:3300/test", headers, body, "POST")
if data.get("code") < 300:
    return data.get("body")
return "error: " + data.get("error")

接口返回res = {"headers": {},
               "body": string,
               "code": number,
               "error": string}
"""

"""
mssiAuthData参数样例
{
    "header":{}, // 连接器认证header参数
    "path": {}, // 连接器认证path参数
    "query":{}, // 连接器认证query参数
    "body":{}, // 连接器认证body参数
    "host":"https://demo.com // API主机地址
}
"""


def extractRequestParam(rawValue, encoded, defaultValue):
    if encoded and rawValue:
        rawValue = str(base64.b64decode(rawValue), "utf-8")
    return json.loads(rawValue) if rawValue else defaultValue


## 请勿对下面的函数做修改
def handler(event, context):
    """
    函数是方法的入口
    :param event: 执行事件（event），包含用户定义的函数参数以及所选择的的连接器认证相关参数
    :param context: Runtime提供的函数执行上下文
    :return:
    """

    isBase64Encoded = event.get('isBase64Encoded', False)
    inputData = extractRequestParam(event.get('body'), isBase64Encoded, {})  # 用户定义的函数参数数据
    mssiAuthData = extractRequestParam(event.get('mssiAuthData'), isBase64Encoded, {})  # 连接器认证数据
    mssiAuthData["securityToken"] = context.getToken()

    dataExtendConfig = extractRequestParam(event.get('dataExtendConfig'), isBase64Encoded, {})  # 流步骤扩展参数

    origin_time = inputData.get('time')
    print(origin_time)
    # 字符串转datetime
    dt_obj = datetime.datetime.strptime(origin_time, '%Y-%m-%d %H:%M:%S')

    # datetime转字符串
    formatted_str = dt_obj.strftime('%d/%m/%Y %H:%M:%S')

    result = {"formatted_time":formatted_str}
    return json.dumps(result)
```


图 8-52 创建函数



示例

1. **开发语言**

   代码节点以Python语言为例。

2. **Python**

   基于Python 3.11.3的标准库，大多数模块都能正常运行，如下面白名单所示模块，不在白名单中的模块可能不能正常运行。

   - 三方库白名单

     sys,time,numpy,warnings,enum,os,functools,collections,types,datetime,numbers,abc,io,executor_sdk,contextlib,dataclasses,math,operator,pickle,contextvars,_contextvars,ast,re,ctypes,copyreg,weakref,textwrap,platform,typing,__future__,sympy,mpmath,bisect,cmath,colorsys,keyword,linecache,timeit,gc,random,decimal,_decimal,fractions,flint,gmpy2,unicodedata,tokenize,gmpy,copy,inspect,string,struct,importlib,array,shutil,pathlib,tempfile,subprocess,json,xml.etree.ElementTree,uuid,_uuid,urandom

   - 内置函数白名单

     exec,print,id,issubclass,compile,__build_class__,hasattr,eval,chr,next,ord,callable,repr,sorted,iter,min,max,weakref,all,any,hash,locals,sum,vars,open,abs,round,divmod,pow,delattr

3. 配置示例

   以数学计算示例代码为例：

```
# -*- coding:utf-8 -*-
import json
import base64
def extractRequestParam(rawValue, encoded, defaultValue):
    if encoded and rawValue:
        rawValue = str(base64.b64decode(rawValue), "utf-8")
    return json.loads(rawValue) if rawValue else defaultValue
def math(args: dict) -> dict:
    # 注意在输入参数中定义名为input1的变量
    input1 = args.get('input1')
    try:
        input1 = int(input1)
        return {
            # 注意输出参数中定义res变量
            'res': input1 * input1
        }
    except Exception as e:
        return {
            # 注意输出参数中定义res变量
            'res': "输入类型错误或者数字大小超出限制"
        }
## 请勿对下面的函数名做修改
def handler(event, context):
    """
```



```
    函数是方法的入口
    :param event: 执行事件（event），包含用户定义的函数参数以及所选择的的连接器认证相关参数
    :param context: Runtime提供的函数执行上下文
    :return:
    """
    isBase64Encoded = event.get('isBase64Encoded', False)
    inputData = extractRequestParam(event.get('body'), isBase64Encoded, {}) # 用户定义的函数参数数据
    mssiAuthData = extractRequestParam(event.get('mssiAuthData'), isBase64Encoded, {}) # 连接器认证数据
    mssiAuthData["securityToken"] = context.getToken()
    result = {'output1':math(inputData.get('input1'))}
    return json.dumps(result)
```

步骤5 节点配置完成后，单击“确定”。

步骤6 连接代码节点和其他节点。

----结束

创建自定义依赖包

函数运行环境内置了常用公共依赖包，支持在函数中引用；同时也支持用户创建自定义依赖包。本节介绍如何创建自定义依赖包。

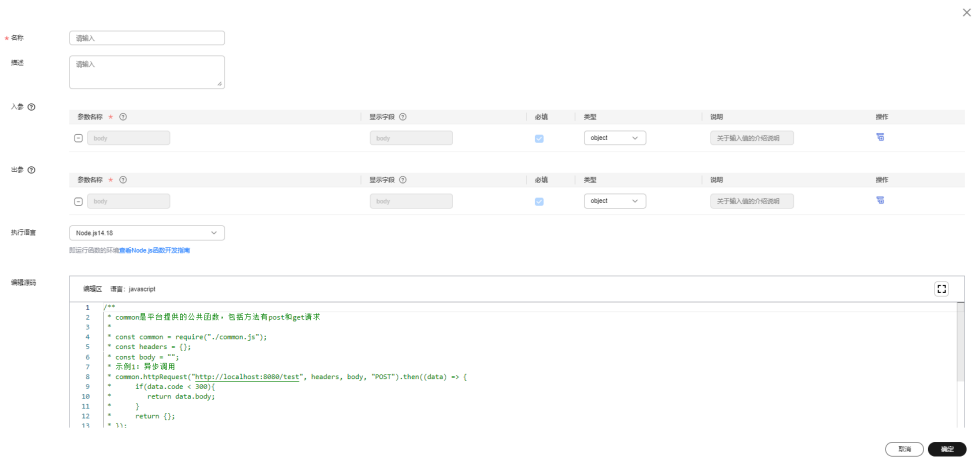
步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏选择“开发中心 > 应用管理 > 工作流应用”，单击您创建的工作流。

步骤3 单击“代码”节点，进入节点配置界面，“执行方式”选择为“FunctionGraph”。

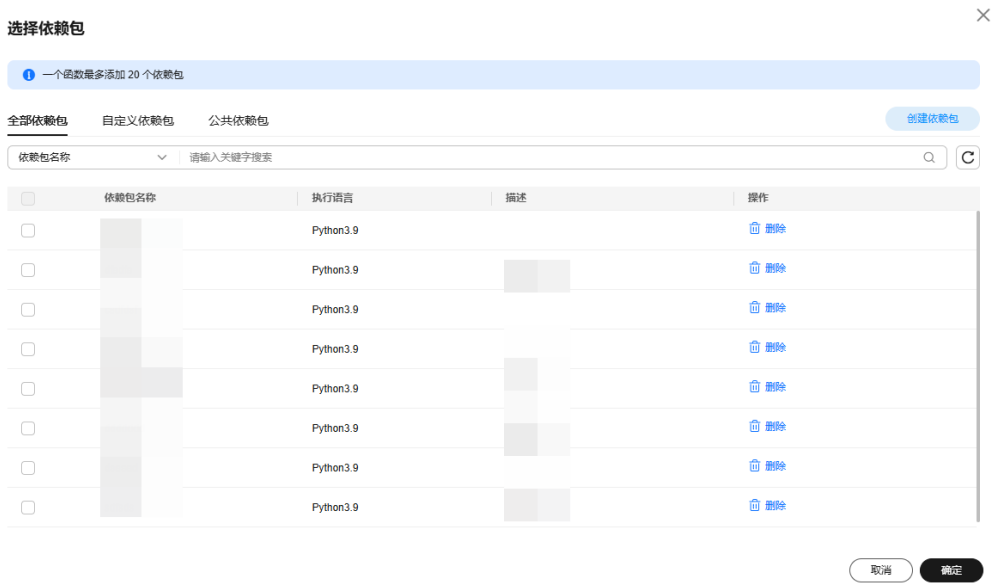
步骤4 选择已有函数，进入编辑函数界面，如图8-53所示。

图 8-53 创建函数



步骤5 单击“依赖包”后的“添加”，如图8-54所示。

图 8-54 选择依赖包



步骤6 单击“创建依赖包”，设置依赖包的基本配置信息，具体的参数说明如**表8-17**所示。

表 8-17 新建依赖包参数说明

参数	说明
依赖包名称	自定义依赖包的名称，支持英文、数字、下划线，仅支持以英文开头，长度为2-32个字符。
执行语言	运行函数的环境，当前仅支持Python3.9、Node.js14.18。
描述（可选）	依赖包的描述信息，最多支持200个字符。
上传（支持多个文件）	上传.zip格式文件，文件大小限制为10MB以内。 上传文件时，如果文件中包含敏感信息（如账户密码等），请您自行加密，防止信息泄露。

步骤7 单击“确定”。

创建完成后，可以在代码节点中添加并使用该依赖包。

----结束

8.9.3 循环

循环节点提供了循环执行节点的能力，可在循环体内配置需要循环的节点，用户可以通过在循环体内编排节点多次执行处理任务。

前提条件

已完成工作流搭建，如果未搭建工作流，请参考**搭建工作流**。

约束与限制

循环节点内不支持配置问答节点、提问器节点、Agent节点、输入节点或带有这些节点的子工作流。

节点说明

循环是一种常见的控制机制，用于重复执行一系列任务，直到满足某个条件为止。工作流提供循环节点，当需要重复执行一些操作，或循环处理一组数据时，可以使用循环节点实现。

配置循环节点

步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏选择“开发中心 > 应用管理 > 工作流应用”，单击您创建的工作流。

步骤3 单击“添加节点”并选择“循环”节点。

步骤4 拖动其他需要循环执行的节点到循环体画布内部并编排（循环内执行需从循环输入节点开始，输出连接到循环输出节点，暂不支持交互式节点）。

说明

只能在“添加节点”中添加，已添加至画布中的循环节点无法拖拽。

步骤5 参照表8-18，完成循环节点的配置。

说明

- 单击图标，可修改循环节点名称，修改完成后单击名称旁边的进行保存。
- 单击图标，可重命名循环节点名称，复制一个循环节点或删除循环节点。

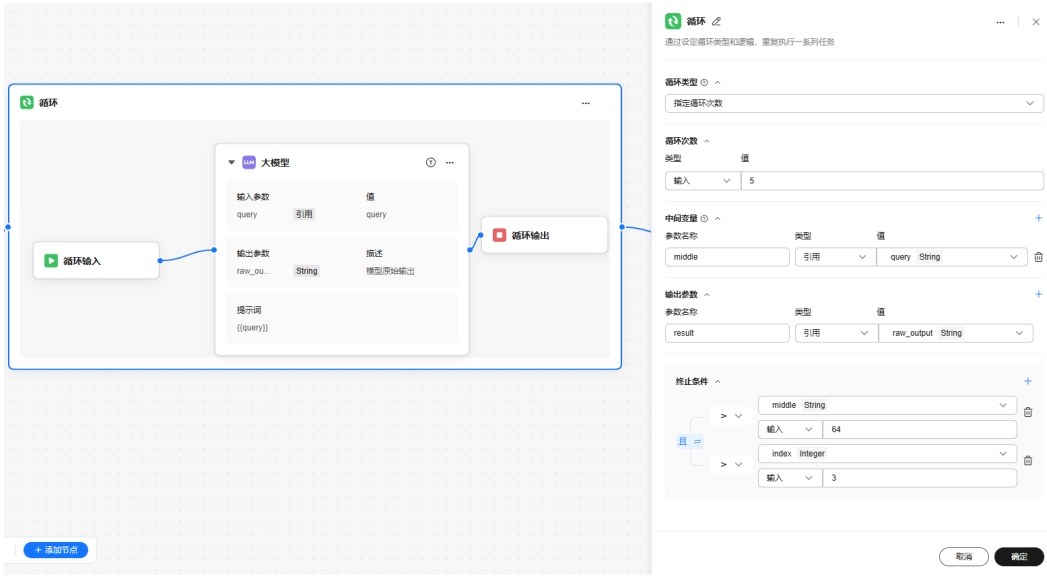
表 8-18 循环节点配置说明

配置类型	参数名称	参数说明
循环类型配置	使用数组循环	使用数组循环类似编程语言中的for语法结构。遍历循环用于遍历一个已知的序列，对序列中的每个元素执行一系列相同的步骤。 使用数组循环时，需要指定arr_loop_var的值，此参数仅支持引用上游节点的输出，且必须为数组格式。使用数组循环模式下执行循环节点时，循环的次数取决于循环数组引用的数组长度。 使用数组循环时，循环节点会遍历数组中的每个元素，每次循环都会将当前循环到的元素赋值给内置变量。内置变量仅限循环节点内部使用。目前支持的内置变量如下： • item：数组元素，即当前循环到的数组元素。 • index：数组索引，index+1为当前循环的轮次。

配置类型	参数名称	参数说明
	指定循环次数	指定循环次数模式通常用于批量、顺序处理数据的场景，需要同时设置循环次数。 循环次数默认为5次，支持设置为1~1000次。 使用参考： 回合制游戏，3局2胜可将循环次数设置为3。 网络爬虫爬取前1000个商品信息，循环次数设置为1000。
变量配置	类型/值	此参数只有在使用数组循环时支持配置，名称固定为arr_loop_var，仅支持引用上游节点输出。
	中间变量	循环节点支持设置中间变量，此变量可作用于每一次循环。中间变量通常和循环体中的设置变量节点搭配使用，在每次循环结束后为中间变量设置一个新的值，并在下次循环中使用新值。 当单击+图标时，可新增中间变量。 ● 参数名称：只允许输入字母、数字、下划线，且不能以数字开头。 ● 类型、值：支持“引用”和“输入”两种类型。 – 引用：支持用户选择工作流中已包含的前置节点的输出变量值和全局配置中的记忆变量。 – 输入：支持用户自定义取值。
	输出参数	循环节点的输出参数可设置为循环体的执行结果集合，表示当数组中所有元素运行完毕之后，将所有循环的运行结果打包输出给下游。也支持设置为中间变量的取值。 当单击+图标时，可新增输出参数。

配置类型	参数名称	参数说明
终止条件	表达式	分支由**[判断参数 比较条件 比较参数]**组成一个条件表达式。 当单击+图标时，可新增终止条件。 • 判断参数：条件表达式上半部分，需要选择来自前序节点的输出参数。 • 比较条件：条件表达式左侧，当前支持的比较条件有：长度大于、长度大于等于、长度小于、长度小于等于、等于、不等于、包含、不包含、为空、不为空。针对不同的判断参数类型，前端将展示不同的比较条件。 • 比较参数：条件表达式下半部分，支持“引用”和“输入”两种类型。 – 引用：支持用户选择工作流中已包含的前置节点输出变量值及全局配置中的记忆变量。 – 输入：支持用户自定义取值。 • 添加条件：单击“+”，在当前条件分支中添加多个条件表达式，多个条件表达式之间通过“且”或“或”来连接。 – 单击“且”或“或”，可以切换该分支表达式的运算逻辑。

图 8-55 循环节点配置示例



步骤6 节点配置完成后，单击“确定”。

步骤7 连接循环节点和其他节点。

----结束

8.9.4 意图识别

意图识别节点主要是让应用理解用户自然语言表达的意图或目的，可用于需要对用户问题进行分类，或者提供综合类功能有不同分支处理的场景。

前提条件

已完成工作流搭建，如果未搭建工作流，请参考**搭建工作流**。

节点说明

意图识别节点通过对用户输入进行推理分析，匹配预定义的意图关键字类别，并根据匹配结果引导至相应的处理流程，该节点通常位于工作流的前置位置。

意图识别节点支持普通模式或高级模式运行。

- 普通模式：适用于对少量意图进行分类的场景。
- 高级模式：适用于对大量可归类意图进行分类的场景。

配置意图识别节点

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏选择“开发中心 > 应用管理 > 工作流应用”，单击您创建的工作流。

步骤3 单击“添加节点”并选择“意图识别”节点。

步骤4 通过单击该节点打开节点配置页面。

步骤5 参照**图8-56**和**表8-19**，完成意图识别节点的配置。

说明

- 单击图标，可修改意图识别节点名称，修改完成后单击名称旁边的进行保存。
- 单击图标，可重命名意图识别节点名称，复制一个意图识别节点或删除意图识别节点。
- 单击图标，可对意图识别节点进行测试。

表 8-19 意图识别节点配置说明

配置类型	参数名称	参数说明	配置示例
模型配置	模型配置	用于配置进行意图识别的大模型，可选择平台已接入的任一模型。	/
	温度	当单击图标时，可进行该参数设置。 用于控制生成结果的随机性。调高温度，会使得模型的输出更具多样性和创新性；降低温度，会使输出内容更加遵循指令要求，但同时也会减少模型输出的多样性。	0.5

配置类型	参数名称	参数说明	配置示例
	核采样	当单击图标时，可进行该参数设置。 模型在输出时会从概率最高的词汇开始选择，直到这些词汇的总概率累积达到核采样值，核采样值可以限制模型选择这些高概率的词汇，从而控制输出内容的多样性。建议不要与温度同时调整。	0.5
	最大回复长度	当单击图标时，可进行该参数设置。 控制模型输出的Tokens长度上限。通常100Tokens约等于150个中文汉字。	131072
参数配置	输入参数	• 参数名称：默认名称input，为固定值，不可编辑。 • 类型、值：支持“引用”和“输入”两种类型。 – 引用：支持用户选择工作流中已包含的前置节点的输出变量值和全局配置中的记忆变量。 – 输入：支持用户自定义取值。	一般选择“引用”开始节点的输入参数“query”，即对用户输入进行意图识别。
意图配置	意图1	用于配置相关意图关键字信息，用户可以添加意图，意图类别默认为意图1、意图2...，意图数量最多为20个。 在意图输入框中输入意图描述信息，描述信息为针对该类别的描述语句或者关键词，也将作为大模型进行推理和分类的依据。	意图的设置和工作流中定义的处理流程相关，例如一个旅游助手工作流，提供天气查询、预订机票、预订酒店等能力，根据用户输入执行上述任一功能。按照对应能力定义意图关键字“天气查询”、“预订机票”、“预订酒店”。
	其他意图	用于控制用户输入意图无法识别的场景，如果推理分析后无法匹配预定义的意图分类，会默认走其他意图对应分支执行后续流程。	其他意图主要用于处理上述定义意图无法匹配时的兜底逻辑，例如意图无法识别时需要返回一个兜底回复，可以在其他意图后接一个消息节点，消息节点中定义兜底回复的内容。定义意图无法识别时，触发“其他意图”分支，执行消息节点返回兜底消息。

配置类型	参数名称	参数说明	配置示例
高级配置	提示词	高级可选配置项，提供进阶开发者修改提示词，如果不配置将会使用系统默认值。提示词的撰写可能影响到意图识别节点的准确性。 • 当单击“保存到模板”，填写“模板名称”、选择“行业”和“标签”后，可将提示词创建成模板并保存到我的提示词。 • 当单击  图标时，可对系统提示词进行智能优化。 • 当单击  图标时，系统会弹出“提示词广场”窗口，可在“预制提示词”或“我的提示词”页签中进行选择。	高级配置，可使用默认的提示词。当意图识别效果没有达到预期时，可以调整提示词优化效果。例如可以在提示词中补充“用户提问飞机时，识别为预订机票功能。”，提升“预订机票”意图识别成功率。
	历史对话轮次	选择是否打开历史对话引用功能，默认为0即不会引用对话历史，配置N轮即可记录N轮对话的内容。	-
	辅助识别	开启辅助识别后，优先通过知识库分类样例的精确匹配进行意图识别，提升意图识别节点的分类能力。 • 意图样例知识库：开启辅助识别，用户需要先创建分类样例知识库，向知识库上传意图FAQ，并选择配置该知识库。 • 过滤标签：可填写意图样例知识库上传FAQ时打的标签值，表示在该标签范围内进行FAQ检索匹配。如果不填写，则默认在整个知识库范围下做FAQ检索匹配。 • 匹配阈值：当分类样例的匹配度低于设置阈值时，会采用默认的大模型进行意图识别分类。阈值范围为0到1。	创建“词语分类”样例知识库，上传多对FAQ： 问题：三国演义；答案：文学作品 问题：光刻机；答案：科技 在意图识别节点配置意图为“文学作品”、“科技”，辅助识别选择“词语分类”知识库，匹配阈值设置0.9。
参数配置	输出参数	输出参数为判断节点最后一轮的输出。	-

图 8-56 意图识别节点配置示例 1

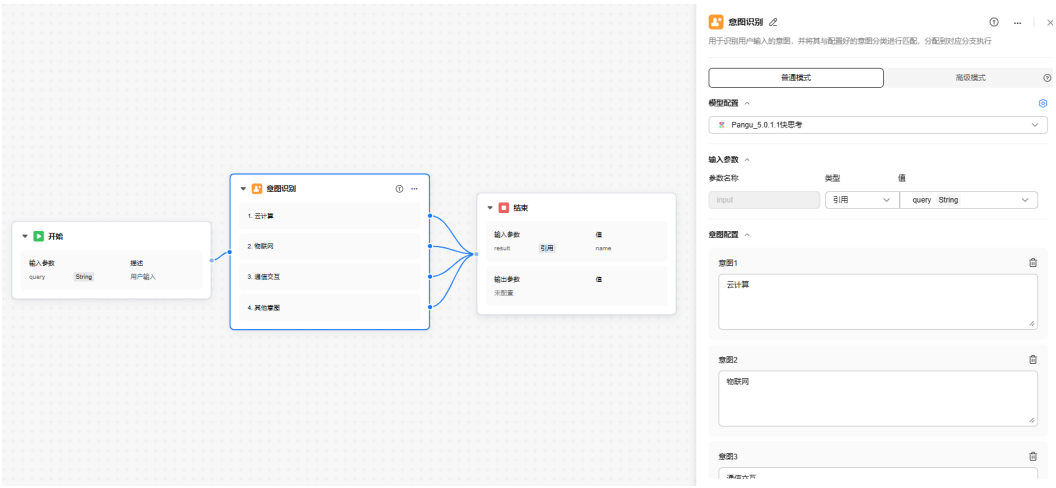


图 8-57 意图识别节点配置示例 2

步骤6 节点配置完成后，单击“确定”。

步骤7 连接意图识别节点和其他节点。

----结束

示例

以提供综合功能，对用户问题进行不同分支处理的工作流为例，通过意图识别节点对用户输入进行分类，流转至不同的功能模块进行处理。

比如提供翻译功能的工作流，节点配置如下：

图 8-58 配置示例

意图识别节点：

工作流将用户问题分为翻译、闲聊两个类别，节点的意图配置添加意图1的类别描述为“文本翻译”，类别后面连接翻译插件节点，实现翻译功能。

意图2的类别描述为“用户闲聊”，类别后面连接大模型节点，实现闲聊功能。

默认的其他意图类别后面连接消息节点，在消息节点中配置默认回复内容，实现未识别意图场景下的兜底回复。

8.9.5 高级意图识别

意图识别节点主要是让应用理解用户自然语言表达的意图或目的，适用于编排超过20个以上意图的分支逻辑。

当意图分支比较多如大于100时，建议使用高级模式。 在独立的页面配置意图分支信息，通过选择子工作流的交互方式完成业务配置。

前提条件

已完成工作流搭建，如果未搭建工作流，请参考**搭建工作流**。

节点说明

意图识别节点支持高级模式运行，适用于对大量可归类意图进行分类的场景。

配置高级意图节点

步骤1 配置意图包。

1. 单击平台左侧菜单 “配置管理 > 意图管理” 新建意图包。
2. 在意图包中添加意图分类，分类信息包含名称和示例。

图 8-59 配置意图包

步骤2 在工作流应用编辑页面，单击“添加节点”，选择“高级意图识别”节点，单击该节点以打开节点配置页面。

步骤3 参照意图模式配置说明，完成配置。

表 8-20 高级意图识别节点配置说明

配置类型	参数名称	参数说明
模型配置	模型配置	用于配置进行意图识别的大模型，可选择平台已部署的任一模型。
	温度	当单击图标时，可进行该参数设置。 用于控制生成结果的随机性。调高温度，会使得模型的输出更具多样性和创新性；降低温度，会使输出内容更加遵循指令要求，但同时也会减少模型输出的多样性。
	核采样	当单击图标时，可进行该参数设置。 模型在输出时会从概率最高的词汇开始选择，直到这些词汇的总概率累积达到核采样值，核采样值可以限制模型选择这些高概率的词汇，从而控制输出内容的多样性。建议不要与温度同时调整。
	最大回复长度	当单击图标时，可进行该参数设置。 控制模型输出的Tokens长度上限。通常100Tokens约等于150个中文汉字。
参数配置	输入参数	• 参数名称：默认名称input，为固定值，不可编辑。 • 类型、值：支持“引用”和“输入”两种类型。 – 引用：支持用户选择工作流中已包含的前置节点的输出变量值和全局配置中的记忆变量。 – 输入：支持用户自定义取值。
意图配置	意图包	选择已经配置的意图包。

配置类型	参数名称	参数说明
高级配置	提示词	高级可选配置项，提供进阶开发者修改提示词，如果不配置将会使用系统默认值。提示词的撰写可能影响到意图识别节点的准确性。 • 当单击“保存到模板”，填写“模板名称”、选择“行业”和“标签”后，可将提示词创建成模板并保存到我的提示词。 • 当单击  图标时，可对系统提示词进行智能优化。 • 当单击  图标时，系统会弹出“提示词广场”窗口，可在“预制提示词”或“我的提示词”页签中进行选择。
	历史对话轮次	选择是否打开历史对话引用功能，默认为0即不会引用对话历史，配置N轮即可记录N轮对话的内容。
参数配置	输出参数	输出参数为判断节点最后一轮的输出。

图 8-60 意图识别节点配置示例

步骤4 节点配置完成后，单击“确定”。

步骤5 单击意图动作节点，配置分支对应的处理逻辑。

图 8-61 配置处理逻辑

步骤6 配置子工作流的输入参数。

图 8-62 配置输入参数

步骤7 单击“确定”，完成意图动作节点配置。

步骤8 连接意图动作节点和其他节点。

----结束

8.10 工具节点

8.10.1 插件

插件节点是工作流中实现第三方能力调用的核心组件。

前提条件

- 已完成工作流搭建，如果未搭建工作流，请参考**搭建工作流**。
- 如果需要添加共享插件，请确保已有他人共享的插件。
- 仅**Versatile企业版**支持使用他人共享的应用。**Versatile基础版（限时免费）**不支持该能力。

节点说明

作为功能扩展的重要载体，该节点允许通过调用插件来执行特定功能任务。每个插件实质上是经过标准化封装的API工具集合，提供即插即用的模块化服务，拓宽工作流的能力边界，完成更复杂的任务。

插件类型包括预置插件和个人插件。

- 预置插件：平台预置了代码解释器插件，能够执行输入的代码，得到运行结果。支持开发者直接将插件添加到工作流或应用中，丰富其能力。
- 个人插件：平台允许开发者创建自定义插件，支持将API通过配置方式快速创建为插件，提供给工作流或应用调用。
- 团队共享：平台支持不同空间之间的插件共享，用户可以查看本空间共享给其他团队的资源，也能查看其他空间共享给本空间的资源，详细操作请参见**使用资产中心的插件资源**。
- 如果“我的插件”和“插件广场”不满足用户需求，可以单击左上角“创建插件”，详细参数配置请参见**基于API创建一个插件**，插件创建成功后在“插件创建成功”界面单击“确定”即可直接添加插件。

配置插件节点

步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏选择“开发中心 > 应用管理 > 工作流应用”，单击您创建的工作流。

步骤3 单击“添加节点”并选择“插件”节点。

步骤4 在“添加插件工具”窗口，选择“我的插件”或“插件广场”，单击目标插件右侧的 展开工具列表，在展开的列表中单击目标工具右侧的“添加”将插件工具添加至画布中。

平台提供的插件分为免费和付费两种:

- **免费插件：**免费插件无需购买，无需鉴权的插件可以直接使用，未鉴权的插件设置鉴权后即可使用。设置鉴权请参考使用平台精选的插件。
- **付费插件：**付费插件需要先购买并设置鉴权后才能使用，单击“获取鉴权信息”可跳转至购买和获取API Key的页面。

步骤5 单击画布中已添加的“插件”节点，参照表8-21，完成插件节点的配置。

说明

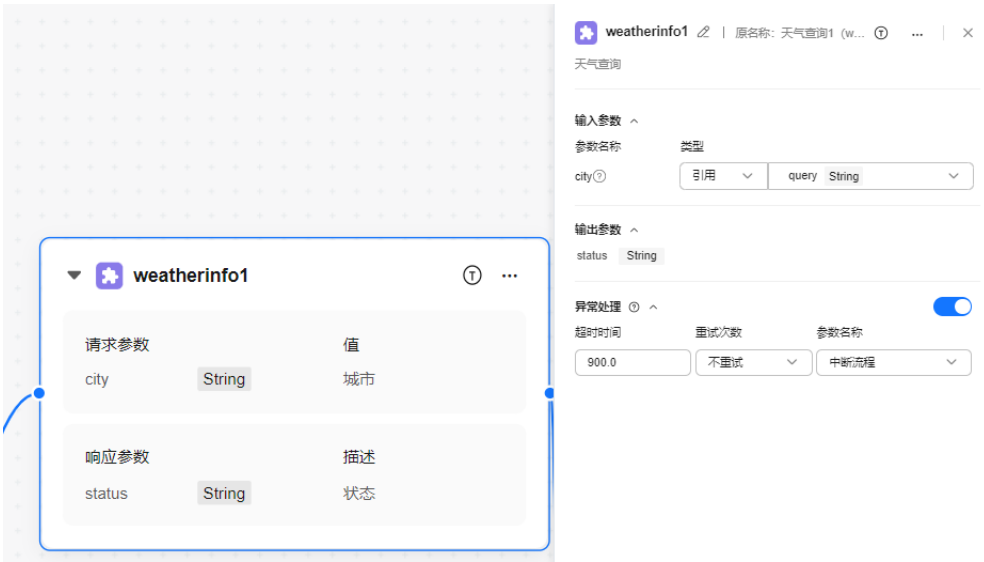
- 单击图标，可修改插件节点名称，修改完成后单击名称旁边的进行保存。
- 单击图标，可查看插件详情，重命名插件节点名称，复制一个插件节点或删除插件节点，同时支持查看插件详情。
- 单击图标，可对插件节点进行测试。

表 8-21 插件节点配置说明

配置类型	参数名称	参数说明	配置示例
参数配置	输入参数	• 参数名称：从插件元信息中导入，用户无需手动添加。 • 类型、值：支持“引用”和“输入”两种类型。 – 引用：支持用户选择工作流中已包含的前置节点的输出变量值和全局配置中的记忆变量，适用于需要从前置节点输出中获取插件入参的场景。 – 输入：支持用户自定义取值，适用于插件入参取值固定的场景。 **说明** 如果插件中设置了默认参数值，这些值将自动填充到输入框中，并且用户可以对其进行修改。	插件的输入参数需要从前置节点中获取时，配置“引用”。 插件的输入参数固定时，如翻译插件要将内容翻译成英文，插件入参to表示翻译后内容的语种，此时应该配置“输入”并赋值“en”。
	输出参数	输出参数所有信息从插件元信息中自动导入，用户无需手动修改。	-

配置类型	参数名称	参数说明	配置示例
	异常处理	支持对节点的异常（如超时、调用失败等情况）进行处理，包括超时时间、重试次数、异常处理方式。 “超时时间”：支持用户配置超时时间，取值范围0.1~900，默认900s。 “重试次数”：支持配置重试次数（不重试、重试1次、重试2次、重试3次），系统默认不重试。 “异常处理方式”：配置异常处理方式。 ● 中断流程：节点发生异常后，直接中断流程，不再运行后续节点。 ● 返回设定内容：节点发生异常后，工作运行不会中断，用户可自定义设置需要返回的输出字段内容，必须是输出参数中已定义的字段，且格式为合法的JSON格式。 ● 执行异常流程：节点发生异常后，工作流不会中断，而是会执行异常处理流程。用户可以在该运行异常的节点前新增节点，并为新增的异常分支配置相应的处理流程。	“超时时间”：900。 “重试次数”：不重试。 “异常处理方式”：中断流程。

图 8-63 插件节点配置示例



步骤6 连接插件节点和其他节点。

步骤7 节点配置完成后，单击“确定”。

----结束

8.10.2 MCP 服务

MCP服务节点是工作流中实现第三方能力调用的核心组件之一。

前提条件

已完成工作流搭建，如果未搭建工作流，请参考**搭建工作流**。

节点说明

作为功能扩展的重要载体，该节点允许通过调用MCP服务来执行特定功能任务。每个MCP服务实质上是一个工具集合，可以提供模块化服务来拓宽工作流的能力边界，完成更复杂的任务。

MCP服务类型目前只支持预置服务。

预置服务：平台预置了“高德地图”、“车票查询工具”、“必应搜索”等多种实用MCP服务，安装后可以一键集成调用。支持开发者在工作流或应用中添加预置MCP服务，丰富其能力。

配置 MCP 服务节点

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 单击左侧导航栏“开发中心 > 应用管理 > 工作流应用”，单击您创建的工作流。

步骤3 单击“添加节点”并选择“MCP服务”节点。

步骤4 单击+，将所需MCP服务添加至画布中，其中有些“预置服务”不能直接添加，需要单击“立即开通”，开通服务后即可添加至画布中。

也可在“添加MCP服务”窗口中，通过单击“创建MCP”，您可以选择基于官方预置MCP或空白模板进行创建，具体步骤参见**创建MCP服务**。创建成功后，通过单击提示信息中的“直接添加”，可立即将新创建的MCP添加至当前工作流中。

也可在“添加MCP服务”窗口中，通过单击“创建MCP”后的下拉框选择创建MCP服务，您可以选择基于官方预置MCP或空白模板进行创建，具体步骤参见**创建MCP服务**，创建成功后，通过单击提示信息中的“直接添加”，可立即将新创建的MCP添加至当前工作流中。或通过第三方服务安装，具体步骤可参见**使用资产中心的MCP资源**。

步骤5 连接MCP服务节点和其他节点。

步骤6 单击画布中已添加的“MCP服务”节点，参照**表1 MCP服务节点配置说明**，完成MCP服务节点的配置。

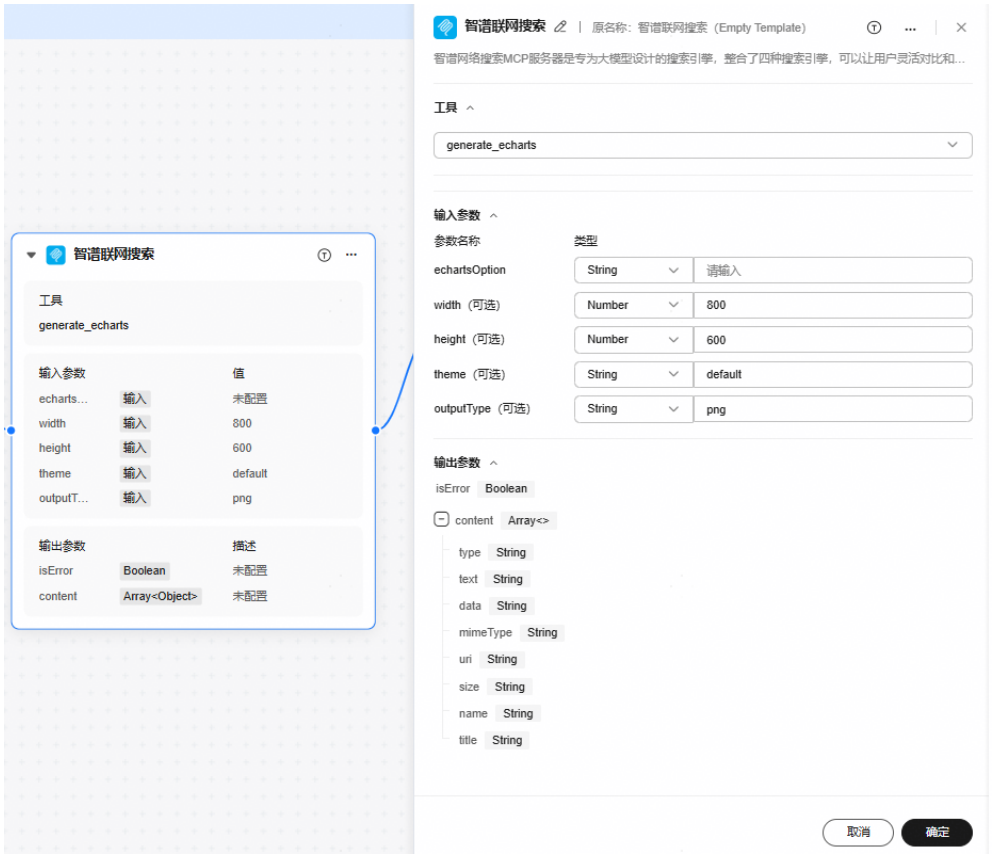
说明

- 单击图标，可修改MCP服务节点名称，修改完成后单击名称旁边的进行保存。
- 单击图标，可重命名MCP服务节点名称，复制一个MCP服务节点或删除MCP服务节点，同时支持查看MCP服务详情。
- 单击图标，可对MCP节点进行测试。

表 8-22 MCP 服务节点配置说明

配置类型	参数名称	参数说明	配置示例
参数配置	工具	支持从当前MCP服务所包含的工具列表中选择一个作为工作流运行到该节点时会执行的工具。	-
	输入参数	• 参数名称、类型：从插件元信息中导入，用户无需手动添加。 • 值：支持“引用”和“输入”两种类型。 – 引用：支持用户选择工作流中已包含的前置节点的输出变量值和全局配置中的记忆变量，适用于需要从前置节点输出中获取插件入参的场景。 – 输入：支持用户自定义取值，适用于MCP服务入参取值固定的场景。	MCP服务工具的输入参数需要从前置节点中获取时，配置“引用”。 MCP服务工具的输入参数固定时，如翻译工具要将内容翻译成英文，入参to表示翻译后内容的语种，此时应该配置“输入”并赋值“en”。
	输出参数	输出参数所有信息从MCP服务元信息中自动导入，用户无需手动配置。	-

图 8-64 MCP 服务节点配置示例



步骤7 节点配置完成后，单击“确定”。

----结束

8.11 消息管理节点

8.11.1 消息

消息节点可提供中间过程的消息输出能力，通过定义一段文本内容，在工作流执行过程中向用户发送该消息。

前提条件

已完成工作流搭建，如果未搭建工作流，请参考**搭建工作流**。

节点说明

通常情况下，工作流会在执行完毕后通过结束节点输出最终的执行结果。当开发者想要在工作流执行过程中输出中间节点的结果，可以使用该节点。

消息节点支持引用流式和非流式输出参数，如大模型节点、插件节点的输出参数。

配置消息节点

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏选择“开发中心 > 应用管理 > 工作流应用”，单击您创建的工作流。

步骤3 单击“添加节点”并选择“消息”节点。

步骤4 通过单击该节点打开节点配置页面。

步骤5 参照**表8-23**，完成消息节点的配置。

说明

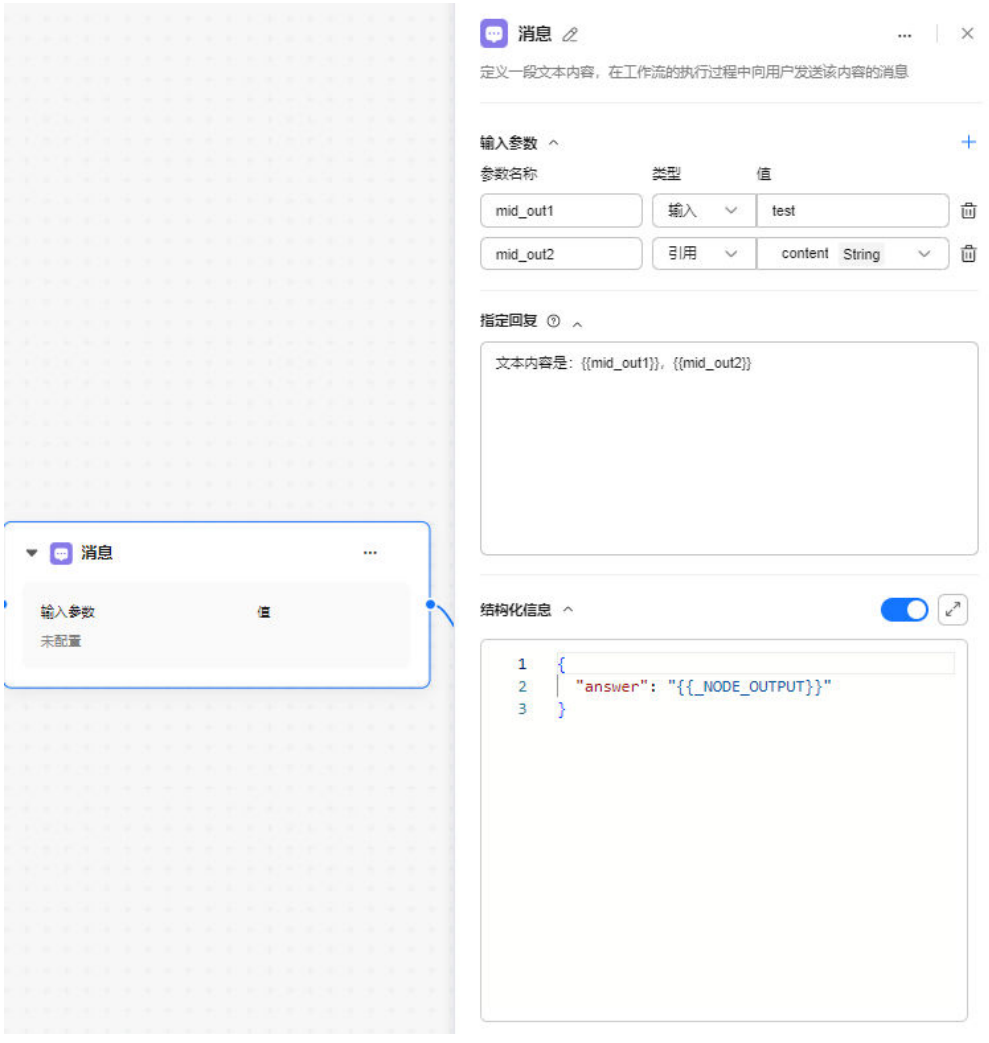
- 单击图标，可修改消息节点名称，修改完成后单击名称旁边的进行保存。
- 单击图标，可重命名消息节点名称，复制一个消息节点或删除消息节点。

表 8-23 消息节点配置说明

配置类型	参数名称	参数说明
参数配置	输入参数	当单击+图标时，可新增输入参数。 • 参数名称：只允许输入字母、数字、下划线，且不能以数字开头。 • 类型、值：支持“引用”和“输入”两种类型。 – 引用：支持用户选择工作流中已包含的前置节点的输出变量值和全局配置中的记忆变量。 – 输入：支持用户自定义取值。

配置类型	参数名称	参数说明
指定回复	指定回复	可撰写指定的回复信息，并支持以{{参数名称}}的形式插入变量。回复信息将在工作流执行到该节点时发送给用户。 可在“结构化信息”中使用{{_NODE_OUTPUT}}引用。
结构化信息	结构化信息	功能开启时，可使用{{_NODE_OUTPUT}}引用“指定回复”中的信息实现结构化输出。

图 8-65 消息节点配置示例



步骤6 节点配置完成后，单击“确定”。

步骤7 连接消息节点和其他节点。

----结束

8.11.2 输入

输入节点提供工作流运行过程中的信息输入。

前提条件

已完成工作流搭建，如果未搭建工作流，请参考**搭建工作流**。

节点说明

在比较复杂的工作流场景中，某些节点的执行往往需要额外的用户输入。如果前置节点中没有获取到这些信息，你可以添加一个输入节点来主动收集信息。工作流执行到输入节点时会暂时中断，直到此节点收集到必要的用户输入。

配置输入节点

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏选择“开发中心 > 应用管理 > 工作流应用”，单击您创建的工作流。

步骤3 单击“添加节点”并选择“输入”节点。

步骤4 通过单击该节点打开节点配置页面。

步骤5 参照**表8-24**，完成变量输入节点的配置。

表 8-24 输入节点配置说明

配置类型	参数名称	参数说明	配置示例
参数配置	输入参数	支持配置一个或多个输入参数，且输入参数可被后置节点引用。 当单击+图标时，可新增输入参数。 • 参数名称：只允许输入字母、数字、下划线，且不能以数字开头。 • 参数类型：可选String、Integer、Number、Boolean、Object、Array、File类型。 • 描述：对于该输入参数的描述。 • 必填：表示添加的参数是否为必须。	当配置一个或多个输入参数后，工作流运行到该节点时会暂时中断，直到用户填写所有输入参数，且每个输入参数均可被后置节点引用。 例如插件节点需要一个输入参数：“city”，可通过在该插件节点的前置节点中配置一个输入节点用于填写“city”的具体值，再在插件节点的“city”参数处引用输入节点相应参数即可。

图 8-66 输入节点配置示例

步骤6 节点配置完成后，单击“确定”。

步骤7 连接输入节点和其他节点。

----结束

8.11.3 提问器

提问器节点为开发者提供了收集用户问题所需信息的功能。

前提条件

已完成工作流搭建，如果未搭建工作流，请参考**搭建工作流**。

节点说明

该节点会循环执行，直到收集到所有必需的信息为止。

配置提问器节点

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏选择“开发中心 > 应用管理 > 工作流应用”，单击您创建的工作流。

步骤3 单击“添加节点”并选择“提问器”节点。

步骤4 通过单击该节点打开节点配置页面。

步骤5 参照**表8-25**，完成提问器节点的配置。

说明

- 单击图标，可修改提问器节点名称，修改完成后单击名称旁边的进行保存。
- 单击图标，可重命名提问器节点名称，复制一个提问器节点或删除提问器节点。
- 单击图标，可对提问器节点进行测试。

表 8-25 提问器节点配置说明

配置类型	参数名称	参数说明
模式偏好	-	效果优先：效果优先模式下，会开启时间增强和反思功能，提参成功率更高，时延会增加。 ● 时间增强：需要提取时间时，可以将自然语言的时间日期提取为YYYY-MM-DD HH:MM:SS标准格式的时间日期，比如：明天 12:30，提取为 2024-04-13 13:30:00。 ● 反思功能：数据提取之后，会让模型再判断是否提取正确，格式是否满足要求，不满足会尽量做一些修正。比如：期望提取电话号码，用户输入：我不记得电话号码，提取出：189******，反思后会认为提取不正确，会继续追问。 速度优先：速度优先模式下时延最低，提参成功率可能无法保障，速度优先模式下不开启时间增强和反思功能。
模型配置	模型选择	选择执行此节点的模型，支持设置模型在此节点中的生成多样性等参数配置，使模型效果更符合你的预期。 提问器模型用于接收用户自然语言，提取用户配置的输出参数，效果优先时还用于提取结果反思和纠正。
	温度	当单击图标时，可进行该参数设置。 用于控制生成结果的随机性。调高温度，会使得模型的输出更具多样性和创新性；降低温度，会使输出内容更加遵循指令要求，但同时也会减少模型输出的多样性。
	核采样	当单击图标时，可进行该参数设置。 模型在输出时会从概率最高的词汇开始选择，直到这些词汇的总概率累积达到核采样值，核采样值可以限制模型选择这些高概率的词汇，从而控制输出内容的多样性。建议不要与温度同时调整。
	最大回复长度	当单击图标时，可进行该参数设置。 控制模型输出的Tokens长度上限。通常100Tokens约等于150个中文汉字。

配置类型	参数名称	参数说明
参数配置	输入参数	设置需要添加到问题中的参数，参数值可以引用前置节点的输出参数，或设置为固定文本内容，可引用多个参数。 当单击+图标时，可新增输入参数。 ● 参数名称：只允许输入字母、数字、下划线，且不能以数字开头。 示例：输入参数为“pre_assigned_meeting_rooms”，希望用户在指定的多个选项中选出一个，后续问题配置为“有以下几个会议室供您选择：{{pre_assigned_meeting_rooms}}，请选择您想预订的会议室”。 ● 类型、值：支持“引用”和“输入”两种类型。 – 引用：用户可以选择工作流中该节点的前置节点的输出变量及全局配置的记忆变量作为取值。 – 输入：用户直接输入变量取值文本。

配置类型	参数名称	参数说明
	输出参数	该参数用于解析大模型节点的输出，并提供给后续节点的输出参数引用，支持多参数提取。 当单击+图标时，可新增输出参数。 ● 默认输出参数。 – USER_RESPONSE： 用户原始输出。 – STATUS：提取状态。 0 正常成功提取，用户无确认。 10 正常成功提取，用户已确认。 100 取到部分参数，用户主动中断，已提参数报错，未提参数按格式置空。 101 取到部分参数，循环超轮次，已提参数报错，未提参数按格式置空。 201 大模型调用异常。 202 反思模块有错误。 ● 参数提取：开启后，可增加需要提取的参数，参数可配置属性如下： – 参数名称：只允许输入字母、数字、下划线、短横线。 – 中文名称：不允许为空。 – 类型：输出参数的类型，可选String、Integer、Number、Boolean。 – 默认值：输出参数的默认值，大模型提取不到参数，并达到最大回复轮数时使用默认值。 – 描述：对于该输出参数的描述。 – 校验：开启后可自定义参数校验规则对输出参数规范性进行校验。规则包括参数名称、校验类型及校验规则。 – 提取：开启后该参数必须提取到或配置了默认值则使用默认值，关闭则该参数允许为空。 ● 引用插件：参数提取可能是给插件使用，通过引用插件，可导入插件的参数信息及校验信息，提升配置效率。
问题配置	问题	该参数将在对话框中原样呈现给用户。如未配置此处，将由大模型根据输出参数描述，自动生成包含所有问题关键词的一个问题。 如：请问你的名字是什么? 可通过Jinja语法在问题中使用输入参数 如：请问你是哪个班级的，可选班级有{{classes}}（ classes先在input参数配置好 ）。
	最大回复轮数	该参数用于设置与模型的最大交互次数，超过最大回复轮数还没有提取到参数则跳出提问器。

配置类型	参数名称	参数说明
高级配置	允许用户退出交互	开启后，如果用户在与提问器的对话交互中，表达“中止对话”类的意图，系统会自动结束当前提问，并跳转至结束节点。
	输出参数确认	开启后，如果用户希望提问器参数提取完毕后进行用户确认，则开启此功能。
	提取约束	提供大模型额外的约束信息，用于更准确的提取参数，例如指定被提取参数的格式要求。 ● 当单击“保存到模板”，填写“模板名称”、选择“行业”和“标签”后，可将提示词创建成模板并保存到我的提示词。 ● 当单击  图标时，可对系统提示词进行智能优化。 ● 当单击  图标时，系统会弹出“提示词广场”窗口，可在“预制提示词”或“我的提示词”页签中进行选择。 举例：用户希望提取电话号码tel_number，约束里面可以写tel_number必须是11位数字。
	追问模式	追问模式用来配置，在多次交互过程中，系统返回的参数追问语句生成模式。 ● 默认：使用默认内置追问模板生成追问语句，每次追问内容相同。 ● 智能追问：使用大模型生成语义良好，表达丰富的追问语句，每次追问内容丰富多变。 ● 自定义追问：按照自定义模板配置生成追问语句。‘{unextracted_cn_field_names}’不可修改或删除。每次追问内容相同。 例如，要提取名字和年龄参数 ● 默认：请您提供名字和年龄相关的信息。 ● 智能追问：您好，需要获取您的名字和年龄 。（模型生成，内容不固定。） ● 自定义追问：（自定义追问模板配置为：请问你的如下信息：{unextracted_cn_field_names}）。 请问你的如下信息：名字和年龄。
	追问显示枚举值	开启后，如果参数设置了枚举值校验，将在提问器的追问中，提示设定的参数可选枚举值。
	示例配置	给大模型一段预期的参数提取示例，增强大模型对参数提取场景的理解。模板：输入query：我要坐飞机去呼和浩特学习培训提取参数：{"location":"呼和浩特","traveltool":"飞机"}

图 8-67 提问器节点配置示例

步骤6 节点配置完成后，单击“确定”。

步骤7 连接提问器节点和其他节点。

----结束

8.11.4 问答

问答节点可提供中间过程的向用户提问的能力。

前提条件

已完成工作流搭建，如果未搭建工作流，请参考**搭建工作流**。

节点说明

通过指定问题回复用户后，接收到的用户回答作为参数向后续流转，创建步骤请详见**创建知识库**。

配置问答节点

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏选择“开发中心 > 应用管理 > 工作流应用”，单击您创建的工作流。

步骤3 单击“添加节点”并选择“问答”节点。

步骤4 通过单击该节点打开节点配置页面。

步骤5 参照**表8-26**，完成问答节点的配置。

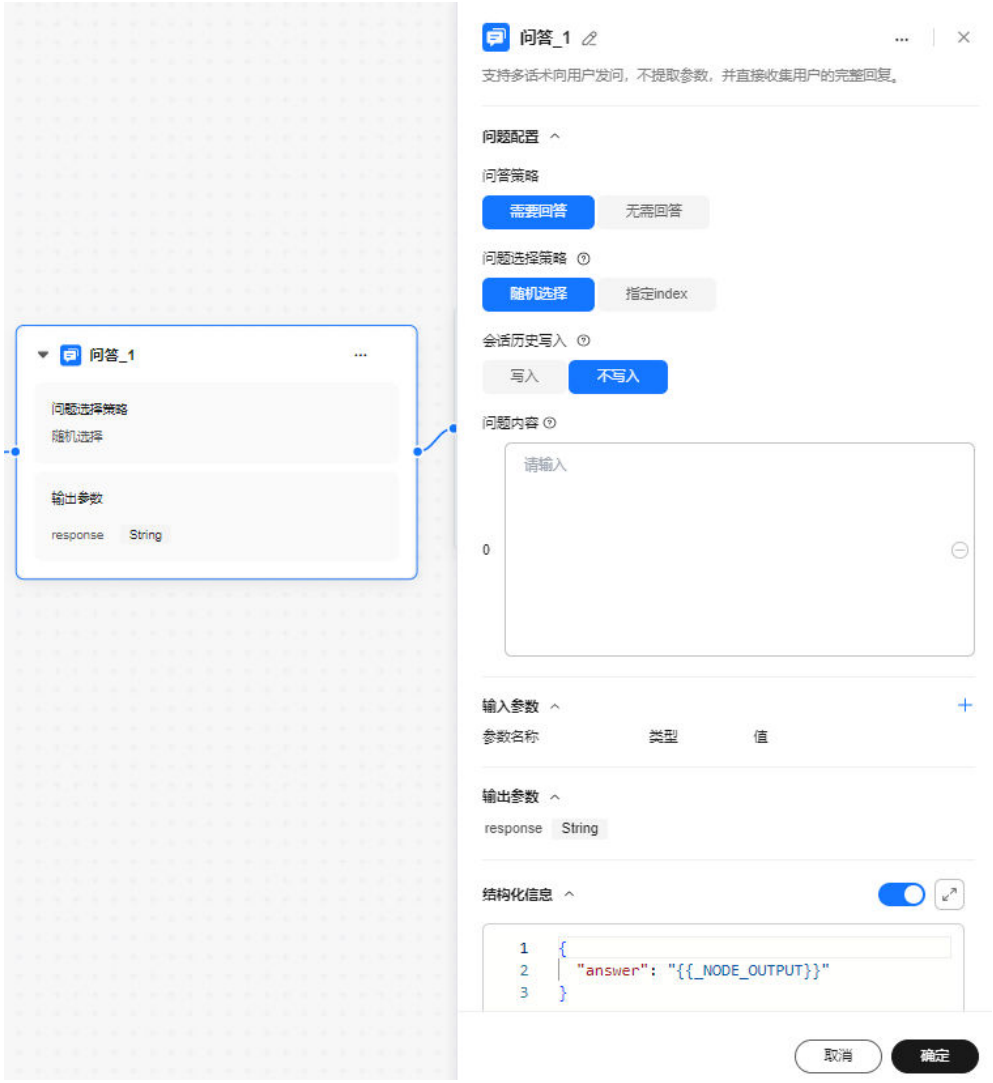
说明

- 单击图标，可修改问答节点名称，修改完成后单击名称旁边的进行保存。
- 单击图标，可重命名问答节点名称，复制一个问答节点或删除问答节点。
- 单击图标，可对问答节点进行测试。

表 8-26 问答节点配置说明

配置类型	参数名称	参数说明
模式配置	问答策略	可选择“需要回答”或“无需回答”，如“需要回答”，执行到问答节点后，会在需要回答节点处停止等待下一轮输入，并将下一轮输入作为问答节点的输出参数。如“无需回答”，执行到问答节点后，会将配置好的问题作为问答节点的输出参数。
	问题选择策略	问题选择策略支持在已配置的问题中“随机选择”或“指定index”进行回复。
	会话历史写入	可选择“写入”或“不写入”。 ● 写入：输出写入会话历史。一般用于记录用户与机器人的对话历史，以便模型能更好理解用户意图。 ● 不写入：输出不写入会话历史。一般机器与机器的交互历史无需记录，提升历史对话信息准确性，减少历史长度。
问题内容	问题内容	问题内容为问答节点问题配置，支持配置多个。 可在“结构化信息”中使用{{_NODE_OUTPUT}}引用。
输入参数	输入参数	● 参数名称：固定为index，仅在“指定index模式生效”。 ● 类型、值：支持“引用”和“输入”两种类型。 – 引用：支持用户选择工作流中已包含的前置节点的输出变量值和全局配置中的记忆变量。 – 输入：支持用户自定义取值。
输出参数	输出参数	该参数用于问答节点的输出，并提供给后续节点的输出参数引用，支持多参数提取。
结构化信息	结构化信息	功能开启时，可使用{{_NODE_OUTPUT}}引用“问题内容”中的信息实现结构化输出。

图 8-68 问答节点配置示例



步骤6 节点配置完成后，单击“确定”。

步骤7 连接问答节点和其他节点。

----结束

8.11.5 对象提取

对象提取节点用于提取指定对象中的参数，并支持配置子工作流进行参数的校验与校准，以及触发用户交互流程。

当工作流包含众多节点且交互复杂，导致难以理解和维护时，用户可以利用该节点来提取指定对象中的参数。该节点不仅支持配置子工作流以进行参数的校验与校准，还能发起用户交互，从而简化复杂工作流的管理和维护。

前提条件

已完成工作流搭建，如果未搭建工作流，请参考**搭建工作流**。

节点说明

通过使用“对象提取节点”，用户可以更高效地管理和维护工作流，提高工作效率，同时减少配置错误的可能性。

配置对象提取

步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏选择“开发中心 > 应用管理 > 工作流应用”，单击您创建的工作流。

步骤3 单击“添加节点”并选择“对象提取”节点。

步骤4 通过单击该节点打开节点配置页面。

步骤5 参照表8-27，完成对象提取节点的配置。

说明

- 单击图标，可修改对象提取节点名称，修改完成后单击名称旁边的进行保存。
- 单击图标，可重命名对象提取节点名称，复制对象提取节点或删除对象提取节点。
- 单击图标，可对对象提取节点进行测试。

表 8-27 对象提取节点配置说明

配置类型	参数名称	参数说明
输入参数	输入参数	支持配置一个或多个输入参数，且输入参数可被后置节点引用。 当单击图标时，可新增输入参数。 • 参数名称：只允许输入字母、数字、下划线，且不能以数字开头。 • 类型、值：支持“引用”和“输入”两种类型。 – 引用：用户可以选择工作流中该节点的前置节点的输出变量及全局配置的记忆变量作为取值。 – 输入：支持用户自定义取值。

配置类型	参数名称	参数说明
上下文变量	上下文变量	上下文变量可被子工作流引用，并在整个会话内生效，作为本节点的输出参数。 当单击+图标时，可新增上下文变量。 • 变量名称：只允许输入字母、数字、下划线，且不能以数字开头。 • 变量类型：支持String、Integer、Number、Boolean、Object、Array<String>、Array<Number>、Array<Integer>、Array<Boolean>、Array<Object>多种类型的变量。 新增上下文变量： – 第一层级：变量类型默认为**Object**，支持修改为其他类型，修改后的类型不支持添加分支；其中**Object**类型参数最多支持5层嵌套。 – 分支：变量类型默认为**String**，支持修改为其他类型，不支持添加分支。 • 类型：支持“引用”和“输入”两种类型。 • 描述（可选）：对于该上下文变量的描述。 • 单击 可删除已添加的上下文变量。 单击 图标，可通过对象模板添加上下文变量，对象模版创建详见**对象管理**页面。
领域对象	领域对象	领域对象可被子工作流引用，并在整个会话内生效。 当单击+图标时，可新增领域对象。该参支持添加多个领域对象。 • 变量名称：只允许输入字母、数字、下划线，且不能以数字开头。 • 变量类型：支持String、Integer、Number、Boolean、Object、Array<String>、Array<Number>、Array<Integer>、Array<Boolean>、Array<Object>。 新增领域对象： – 第一层级：变量类型默认为**Object**，支持修改为其他类型，修改后的类型不支持添加分支；其中**Object**类型参数最多支持5层嵌套。 – 分支：变量类型默认为**String**，支持修改为其他类型，不支持添加分支。 • 描述（可选）：对于该领域对象的描述。 • 单击 可删除已添加的领域对象。 单击 图标，可通过对象模板添加领域对象，对象模版创建详见**对象管理**页面。

配置类型	参数名称	参数说明
	对象处理流（可选）	模型在完成对象提取后，按定义顺序依次执行多个对象处理流，且执行时机晚于扩展工作流“模型提参后”。 当单击+图标时，可新增对象处理流。一个领域对象支持添加多个工作流，支持根据业务执行顺序拖动工作流。 已添加的工作流如果有更新，支持在对象提取节点中进行升级并查看配置。 ● 添加工作流后，单击可配置子工作流，设置允许用户配置工作流节点的进入条件。 – 输入参数：调用子工作流开始节点的参数，用户不支持选择。 参数名称：为子工作流开始节点的参数名称，不可编辑。 类型与子工作流的开始节点类型保持一致，值支持用户自定义取值；同时系统支持自动填入相同类型和名称的参数值。 – 输出参数：工作流节点的输出结构取决于子工作流定义的输出结构，不支持自定义设置。 – 上下文变量：变量名称、类型默认全部读取子工作流中的全局配置中的记忆变量参数，不支持修改和删除。上下文变量值支持用户自定义取值，也可不填参数值，同时系统支持自动填入相同类型和名称的参数值。 – 进入条件（可选）：进入本工作流条件，若不配置，默认进入。 由**[判断参数 比较条件 比较参数]**组成一个条件表达式。 判断参数：条件表达式左边部分，需要选择来自前序节点的输出参数。 比较条件：条件表达式中间部分，当前支持的比较条件有：长度小于、长度小于等于、等于、不等于、包含、不包含、为空、不为空。 比较参数：条件表达式右边部分，支持“引用”和“输入”两种类型。 ■ 引用：支持用户选择工作流中已包含的前置节点输出变量值及全局配置中的记忆变量。 ■ 输入：支持用户自定义取值。 添加条件：单击+，在当前条件分支中添加多个条件表达式，多个条件表达式之间通过“且”或“或”来连接。 单击“且”或“或”，可以切换该分支表达式的运算逻辑。 – 备注（可选）：描述子工作流节点功能。

配置类型	参数名称		参数说明
			• 单击 进入工作流版本预览界面，可预览子工作流、节点配置及版本信息。 • 单击 可删除已添加的子工作流。
模型提参配置	选择配置	选择模型	选择已配置的大语言模型。
		温度	当单击图标时，可进行该参数设置，推荐使用默认值。 用于控制生成结果的随机性。调高温度，会使得模型的输出更具多样性和创新性；降低温度，会使输出内容更加遵循指令要求，但同时也会减少模型输出的多样性。
		核采样	当单击图标时，可进行该参数设置，推荐使用默认值。 模型在输出时会从概率最高的词汇开始选择，直到这些词汇的总概率累积达到核采样值，核采样值可以限制模型选择这些高概率的词汇，从而控制输出内容的多样性。建议不要与温度同时调整。
		历史对话轮数	当单击图标时，可进行该参数设置，推荐使用默认值。 设置带入模型上下文的对话历史轮数。轮数越多，多轮对话的相关性越高，但消耗的Token也越多。
		最大回复长度	当单击图标时，可进行该参数设置，推荐使用默认值。 控制模型输出的Tokens长度上限。通常100Tokens约等于150个中文汉字。
		重复语句惩罚	当单击图标时，可进行该参数设置，推荐使用默认值。 当该值为正时，会阻止模型频繁使用相同的词汇和短语，从而增加输出内容的多样性。

配置类型	参数名称	参数说明
	参数提取配置	模型依据提示词和参数描述尝试进行对象提取，若未能提取，参数提取的变量名称为空。 当单击+图标时，可新增一条参数。一个模型支持添加多个参数。 ● 变量名称：与提取的参数名称保持一致，支持用户自定义填写。只允许输入字母、数字、下划线，且不能以数字开头。 ● 变量类型：与提取的参数名称的类型保持一致，支持切换其他类型。支持String、Integer、Number、Boolean、Object、Array<String>、Array<Number>、Array<Integer>、Array<Boolean>、Array<Object>。 ● 描述（可选）：描述变量功能。 ● 单击 可删除已添加的子工作流。
	提示词配置	配置输入给大模型的提示词，系统级提示词，用于指导模型按要求进行回复。支持使用{{variable}}格式引用当前节点输入参数中已定义好的参数。最终替换后的内容会传递给模型。 ● 当单击 图标时，可对提示词进行智能优化。 ● 当单击 图标时，系统会弹出“提示词广场”窗口，可在“预置提示词”或“我的提示词”页签中进行选择。

配置类型	参数名称	参数说明
扩展工作流（可选）	扩展工作流	当单击+图标时，可新增扩展工作流，支持添加多个工作流。已添加的工作流如果有更新，支持在对象提取节点中进行升级并查看配置。 • 添加工作流后，单击“执行时机”下拉框，配置子工作流的执行时机。 – 首次进入：进入节点后执行，且只执行一次。 – 模型提参后：大模型完成对象提取后执行。 – 退出条件判断前：退出条件判断前执行。 多个相同执行时机的扩展工作流按照定义先后顺序依次执行。 • 添加工作流后，单击可配置子工作流，设置允许用户配置工作流节点的进入条件。 – 输入参数：配置子工作流开始节点的输入参数，类型与子工作流的开始节点类型保持一致，值支持用户自定义取值，同时系统支持自动填入相同类型和名称的参数值。 – 输出参数：工作流节点的输出结构取决于子工作流定义的输出结构，不支持自定义设置。 – 上下文变量（可选）：支持将参数提取节点的领域对象、上下文变量传入子工作流，并可在子工作流内部通过变量赋值节点修改。同时需要先在子工作流全局变量中定义记忆变量，同时系统支持自动填入相同类型和名称的参数值。 – 进入条件（可选）：进入本工作流条件，若不配置，默认进入。 由**[判断参数 比较条件 比较参数]**组成一个条件表达式。 判断参数：条件表达式左边部分，需要选择来自前序节点的输出参数。 比较条件：条件表达式中间部分，当前支持的比较条件有：长度小于、长度小于等于、等于、不等于、包含、不包含、为空、不为空。 比较参数：条件表达式右边部分，支持“引用”和“输入”两种类型。 ▪ 引用：支持用户选择工作流中已包含的前置节点输出变量值及全局配置中的记忆变量。 ▪ 输入：支持用户自定义取值。 添加条件：单击+，在当前条件分支中添加多个条件表达式，多个条件表达式之间通过“且”或“或”来连接。 单击“且”或“或”，可以切换该分支表达式的运算逻辑。

配置类型	参数名称	参数说明
		– 备注（可选）：描述子工作流节点功能。 ● 单击 进入工作流版本预览界面，可预览子工作流、节点配置及版本信息。 ● 单击 可删除已添加的子工作流。
条件配置（可选）	模型提参条件	使用模型对象提取的条件，若不配置，默认使用。 ● 进入条件（可选）：进入本工作流条件，若不配置，默认进入。 由**[判断参数 比较条件 比较参数]**组成一个条件表达式。 判断参数：条件表达式左边部分，需要选择来自前序节点的输出参数。 比较条件：条件表达式中间部分，当前支持的比较条件有：长度小于、长度小于等于、等于、不等于、包含、不包含、为空、不为空。 比较参数：条件表达式右边部分，支持“引用”和“输入”两种类型。 – 引用：支持用户选择工作流中已包含的前置节点输出变量值及全局配置中的记忆变量。 – 输入：支持用户自定义取值。 添加条件：单击+，在当前条件分支中添加多个条件表达式，多个条件表达式之间通过“且”或“或”来连接。 单击“且”或“或”，可以切换该分支表达式的运算逻辑。

配置类型	参数名称	参数说明
	节点退出条件	节点退出条件：退出本节点的条件，若不配置，执行一轮后退出。 ● 进入条件（可选）：进入本工作流条件，若不配置，默认进入。 由**[判断参数 比较条件 比较参数]**组成一个条件表达式。 判断参数：条件表达式左边部分，需要选择来自前序节点的输出参数。 比较条件：条件表达式中间部分，当前支持的比较条件有：长度小于、长度小于等于、等于、不等于、包含、不包含、为空、不为空。 比较参数：条件表达式右边部分，支持“引用”和“输入”两种类型。 – 引用：支持用户选择工作流中已包含的前置节点输出变量值及全局配置中的记忆变量。 – 输入：支持用户自定义取值。 添加条件：单击+，在当前条件分支中添加多个条件表达式，多个条件表达式之间通过“且”或“或”来连接。 单击“且”或“或”，可以切换该分支表达式的运算逻辑。
	节点异常条件	进入异常分支条件，若不配置，无异常分支。添加节点异常条件后需要在对象提取参数后设置异常处理分支。 ● 进入条件（可选）：进入本工作流条件，若不配置，默认进入。 由**[判断参数 比较条件 比较参数]**组成一个条件表达式。 判断参数：条件表达式左边部分，需要选择来自前序节点的输出参数。 比较条件：条件表达式中间部分，当前支持的比较条件有：长度小于、长度小于等于、等于、不等于、包含、不包含、为空、不为空。 比较参数：条件表达式右边部分，支持“引用”和“输入”两种类型。 – 引用：支持用户选择工作流中已包含的前置节点输出变量值及全局配置中的记忆变量。 – 输入：支持用户自定义取值。 添加条件：单击+，在当前条件分支中添加多个条件表达式，多个条件表达式之间通过“且”或“或”来连接。 单击“且”或“或”，可以切换该分支表达式的运算逻辑。

配置类型	参数名称	参数说明
备注（可选）	备注	用于描述对象提取节点实现的功能，可在画布中的节点内展示，支持多种数据类型，长度0~512个字。

图 8-69 对象提取节点配置示例

步骤6 节点配置完成后，单击“确定”。

步骤7 连接对象提取节点和其他节点。

----结束

8.11.6 异常

异常节点功能允许用户根据业务需求灵活设置和抛出详细的异常信息。用户可以自定义异常信息，在遇到问题时提供清晰的反馈信息。

前提条件

已完成工作流搭建，如果未搭建工作流，请参考**搭建工作流**。

节点说明

用户可以根据具体的业务需求，直接在消息体中输入异常信息，或者选择已创建的消息分类中的异常消息模板。

配置异常节点

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏选择“开发中心 > 应用管理 > 工作流应用”，单击您创建的工作流。

步骤3 单击“添加节点”并选择“异常”节点。

步骤4 单击画布中已添加的“异常”节点，可以根据以下方式在异常码抛出中输入异常信息。单击右侧的 可以扩展内容输入框，方便输入更多详细信息。

- 直接输入：直接在消息体中输入异常信息，仅支持JSON格式。
- 插入创建好的消息模板：单击 ，在展开的“插入信息模板”弹框中选择创建好的异常信息，单击“确定”。

步骤5 节点配置完成后，单击“确定”。

步骤6 连接异常节点与其他节点。使用鼠标拖拽连线，将异常节点与前面和后面的节点连接起来。

说明

- 单击 图标，可修改异常节点名称，修改完成后单击名称旁边的 进行保存。
- 单击 图标，可重命名异常节点名称，复制一个异常节点或删除异常节点。

----结束

8.12 数据&知识节点

8.12.1 变量赋值

变量赋值节点，将特定的值赋给变量，可以实现数据的动态更新和传递，使工作流能够根据实时数据做出相应的处理和决策。

前提条件

已完成工作流搭建，如果未搭建工作流，请参考**搭建工作流**。

节点说明

变量赋值节点支持在循环节点内部使用，通过变量赋值节点，将特定的值赋给中间变量，可以实现循环过程中数据的动态更新和传递。

配置变量赋值节点

步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏选择“开发中心 > 应用管理 > 工作流应用”，单击您创建的工作流。

步骤3 单击“添加节点”并选择“变量赋值”节点。

步骤4 通过单击该节点打开节点配置页面。

步骤5 参照表8-28，完成变量赋值节点的配置。

说明

- 单击图标，可修改变量赋值节点名称，修改完成后单击名称旁边的进行保存。
- 单击图标，可重命名变量赋值节点名称，复制一个变量赋值节点或删除变量赋值节点。

表 8-28 变量赋值节点配置说明

配置类型	参数名称	参数说明
循环节点外变量赋值节点配置	变量赋值	变量赋值节点变量名称仅支持全局配置中记忆变量引用，值可支持引用或者输入两种。 类型、值：支持“引用”和“输入”两种类型。 • 引用：支持用户选择工作流中已包含的前置节点的输出变量值以及全局配置的中的记忆变量。 • 输入：支持用户自定义取值。
循环节点中变量赋值配置	变量赋值	变量赋值节点支持在循环体内部引用，只支持更改循环体中间变量的值，被赋值变量仅支持选择中间变量，值可支持引用或输入两种。适用于循环过程中动态更新中间变量，自定义循环逻辑中进行参数传递的场景。 类型、值：支持“引用”和“输入”两种类型。 • 引用：中间变量的值需要引用上游节点输出时勾选此项，支持用户选择工作流中已包含的前置节点的输出变量值以及循环体内置变量，包括index、item（数组循环）以及中间变量，适用于循环过程中修改中间变量的值为变量的场景。 • 输入：支持用户自定义取值，适用于循环过程中修改中间变量的值为固定值场景。

图 8-70 变量赋值节点配置示例

图 8-71 变量赋值节点在循环节点中配置示例

步骤6 节点配置完成后，单击“确定”。

步骤7 连接变量赋值节点和其他节点。

----结束

8.12.2 变量聚合

变量聚合节点能够对多个分支的输出进行聚合处理，方便后置节点统一配置。

前提条件

已完成工作流搭建，如果未搭建工作流，请参考**搭建工作流**。

节点说明

如果工作流中设计了多个分支，往往需要一个节点来汇总所有分支的输出结果。在这种场景下，你可以使用变量聚合节点聚合多路分支的输出变量，变量聚合节点会读取多路分支中第一个不为空的值，供流程下游的节点使用和操作，不用额外处理未运行分支的输出结果，简化了数据流的管理。

配置变量聚合节点

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏选择“开发中心 > 应用管理 > 工作流应用”，单击您创建的工作流。

步骤3 单击“添加节点”并选择“变量聚合”节点。

步骤4 通过单击该节点打开节点配置页面。

步骤5 参照表8-29，完成变量聚合节点的配置。

说明

- 单击图标，可修改变量聚合节点名称，修改完成后单击名称旁边的进行保存。
- 单击图标，可重命名变量聚合节点名称，复制一个变量聚合节点或删除变量聚合节点。

表 8-29 变量聚合节点配置说明

配置类型	参数名称	参数说明
参数配置	输出参数	• 参数名称：固定为Group1，如果有多个分组则根据分组数量递增为Group2、Group3等。 • 参数类型：取决于对应聚合分组的变量数据类型。
聚合策略	-	通过指定策略对每个分组中的所有变量进行聚合处理，同一组内的变量实施相对应的聚合策略。 目前聚合策略仅支持设置为“返回每个分组中第一个非空值”，支持拖动变量、调整变量位置。例如组内按顺序设置三个变量output1、output2和output3，将其聚合为一个变量Group1，如果output1不为空，则用output1的值为Group1赋值；如果output1为空，则取output2的值，依次类推。
聚合分组	-	默认只有一个分组Group1，对应一个输出变量Group1。分组中所有变量类型相同。如果需要输出多个变量，可以添加多个分组，依次递增为Group2、Group3等。
聚合变量	-	在聚合分组中选择需要聚合的变量，每个分组只能聚合一种数据类型的变量。例如将多个String类型的变量聚合为一个String变量、将多个Integer类型的变量聚合为一个Integer变量。

图 8-72 变量聚合节点配置示例

步骤6 节点配置完成后，单击“确定”。

步骤7 连接变量聚合节点和其他节点。

----结束

8.12.3 知识检索

知识检索是一种从大量信息源中找到所需知识或信息的过程。

前提条件

已完成工作流搭建，如果未搭建工作流，请参考**搭建工作流**。

节点说明

知识检索节点可以基于用户的输入，从指定知识库内召回匹配的信息，并将匹配结果以列表形式返回。该节点支持选择用户创建的知识库，创建步骤请详见**创建知识库**。

配置知识检索节点

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏选择“开发中心 > 应用管理 > 工作流应用”，单击您创建的工作流。

步骤3 单击“添加节点”并选择“知识检索”节点。

步骤4 通过单击该节点打开节点配置页面。

步骤5 参照**表8-30**完成大模型节点的配置。

说明

- 单击图标，可修改知识检索节点名称，修改完成后单击名称旁边的进行保存。
- 单击图标，可重命名知识检索节点名称，复制一个知识检索节点或删除知识检索节点。

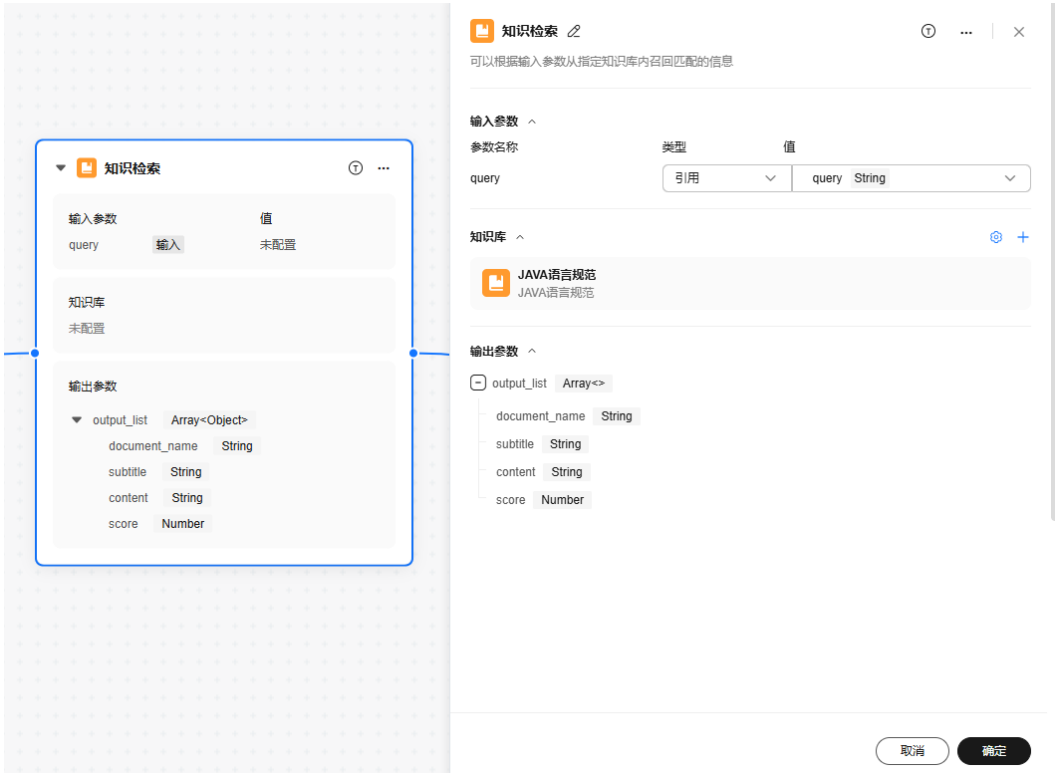
表 8-30 知识检索节点配置说明

配置类型	参数名称	参数说明
参数配置	输入参数	• 参数名称：输入参数固定只有1个，参数名称为query且不可修改，类型是字符串，表示待知识检索的问题。 • 类型、值：支持“引用”和“输入”两种类型。 – 引用：支持用户选择工作流中已包含的前置节点的输出变量值以及全局配置中的记忆变量，限制String类型，适用于需要从前置节点输出中获取知识检索问题的场景。 – 输入：支持用户自定义输入问题，适用于知识检索问题固定的场景。

配置类型	参数名称	参数说明
	知识库	支持选择用户所创建的知识库。 **说明** ● Versatile基础版（限时免费）：最多支持关联1个知识库。 ● Versatile企业版：最多支持关联3个知识库。
	检索策略	文档检索的方式，有如下几种检索策略： ● **语义检索**：使用向量检索技术检索，对文档及结构化数据中知识进行检索，召回与用户意图相关性高的切片内容，推荐在需要结合上下文相关性、并对用户意图理解场景中使用。 ● **关键词检索**：使用倒排检索技术，对文档及结构化数据中知识进行检索，召回与Query关键词匹配度高的切片内容，推荐在需要用户提问关键词匹配度高的场景中使用。 ● **混合检索**：使用向量检索和关键词检索两种策略混合检索知识库，推荐在需要兼顾用户意图理解及关键词匹配度场景中使用。
	相关度阈值	得分低于相关度阈值的搜索结果会被过滤，可以参考知识库命中测试的相关度分值调整该阈值。 取值范围为0~1。
	topk召回数量	从知识库中召回的最大切片数量，如topk召回数量为5，则得分不在前5的切片将被过滤。 取值范围为1~50。
	FAQ直出阈值	FAQ检索超过阈值的结果将直接返回，不再进行文档检索。如果没有超过阈值的结果，将进行文档检索。 取值范围为0~1。 启用FAQ功能后，系统将优先检索FAQ数据。若未命中结果，则会继续查询切片内容，可能会带来一定的性能开销。当FAQ检索结果超过预设阈值时，将直接提交给大模型进行总结，不再进行文档检索。若未超过阈值，则将继续进行文档检索。
	查看图片	开启后此功能后，当知识库支持图片检索时，可查看检索结果中的图片信息。

配置类型	参数名称	参数说明
输出参数	-	知识检索节点的输出是一个对象数组，参数名是output_list，表示所有满足检索要求的知识切片。数组中对象有四个属性： • document_name，知识切片所在的知识文档名称。 • subtitle，知识切片子标题。 • content，知识切片的内容。 • score，知识切片的匹配度得分，output_list中的元素按照得分由高到低排序。 后续节点引用该输出参数，可以引用output_list，此时将获取全量的检索结果，包括文档名、切片子标题、切片内容和分数。也可以直接引用切片的属性，比如content，此时将获取output_list中第一条记录的切片内容。

图 8-73 知识检索节点配置示例



步骤6 节点配置完成后，单击“确定”。

步骤7 连接知识检索节点和其他节点。

----结束

8.13 数据库

8.13.1 数据查询

数据查询节点用于查询数据库数据，用户可配置查询条件和查询方式。

前提条件

已完成工作流搭建，如果未搭建工作流，请参考**搭建工作流**。

节点说明

数据查询节点主要用于从数据库中提取数据。它支持连接配置、查询构建（包括SQL查询和参数化查询）、数据预览。

配置数据查询节点

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 单击左侧导航栏“开发中心 > 应用管理 > 工作流应用”，单击您创建的工作流。

步骤3 单击“添加节点”并选择“数据查询”节点，将所需数据查询添加至画布中。

步骤4 连接数据查询节点和其他节点。

步骤5 在工作流画布中单击“数据查询”节点，打开配置页面。

图 8-74 数据查询节点配置界面

步骤6 单击画布中已添加的“数据查询”节点，参照**表8-31**，完成“数据查询”节点的配置。

说明

- 单击图标，可修改数据查询节点名称，修改完成后单击名称旁边的进行保存。
- 单击图标，可查看重命名数据查询节点名称，复制一个数据查询节点或删除数据查询节点。
- 单击图标，可对数据查询节点进行测试。

表 8-31 数据查询节点配置说明

配置类型	参数名称	参数说明
查询数据	数据库	• 单击选择数据库的下拉框可以选择已经接入的数据源。如果没有已接入的数据源，请先接入数据源。 • 单击选择数据表的下拉框可以选择数据库关联的数据表，每个查询数据节点仅支持操作一张数据表。
参数配置	输入参数	支持配置一个或多个输入参数，且输入参数可被后置节点引用。 当单击图标时，可新增输入参数。 • 参数名称：只允许输入字母、数字、下划线，且不能以数字开头。 • 类型、值：支持“引用”和“输入”两种类型。 – 引用：用户可以选择工作流中该节点的前置节点的输出变量作为取值。 – 输入：支持用户自定义取值。
	输出参数	输出符合查询条件及查询字段的数据库数据和字段。
查询条件配置	查询字段	• 用于配置查询后要输出的数据库字段，设置后系统会根据查询字段进行输出，若不设置则系统不会返回数据。 • 支持添加多个查询字段，查询字段类型与数据表一致。

配置类型	参数名称	参数说明
	查询条件	可选参数。 用于设置查询数据库字段的条件，指定待查询数据的范围。 ● 查询条件：支持等于（=）、不等于（!= 或 <>）、模糊匹配（LIKE）、模糊不匹配（NOT LIKE）、属于（IN）、不属于（NOT IN）、为空（IS NULL）、不为空（IS NOT NULL）等运算符。 ● 添加条件：单击+，在当前条件分支中添加多个条件表达式，多个条件表达式之间通过“且”或“或”来连接。单击“且”或“或”，可以切换该分支表达式的运算逻辑。 ● 比较参数：条件表达式右边部分，支持“引用”和“输入”两种类型。 – 引用：支持用户选择工作流中已包含的前置节点输出变量值及全局配置中的记忆变量。 – 输入：支持用户自定义取值。 **说明** ● 当只有一条查询条件时，逻辑关系不支持配置。 ● 当查询条件是时间类型时，支持配置时间。
	字段排序	可选参数。 用于设置查询结果的排序方式。 当单击+图标时，可添加查询字段。 当单击图标时，可删除查询字段。 如果添加了多个查询字段，支持拖动字段来调整字段排序的优先级。同时添加的字段支持切换“升序”和“降序”。
	返回行数	用于限制返回结果行数，默认值为10，最大可设置为1000。
	偏移量	用于设置查询行数起始值，默认从第0行开始查询。

图 8-75 数据查询节点配置示例

步骤7 节点配置完成后，单击“确定”。

----结束

8.14 配置管理

8.14.1 管理意图包

意图包是一种用于封装和管理特定意图的功能模块或配置的集合，通常包含一组预定义的意图、规则、逻辑或数据，同时帮助用户更高效地构建和维护智能体应用，旨在帮助系统或应用更高效地理解和响应用户的需求，提升智能体的灵活性和可扩展性。

本文主要围绕意图包的创建、维护、部署和优化展开，帮助用户高效地管理和使用意图包。

前提条件

已**购买Versatile智能体平台**。

约束与限制

- 单个sheet页的导入数据最大支持200条，大于200则不支持导入。
- 上传文件限xlsx格式，文件不大于5MB，支持下载模板。
- 批量录入时，多个意图样例名称须使用英文逗号分隔，最大支持输入1000字符。
- 导入文件中包含意图包名称与已配置意图包名称不能重复。

创建意图包

步骤1 在“Versatile”页面，打开“开发中心 > 配置管理”下的“意图管理”。

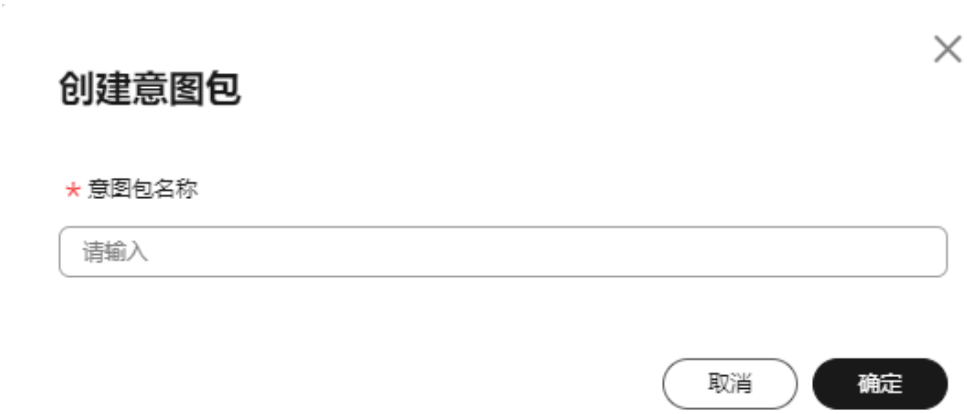
步骤2 单击“创建意图包”。

步骤3 在“创建意图包”界面输入意图包名称，单击“确定”。

说明

- 输入意图包名称不允许重复。
- 支持中英文、数字、下划线和中划线输入。

图 8-76 创建意图包



步骤4 单击“编辑”，进入意图包编辑页面。

图 8-77 编辑意图包

步骤5 单击 ，可在意图包中添加意图分类，分类信息包含名称和样例。

图 8-78 配置意图包

说明

意图管理页面支持对意图包重命名，添加意图、导入意图、批量录入、搜索、添加意图样例等。

- 单击 支持对意图包重命名。
- 单击 可添加意图。
- 单击 导入意图。
- 单击 批量录入 批量导入意图。
- 搜索框中支持输入搜索。

----结束

导入、导出意图包

Versatile支持导出和导入意图包。

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 进入“开发中心 > 配置管理 > 意图管理”页面。

步骤3 导出意图包。

1. 单击页面左上角“导出”。
2. 在“导出意图包”页面选择意图包前的复选框 ，单击“导出”。单击页面左上角“导出”。意图包将以xlsx格式的文件下载至本地。

   说明

   当导出多个意图包时，不同的意图包将在xlsx格式文件的不同sheet页呈现，并以意图包名称命名。

步骤4 导入意图包。

1. 单击页面左上角“导入”。
2. 在“导入”页面，单击“添加文件”选择需要导入xlsx的文件。
3. 单击“导入”，导入成功的意图包将在“意图管理”页面中展示。

说明

在xlsx格式文件的不同sheet页的意图包导入后，在意图管理页面将以单sheet页名称显示。

----结束

8.14.2 消息模板

消息模板是预设的标准化消息结构，您可以在当前页面创建消息模板，并在相关节点中引用这些模板。通过快速引用这些模板，减少重复性工作，从而提高开发效率。

前提条件

已**购买Versatile智能体平台**。

创建消息模板

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏选择“开发中心 > 配置管理 > 消息模板”。

步骤3 在“消息模板”页面，单击左上角的“新建”。

步骤4 在“创建消息模板”的弹框中输入消息的详细信息，具体请参考**表8-32**，配置消息信息。

表 8-32 创建消息参数说明

参数名称	参数说明
消息名称	消息的名称，用于标识和区分不同的消息。
消息分类	选择消息的分类，分为异常和自定义。 **说明** • 异常分类的消息模板可被工作流异常节点引用。 • 自定义节点暂未开放此功能。
可见范围	设置消息模板的可见范围。当前支持以下三种设置。 • 租户内：整个租户内的所有用户均可查看。 • 当前空间内：仅当前空间内的用户可以查看。 • 仅个人：仅创建者本人可见。
消息体	自定义消息的详细信息。单击右侧的 可以扩展内容输入框，方便输入更多详细信息。 **注意** • 消息体仅支持JSON格式。 • 系统提供了校验功能，确保消息的语法正确性。如果消息体存在语法问题，将无法成功创建。

步骤5 单击“确定”可保存设置的消息信息。创建完成后，您可以在异常节点中使用创建好的消息模板。

----结束

更多操作

消息模板建完成后，您可以执行如表8-33的操作。

表 8-33 相关操作

操作	说明
导入	支持批量导入功能。 1. 在“消息模板”页面，单击页面左上角的“导入”。 2. 在“导入信息模板”弹框中，单击“添加文件”选择需要导入的xlsx文件。 上传文件时，请注意以下限制条件： • 上传文件限xlsx格式，文件不超过5MB； • 单个sheet页的导入数据应不超过200条，超过数量不允许导入。 • 在“导入信息模板”弹框中，单击“下载模板”，可以下载消息模板。
导出	在“消息模板”页面，选择需要导出的消息，单击“导出”，导出的消息会以xlsx的格式保存在本地。
编辑	编辑消息模板的内容。在“消息模板”页面，找到需要编辑的消息，单击操作列中的“编辑”，在弹出的“编辑消息模板”对话框中进行内容编辑。
删除	删除消息模板。在“消息模板”页面，勾选需要删除的消息模板，单击“删除”，可以删除消息模板。 • 单个删除：在待删除的消息模板对应的“操作”列下，单击“删除”。 • 批量删除：勾选待删除消息模板，单击页面左上角的“删除”。

8.14.3 对象管理

对象管理介绍如何创建对象，编辑对象以及删除对象。您可以在当前页面创建对象模板，并在相关节点中引用这些模板。通过快速引用这些模板，减少重复性工作，从而提高开发效率。

新建对象

步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在Versatile的左侧导航栏，选择“开发中心 > 配置管理 > 对象管理”。

步骤3 单击页面左上角的“新建对象”。

步骤4 设置新建对象的基本配置信息，具体的参数说明如表1所示。

表 8-34 新建依赖包参数说明

参数	说明
对象名称	自定义对象管理的名称，支持中英文、数字、下划线、中划线，长度 2-64 字符，以中英文、数字开头。
对象变量	自定义对象变量包括变量名称，变量类型和描述（可选）。 • 变量名称：不允许为空，只允许输入字母、数字，下划线，且不能以数字开头。 • 变量类型：支持配置String、File、Integer、Number、Boolean、Object、Array多种类型的参数，其中Object类型参数最多支持5层嵌套。 • 描述（可选）：对变量的详细描述。

步骤5 单击“确定”。

创建完成后，您可以在对象提取节点中的“上下文变量”中，通过选择“对象模板”使用。

----结束

更多操作

新建对象创建完成后，您可以执行如表2的操作。

表 8-35 相关操作

操作	说明
搜索对象	输入部分对象名称后，单击 或按Enter键，符合条件的对象会显示在列表中。
编辑对象	在显示结果中，单击操作列的“编辑”，可以修改对象的名称、对象变量参数。
删除对象	在显示结果中，单击操作列的“删除”，然后在弹出的确认框中单击“确认”。

9 开发多智能体应用

9.1 多智能体应用介绍

在Versatile中创建的单智能体应用，能够处理基本任务，但在进行复杂任务处理时，需要编写详细且冗长的提示词，并添加各种插件、知识库、MCP服务等，增加了调试的复杂性。在单智能体应用中，任意一处改动都有可能影响到整体功能，导致用户在处理实际任务时，处理的结果可能与预期效果有较大出入。

为了解决这一问题，Versatile提供了多智能体应用。多智能体应用具有以下优势：

- 多智能体应用可以灵活应用各种工作流来完成用户任务，支持根据用户意图在不同的工作流之间跳转。
- 多智能体应用支持模型自动控制模式，进一步提升了任务处理的效率和准确性。

适用场景

适用于需要执行多任务处理的场景。例如，在金融领域，应用实现风险评估、投资组合优化、研报分析等多种复杂能力的智能投顾系统。

单智能体与多智能体功能与应用场景差异

单智能体：依赖模型，可以使用插件、工作流、知识库、MCP服务等工具，让模型自主规划，使用不同工具完成指定任务。

多智能体：可配置多个工作流，侧重根据客户意图在不同工作流中进行选择和跳转。

相关文档

- **单智能体应用与多智能体应用是否可以切换?**
- **工作流应用和多智能体应用有什么区别?**
- **在单智能体应用中使用工作流与在多智能体应用中使用工作流有什么区别?**

9.2 创建多智能体应用

多智能体应用允许用户将多个智能体应用或工作流应用进行组合，并按预设模式进行调度协同。多智能体应用的更多介绍请参考**多智能体应用介绍**。

本节介绍配置多智能体应用的流程。

前提条件

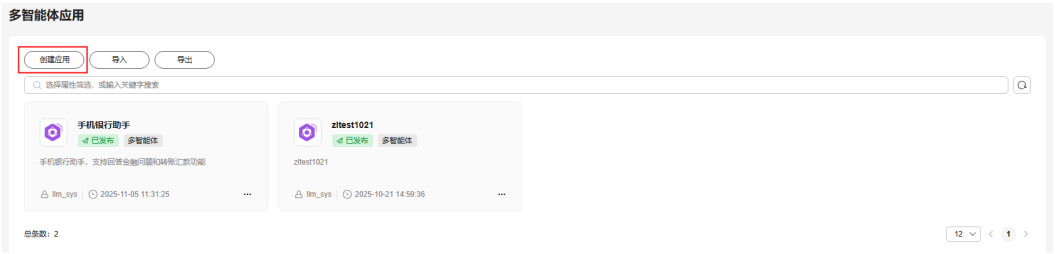
- 已**购买Versatile智能体平台**。
- 已**发布工作流**。
- 登录用户为空间所有者、空间管理员、开发工程师，详细信息请参考**管理团队空间成员**。

创建多智能体应用

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航，选择“开发中心 > 应用管理 > 多智能体应用”，单击“创建应用”。

图 9-1 “创建应用”



步骤3 在“创建应用”页面，配置应用基础信息，具体参数说明请参考**表9-1**。

图 9-2 创建应用

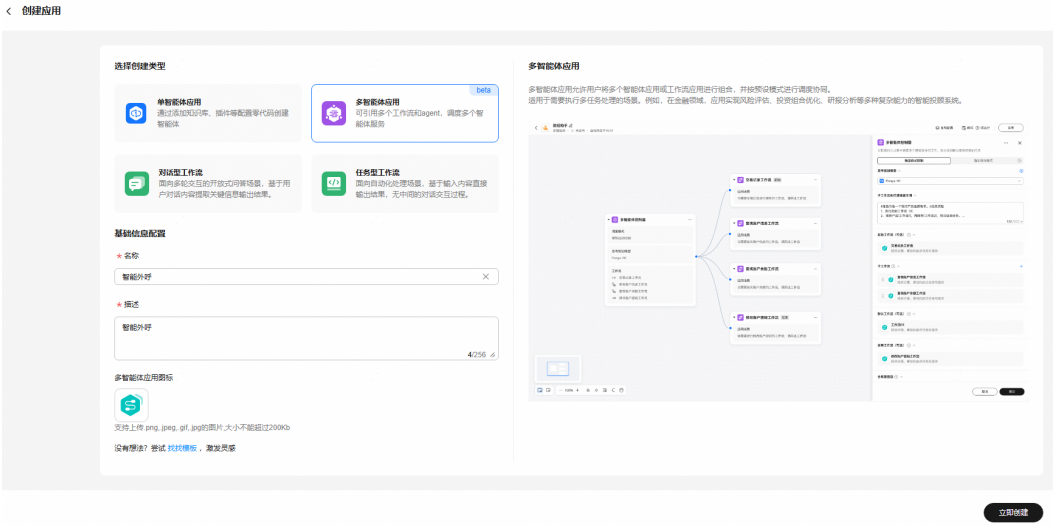


表 9-1 创建多智能体参数说明

参数	说明	示例
名称	多智能体应用的名称。由2~64个字符组成，包含中英文、数字、下划线、中划线、空格，不能以空格开头或结尾。	智能外呼

参数	说明	示例
描述	多智能体的描述信息。由1~256个字符组成。	智能外呼
多智能体应用图标	系统默认多智能体应用图标，用户也可以自定义图标。 1. 鼠标移动至系统默认图标上，单击鼠标左键。 2. 上传已准备好的应用图标。 支持jpg、jpeg、png、gif格式图片，且不大于200KB。	系统默认图标

步骤4 单击“立即创建”。

创建后，进入多智能体应用编辑页面，初始只有一个“多Agent控制器”节点。创建的多智能体应用显示在多智能体应用卡片列表中。

步骤5 设置全局配置。

在多智能体应用编辑页面右上方，单击“全局配置”。

全局配置可配置输入参数和全局变量，这些都可以给工作流的输入参数使用。

图 9-3 全局配置

表 9-2 全局配置参数说明

参数	说明
输入参数	传给工作流的输入参数，且值不可修改。 单击 ＋ ，添加输入参数。 ● 参数名称：由1~32字符组成，包含字母、数字、下划线，不能以数字开头。 ● 类型：类型支持String、Integer、Number、Boolean。默认值String。 ● 描述：由0~256个字符组成。 ● 必填：根据需求去掉勾选。默认勾选。 ● 🗑 ：删除输入参数。 ● ··· ：输入参数默认值。
全局变量	传给工作流的输入参数，工作流如果有相同名称和类型的输出参数会覆盖该值。 单击 ＋ ，添加全局变量。 ● 参数名称：由1~32字符组成，包含字母、数字、下划线，不能以数字开头。 ● 类型：类型支持String、Integer、Number、Boolean。默认值String。 ● 描述：由1~256个字符组成。 ● 🗑 ：删除输入参数。 ● ··· ：输入参数默认值。

步骤6 配置多Agent控制器。

在“多Agent控制器”卡片上，单击鼠标左键，在弹出页面配置参数信息，多Agent控制器参数说明请参考**表9-3**。

图 9-4 配置多 Agent 控制器

表 9-3 多 Agent 控制器参数说明

参数	说明	示例
模型配置	在下拉框中选择该多智能体应用工作使用的模型服务。已接入的模型服务详见**接入模型服务**。 **说明** 模型的标签展示顺序从左到右依次是用户自定义标签、**接入模型**时的“选择标签”、“模型类型”。 • 接入模型时的“选择标签”： – 联网：表示该大模型具备联网搜索能力。 – 思考：表示该大模型具备思维推理能力。 – 工具：表示该大模型支持应用调用外部工具，例如，MCP服务、插件、知识库等。 – default-import：表示该大模型是系统默认模型。 – 免费：表示该平台预置大模型可免费使用。 – 体验：表示该平台预置大模型可以体验，会话轮数最大为20次。 • “模型类型”包含： – 文本：表示该大模型是文本对话类型。 – 视觉：表示该大模型是图像理解类型。 – 嵌入：表示该大模型是文本向量化类型。 – 排序：表示该大模型是文本排序类型。 • 模型状态： – 未验证：表示该大模型未检验鉴权信息，不可使用。 – 成功：表示该大模型鉴权信息校验成功，可以使用。 – 失败：表示该大模型鉴权信息校验失败，不可使用。 在“模型配置”右侧，单击 ，显示如下参数： • **核采样**：模型在输出时，会从概率最高的词汇开始选择，直到这些词汇的总概率累计达到核采样值，这样可以限制模型选择这些高概率的词汇，从而控制输出内容的多样性。建议不要与温度同时调整。 – 数值较低，输出的文本更有确定性。 – 数值较高，输出的文本更有多样性。 • 取值范围0.1~1，默认值为0.5。 • **温度**：调高温度会使得模型的输出更多样性和创新性，反之，降低温度会使输出内容更加遵	DeepSeek-V3

参数	说明	示例
	循指令要求但减少多样性。建议不要与核采样同时调整。取值范围0~1，默认值为0.5。	
子工作流执行逻辑提示词	执行子工作流的提示词。**该提示词会反馈到大模型，大模型识别后，执行对应的子工作流。** 相当于一个角色设定，辅助智能体选择合适的子工作流执行任务。 ● 当单击“保存到模板”，填写“模板名称”、选择“行业”和“标签”后，可将提示词创建成模板并保存到我的提示词。 ● 当单击  图标时，可对系统提示词进行智能优化。 ● 当单击  图标时，系统会弹出“提示词广场”窗口，可在“预制提示词”或“我的提示词”页签中进行选择。	保持默认
意图识别（可选）	该多智能体应用的意图识别能力。 ● **不配置，则由模型决策执行的工作流。** ● **配置后，通过配置的工作流应用进行决策，执行对应的工作流。** 配置意图识别能力操作如下： 1. 在“意图识别（可选）”右侧，单击 ＋ ，在“添加工作流”页面的“当前空间”或“团队共享”选择具有特定输入输出参数的工作流应用。 **说明** – 单击“创建工作流”，创建工作流应用，具体操作请参考**创建工作流**。 – 仅**Versatile企业版**支持使用他人共享的应用。**Versatile基础版（限时免费）**不支持该能力。 2. 单击“确定”。 如果选择的工作流需要删除，单击  。	-

参数	说明	示例
起始工作流（可选）	**起始工作流配置后**，无论全局意图如何改变执行顺序，多智能体应用都会以**此工作流为起点**。 配置起始工作流操作如下： 1. 在“起始工作流（可选）”右侧，单击 ，在“添加工作流”页面的“当前空间”或“团队共享”选择作为起始的工作流应用。 **说明** – 单击“创建工作流”，创建工作流应用，具体操作请参考**创建工作流**。 – 仅**Versatile企业版**支持使用他人共享的应用。**Versatile基础版（限时免费）**不支持该能力。 2. 单击“确定”。 如果选择的工作流需要删除，单击 。	-
子工作流	最多支持添加30个子工作流。 添加子工作流的操作如下： 1. 在“子工作流”右侧，单击 ，在“添加工作流”页面的“当前空间”或“团队共享”，单击 。 如果选择的子工作流需要取消，单击 。 **说明** – 单击“创建工作流”，创建工作流应用，具体操作请参考**创建工作流**。 – 仅**Versatile企业版**支持使用他人共享的应用。**Versatile基础版（限时免费）**不支持该能力。 2. 单击“确定”。 如果选择的子工作流需要删除，单击 。 3. 设置子工作流的执行动作。 支持的执行动作如下： – **继续**：按该工作流的执行结果，继续执行其他的子工作流。 – **终止**：按该工作流的执行结果，调用结束工作流结束任务。 – **等待输入**：按该工作流的执行结果，待用户输入问题后执行任务。	选择对应的工作流

参数	说明	示例
关联智能体	在多智能体应用中，可以将一个多智能体应用用于另外一个多智能体应用中，从而实现多层控制的效果。目前支持2级控制。最多支持添加30个智能体。 关联智能体的操作如下： 1. 在“关联智能体”右侧，单击 ，在“添加智能体”页面的“当前空间”或“团队共享”，单击 。 如果选择的智能体需要取消，单击 。 **说明** 仅**Versatile企业版**支持使用他人共享的应用。 **Versatile基础版（限时免费）**不支持该能力。 2. 单击“确定”。 如果选择的智能体需要删除，单击 。	-
默认工作流（可选）	**当用户问题未匹配到任何子工作流业务意图时，执行当前默认工作流。** 配置默认工作的操作如下： 1. 在“默认工作流（可选）”右侧，单击 ，在“添加工作流”页面的“当前空间”或“团队共享”选择作为默认的工作流应用。 **说明** – 单击“创建工作流”，创建工作流应用，具体操作请参考**创建工作流**。 – 仅**Versatile企业版**支持使用他人共享的应用。**Versatile基础版（限时免费）**不支持该能力。 2. 单击“确定”。 如果选择的工作流需要删除，单击 。 3. 设置工作流的执行动作。 支持的执行动作如下： – **继续**：按该工作流的执行结果，继续执行其他的子工作流。 – **终止**：按该工作流的执行结果，调用结束工作流结束任务。 – **等待输入**：按该工作流的执行结果，待用户输入问题后执行任务。	-

参数	说明	示例
结束工作流（可选）	**结束工作流配置后**，无论全局意图如何改变执行顺序，多智能体应用都会以**此工作流为终点**。 ● 配置结束工作流的操作如下： 1. 在“结束工作流（可选）”右侧，单击  ，在“添加工作流”页面的“当前空间”或“团队共享”选择作为结束的工作流应用。 **说明** ■ 单击“创建工作流”，创建工作流应用，具体操作请参考**创建工作流**。 ■ 仅**Versatile企业版**支持使用他人共享的应用。**Versatile基础版（限时免费）**不支持该能力。 2. 单击“确定”。 如果选择的工作流需要删除，单击  。	-
全局意图	在与智能体交互过程中，用户可能有一些与业务无关的公共意图，例如“不感兴趣”、“非本人”等，可以将这些意图配置到全局意图，并且可以配置该意图对应的动作。 在“全局意图”右侧，单击  ，输入意图名称、处理方式、意图的执行动作。 支持如下处理方式： ● **直接应答**：配置一段文本，输出给用户。 ● **流程跳转**：关联一个工作流完成对应意图需要执行的动作。 支持如下执行动作： ● **继续**：按直接应答/工作流的执行结果，继续执行其他的子工作流。 ● **终止**：按直接应答/工作流的执行结果，调用结束工作流结束任务。 ● **等待输入**：按直接应答/工作流的执行结果，待用户输入问题后执行任务。	-

参数	说明	示例
高级配置	• **最大对话历史轮次**：设置历史对话次数，选择N，记录最近N条会话内容。例如，选择10，记录最近10条会话内容。取值范围0~100，默认值为10。 • **最大跳转次数**：多智能体运行过程中，根据用户意图，会在多个工作流之间跳转，为了**避免工作流之间无限循环跳转**，该参数可限制最大跳转次数。只有业务工作流之间跳转才会计算次数，起始工作流、结束工作流不计算跳转次数。取值范围0~30，默认值为9。 例如，一个多智能体应用含5个工作流，分别为工作流A（起始工作流或默认工作流）、工作流B、工作流C、工作流D、工作流E（结束工作流），根据用户问题先执行工作流A，根据工作流A的结果执行工作流B，根据工作流B的结果执行工作流C，再根据工作流C的结果执行工作流D，最后执行工作流E，相当于跳转了3次。	• 10 • 9

步骤7 单击“确定”。

设置后，进入多智能体编辑页面。

- 在多智能体编辑页面，显示多Agent控制器及添加的工作流、智能体及对应的工作流。

  单击工作流卡片，用户可以自定义意图名称、意图描述。

  单击智能体卡片，用户可以自定义意图名称、意图描述。

  单击智能体对应的工作流卡片，显示意图名称、意图描述、输入输出信息。

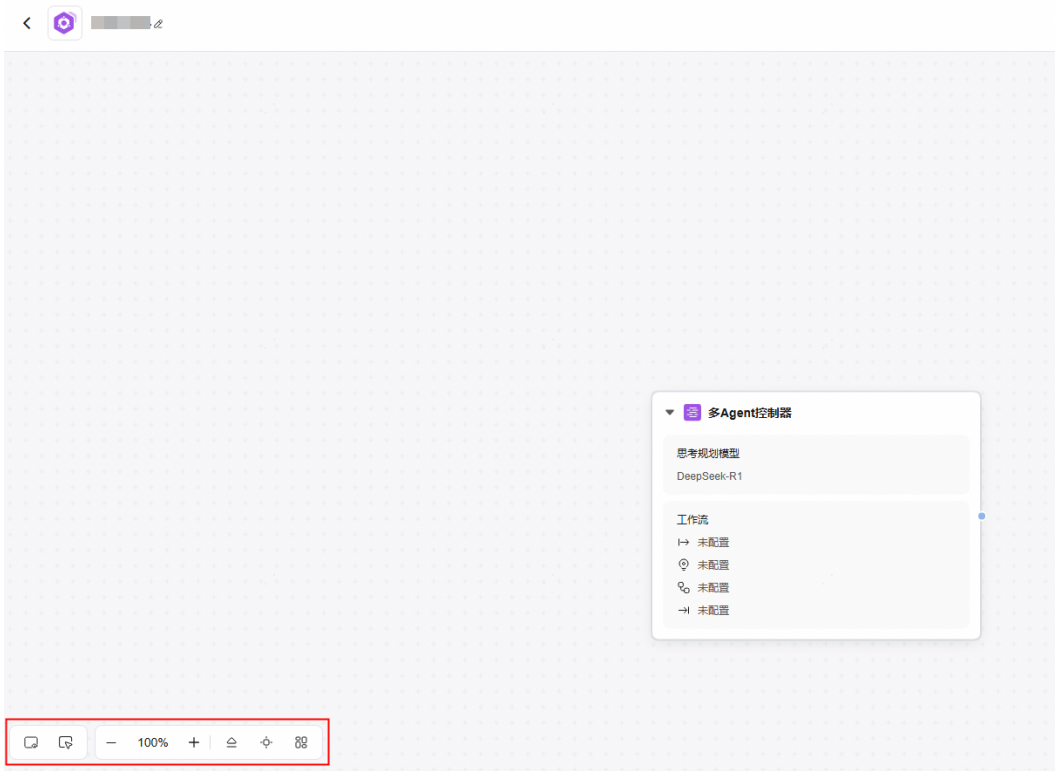
  在工作流或智能体卡片上，单击“ ··· > 工作流详情”或“ ··· > 智能体详情”，查看工作流或智能体详情。

  在工作流或智能体卡片上，单击“ ··· > 复制工作流ID”或“ ··· > 复制智能体ID”，复制工作流或智能体ID。
- 在多智能体编辑页面，可以调试、发布多智能体应用，调试与发布多智能体应用请参考**调试多智能体应用**、**发布多智能体应用为API**。

----结束

相关操作

在多智能体应用编辑页面，对画布的操作如下。



- ：显示/隐藏缩略图。
- ：查看/取消查看画布节点。
- 100% ：缩小/放大画布内容。
- /：全局折叠/展开节点。
- ：画布内容居中显示。
- ：画布内容布局优化。

在多智能体应用卡片列表中，支持的其他操作请参考表9-4。

表 9-4 相关操作

操作	说明
编辑多智能体应用信息	单击待编辑的多智能体应用卡片，进入多智能体编辑页面，在名称右侧单击，可以编辑多智能体应用的名称、描述、图标。
复制多智能体应用至其他空间	将此空间的多智能体应用复制到其他空间。 在待复制的多智能体应用卡片上，单击“ ··· > 复制”，在“复制到”页面，选择待复制的空间，单击“确定”。
复制多智能体应用的ID	在待复制ID的多智能体应用卡片上，单击“ ··· > 复制ID”，该多智能体应用的API接口被调用时，此ID为参数“agent_id”的值。

操作	说明
获取多智能体应用的调用路径	在待获取调用路径的多智能体应用卡片上，单击“ ··· > 调用路径”，在“调用路径”页面，单击“复制路径”。
删除多智能体应用	**注意** 如果应用已经上架，则用户无法 直接删除，必须先手动完成下架操作后 ，才能在应用管理页面自行删除应用。详细操作请参考**更多操作**。 在待删除的多智能体应用卡片上，单击“ ··· > 删除”。

相关文档

多智能体应用实践，请参考**多智能体应用实践**。

9.3 调试多智能体应用

开发者可以在多智能体应用创建完成后，直接与多智能体进行对话，实时观察其执行过程和响应效果，并根据需要对配置进行优化和调整。

创建多智能体应用后，Versatile支持对应用进行预览与调试。

前提条件

- 已**购买Versatile智能体平台**。
- 已**创建多智能体应用**。
- 登录用户为空间所有者、空间管理员、开发工程师，详细信息请参考**管理团队空间成员**。

调试多智能体应用

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航，选择“开发中心 > 应用管理 > 多智能体应用”，单击待调试的多智能体应用卡片。

步骤3 在多智能体应用编辑页面，单击“试运行”。

步骤4 在“试运行配置”页面，输入试运行配置，单击“开始运行”。

多智能体应用设置了全局配置中的输入参数或是使用的工作流应用的开始节点设置了输入参数，才会显示“运行配置”页面。

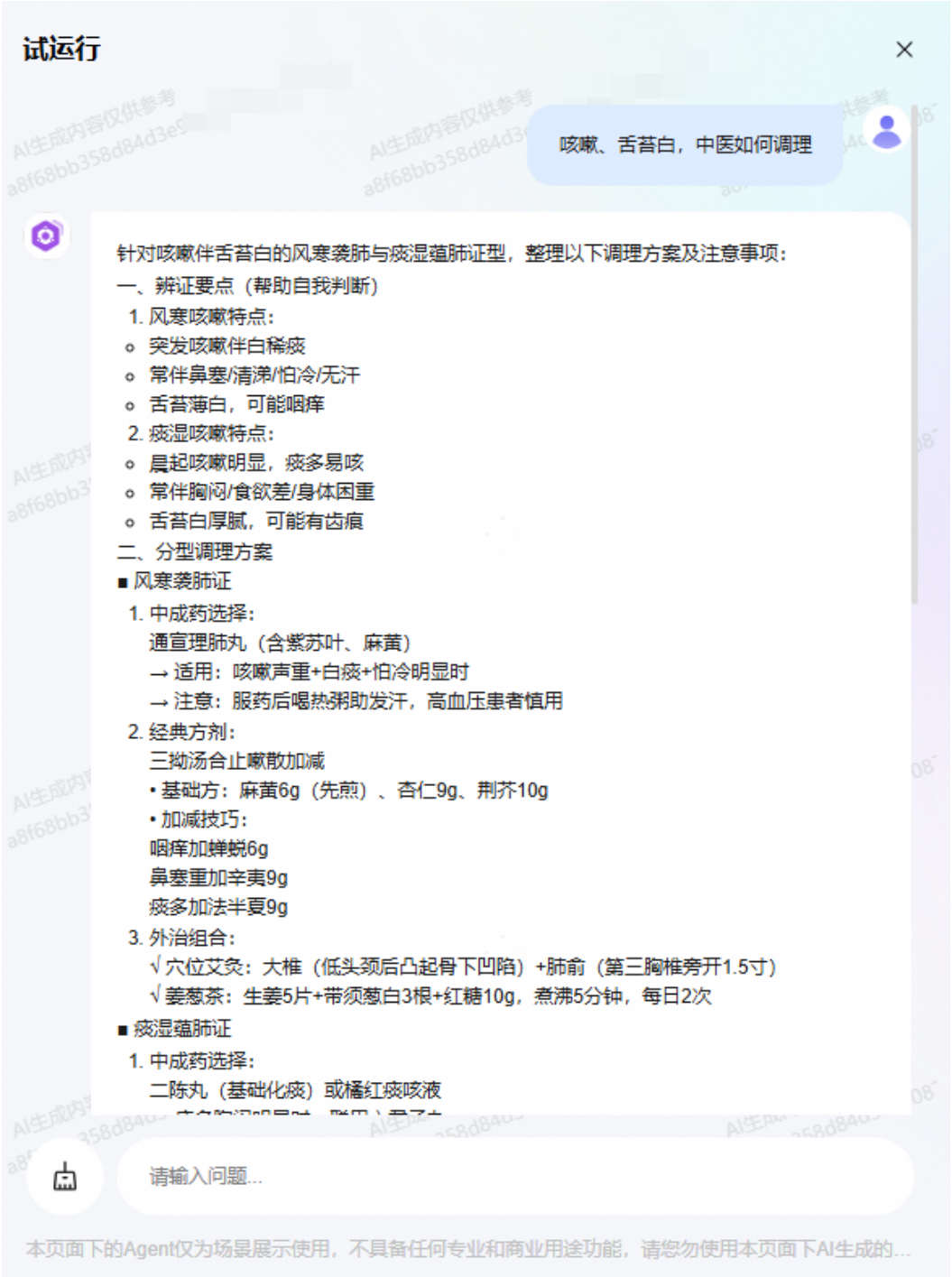
步骤5 在“试运行”页面，输入对话内容与智能体对话，并根据执行过程、响应结果，优化“多Agent控制器”参数配置。

文本输入：在对话输入框输入对话后按Enter键或单击  ，查看应用响应结果。

单击  ，清除本次会话内容，可以开始新的会话。

如果试运行失败，常见报错与解决方案请详见**应用开发常见问题**。如果运行结果返回为空，处理方法请参考**为什么多智能体应用返回为空？**。

图 9-5 试运行



步骤6 在试运行过程中，可以单击右上角 调试 查看调试结果，包括运行结果与调用详情，如**图9-6**所示。

- **运行结果**：运行结果中可以看到应用的执行开始时间、结束时间、运行时间等信息，还能看到输入和输出信息。对于性能的情况有个直观的认识。

- **调用详情**：在多智能体应用会话时，调用链中展示控制器或工作流的详细信息，包括运行的控制器或工作流、控制器或工作流耗时、控制器或工作流的输入和输出信息等。便于开发者快速地追溯操作顺序并精确定位问题。单击“查看子工作流调用链”，可以查看工作流中每个节点的详细信息，包括运行的节点、节点耗时、节点的输入输出。

  说明

  支持在**查看应用调用链信息**页面中，查看该调用链的详细信息，具体操作请参见**使用过滤器筛选信息**。

图 9-6 调试结果示例

----结束

9.4 发布多智能体应用为 API

多智能体应用试运行成功后，可对其进行发布，便于后续使用。

前提条件

- 已**购买Versatile智能体平台**。
- 已**创建多智能体应用**。
- 登录用户为空间所有者、空间管理员、开发工程师，详细信息请参考**管理团队空间成员**。

发布多智能体应用为 API

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航，选择“开发中心 > 应用管理 > 多智能体应用”，单击待发布的多智能体应用卡片。

步骤3 在多智能体应用编辑页面，单击“发布”。

已经发布的多智能体应用，修改后再次发布，显示为“更新发布”。

步骤4 在“发布”页面，配置发布信息，具体参数说明请参考**表9-5**。

图 9-7 发布

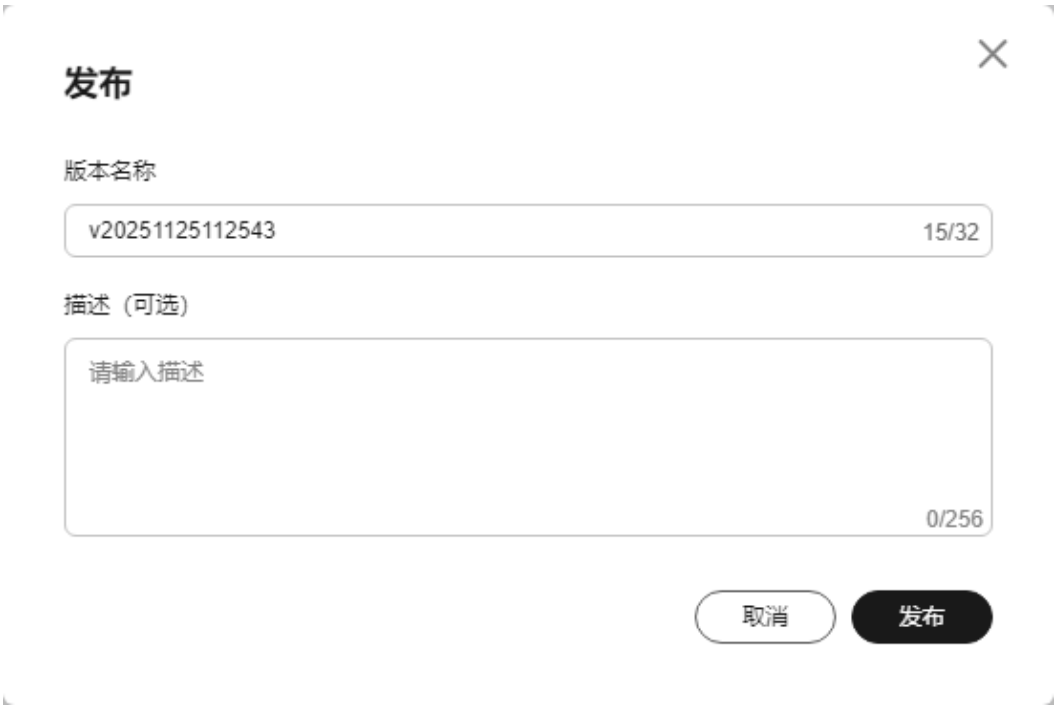


表 9-5 发布多智能体参数说明

参数	说明
版本名称	系统自动生成带年月日的版本名称，以v开头。也可以自定义版本名称，由1~32个字符组成。
描述（可选）	多智能体的描述信息。由0~256的字符组成。

步骤5 单击“发布”。

发布后，在“多智能体应用”页面的卡片上，显示“已发布”。

多智能体发布后，支持通过API调用，具体请参考**通过API调用多智能体应用**。

----结束

相关操作

在多智能体应用编辑页面，支持的其他操作请参考**表9-6**。

表 9-6 多智能体应用相关操作

操作	说明
查看版本历史记录	单击  ，查看版本历史记录。 在版本历史记录中，可以执行如下操作。 • 在“版本ID”右侧，单击  ，可以复制发布的版本ID。版本ID在多智能体应用被API调用时，为“version”参数的值。 • 单击“还原版本”，可以还原到此版本。 **说明** 还原后，当前工作流配置将不再保留，请谨慎操作。 • 单击“删除”，可以删除此发布版本。

相关文档

多智能体应用发布后，可以通过API接口调用，请参考**通过API调用多智能体应用**。

多智能体发布后，还可以共享给其他团队空间查看，具体操作请参考**共享应用**。

9.5 通过 API 调用多智能体应用

Versatile的API调用是应用开发中的强大工具，可以帮助用户快速集成功能和服务，同时支持与其他系统或服务进行交互，提升应用性能和用户体验。合理设计和管理API是确保应用安全和稳定的关键。通过API，用户可以构建功能丰富、高效的应用，满足多样化的用户需求。

通过API调用多智能体的优势：

- 提升开发效率

  API接口可以让不同的应用程序高效交互，大幅度节省数据传输与处理的人力和物力成本，同时可以帮助开发者将更多时间和精力集中在核心业务上。
- 扩展应用范围

  通过API接口的使用，可以实现不同系统、平台和服务之间的交互。这不仅促进了多平台之间资源共享，还能使开发者能够创新更多应用，从而增强企业的竞争力。

前提条件

- 已**购买Versatile智能体平台**。
- 在调用应用前，须确保应用已发布，具体请参考**发布多智能体应用为API**。
- 登录用户为空间所有者、空间管理员、开发工程师，详细信息请参考**管理团队空间成员**。

获取应用 ID 和调用路径

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航，选择“开发中心 > 应用管理 > 多智能体应用”，选择目标多智能体应用。

步骤3 单击“ ··· > 复制ID”，可获取当前应用ID。请记录保存，用于填写调用Agent应用接口的agent_id字段。

图 9-8 获取应用 ID

步骤4 单击“ ··· > “调用路径””，在弹出的“调用路径”页面，单击“复制路径”即可获取调用路径，如图9-9所示。

其中，“bbd49397-fe4f-4ce6-b174-f900e5efbd5d”是随机生成的字符串，在使用时可以替换为其他的字符串。字符串长度为1~64个字符，支持英文字母、数字、中划线、下划线。在多智能体应用被API调用时，为“conversation_id”的值。

图 9-9 获取应用调用路径

----结束

使用 API 调用多智能体应用

使用API调用多智能体应用的操作，请参考**调用智能体应用**。

9.6 导入导出多智能体应用

Versatile支持将多智能体应用在环境中导出，导入至另一个环境中，无需用户重复构建，快速完成多智能体应用跨环境构建或复用。

使用的业务场景如下：

- 从测试环境导出多智能体，到生产环境上部署。
- 在不同的开发环境之间迁移多智能体应用。
- 将多智能体应用下载到本地进行代码归档。
- 作为模板提供给其他客户进行复用。

前提条件

- 已**购买Versatile智能体平台**。
- 登录用户为空间所有者、空间管理员、开发工程师，详细信息请参考**管理团队空间成员**。

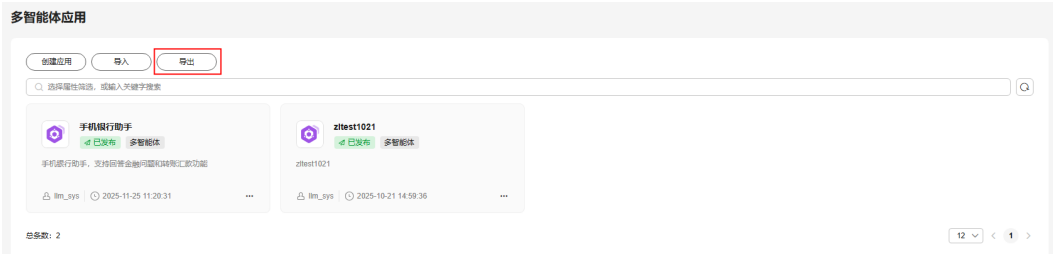
导入多智能体应用

已有从其他Versatile环境导出的多智能体应用的JSONL格式文件。

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航，选择“开发中心 > 应用管理 > 多智能体应用”，单击“导入”。

图 9-10 导入多智能体应用

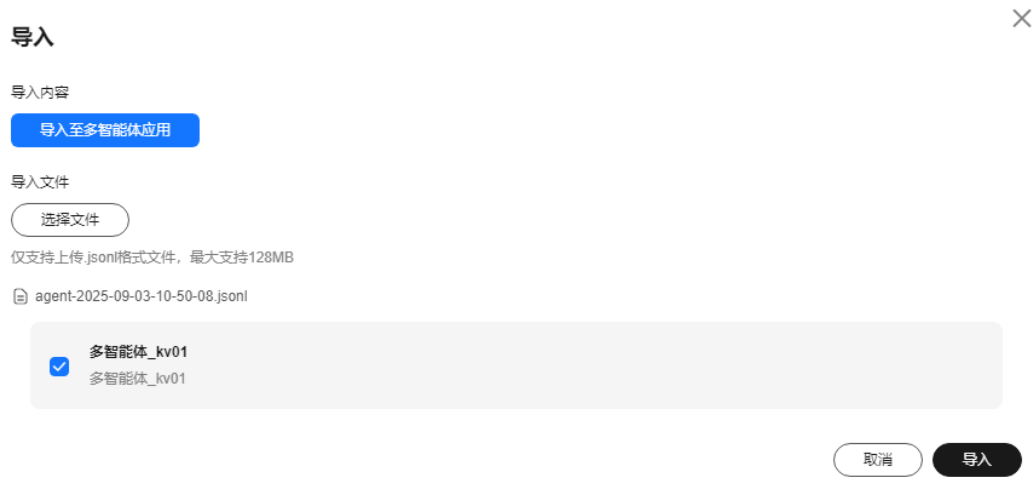


步骤3 在“导入”页面，单击“选择文件”，选择本地已准备好的文件，单击“导入”。

支持JSONL格式文件，最大128MB。

如果一个多智能体应用关联了其他多智能体应用，则在导入时，一并导入关联的多智能体应用。

图 9-11 导入多智能体应用



导入后，导入的多智能体应用显示在多智能体应用卡片列表中，如果名称相同，则会覆盖原有多智能体应用。

----结束

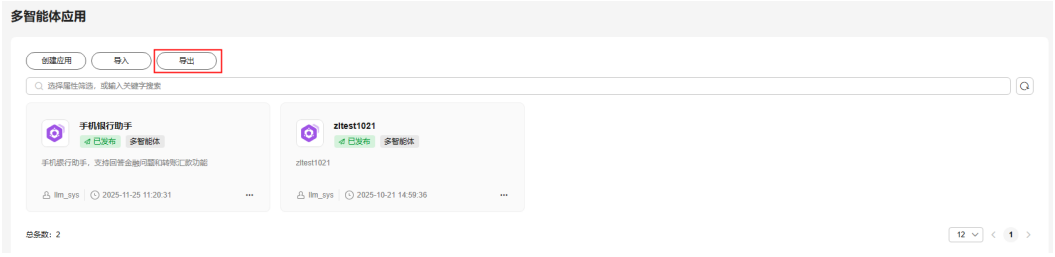
导出多智能体应用

已**创建多智能体应用**或已**导入多智能体应用**。

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航，选择“开发中心 > 应用管理 > 多智能体应用”，单击“导出”。

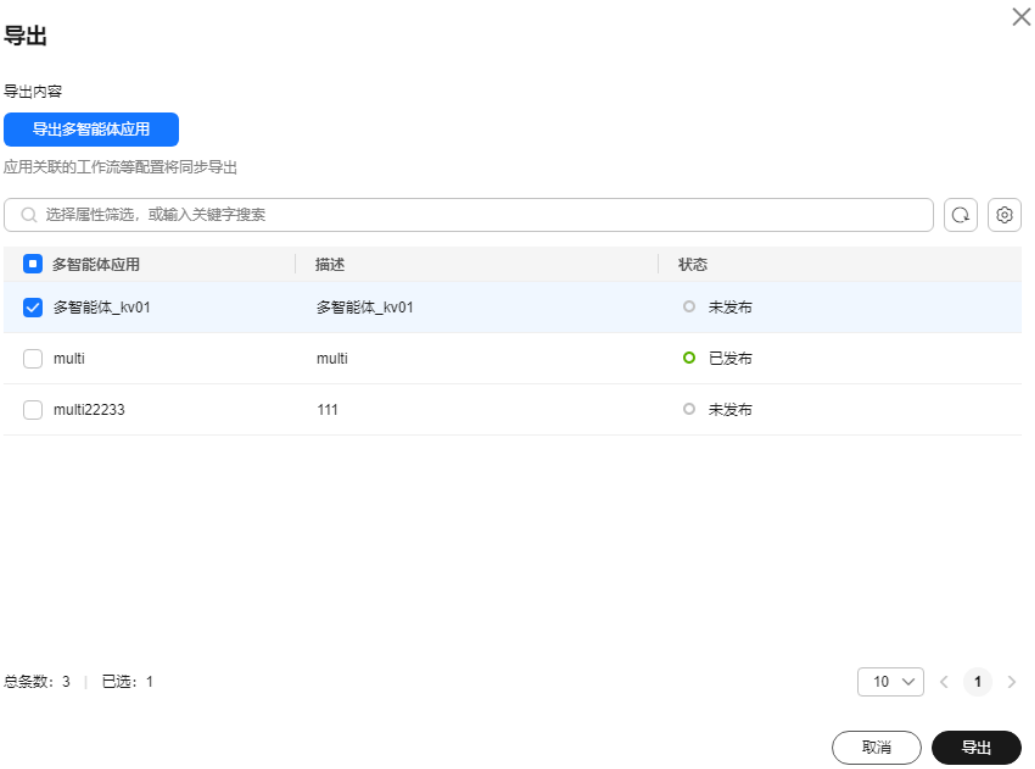
图 9-12 导出多智能体应用



步骤3 在“导出”页面，勾选待导出的多智能体应用左侧复选框，单击“导出”。

如果一个多智能体应用关联了其他多智能体应用，则在导出时，一并导出关联的多智能体应用。

图 9-13 导出多智能体应用



导出的文件以JSONL格式显示在本地。

----结束

10 管理资源

10.1 插件

10.1.1 插件介绍

插件介绍

在Versatile中，插件是智能体能力的重要扩展工具。通过模块化设计，插件能够为智能体提供丰富的专业技能和复杂任务处理能力，帮助其在多样化的实际场景中更高效地满足用户需求。

通过插件接入，用户可以轻松为智能体赋予本身不具备的能力。例如，在对话过程中，模型能够根据提示词自动感知适用的插件，并调用插件完成任务，最终返回执行结果。这种设计让应用能够自动化处理复杂任务，甚至跨领域解决问题，极大提升了智能体的实用性和灵活性。

Versatile支持以下类型的插件，满足不同用户的需求：

- **平台精选**：平台为用户提供了多种无需额外开发的预置插件。例如，“python_interpreter”能够根据用户输入的问题自动生成Python代码，并执行该代码获取结果。此插件为智能体提供了强大的计算、数据处理和分析功能，用户只需将其添加到智能体应用或工作流中，即可扩展功能。

图 10-1 插件广场

- **团队共享插件**：平台支持同一租户下的不同空间共享插件，这不仅增强了插件的功能多样性，还极大地提升了用户的使用体验。通过共享机制，用户可以在不同空间中无缝访问和使用所需的插件，无需重复安装或配置，从而提高了工作效率和资源利用率。
- **自定义插件**：为了满足个性化需求，平台还支持开发者创建自定义插件。通过简单的配置，开发者可以将API快速创建为插件，提供给智能体使用。这种方式让用户能够根据实际需求，为智能体添加专属功能，灵活满足多样化场景。

相关文档

- 在插件使用IP地址，为什么会调试失败?
- 如何配置插件请求参数中的“默认值”？
- 插件的输入参数支持图片格式吗?
- 工作流添加插件节点执行报错?

10.1.2 创建插件

10.1.2.1 基于 API 创建一个插件

本章节将介绍如何通过API创建插件。创建完成后，必须发布才可以被智能体或工作流使用。如何发布插件，请参考**发布插件**。

前提条件

已**购买Versatile智能体平台**。

创建插件

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏中选择“开发中心 > 组件库 > 我的插件”，单击左上角“创建插件”。

步骤3 在“创建插件”页面中的“插件类型”中选择“API类型”，然后根据以下步骤配置插件信息。

1. 在“基本信息”步骤中设置插件的基础信息，请参照表10-1完成信息配置，并单击“下一步”进入配置信息页面。

表 10-1 基本信息

参数	说明	示例
插件图标	单击默认图标按钮，可上传本地图片作为插件的自定义图标。 支持jpg、jpeg、png格式，不超过200KB。	系统默认图标
展示名称	用于标识当前插件，在添加到智能体或工作流后显示的名称。这有助于在智能体、工作流和资产中心中快速搜索和定位该插件。 命名规则： 命名要求：可以包含中文、英文、数字、特殊字符等； 长度限制：1~64个字符。	商品服务
名称	插件的英文名称。 命名规则： 命名要求：字母、数字和下划线（ _ ）的组合，不允许使用其他特殊字符或空格； 长度限制：1~64个字符。	Commodity Service
描述	描述当前插件的类型、功能和适用场景，帮助用户快速了解插件的作用和用途。	这是一个商品查询插件，可以用于查询商品的基础信息、库存和价格。
仅我可见	该功能默认关闭。开启后，仅插件的创建者可见。此设置在插件创建后无法修改。	

2. 在“配置信息”步骤中配置插件信息，请参照表10-2完成配置。

表 10-2 配置信息

参数	说明
协议	API服务接口通信协议。 – **https** – **http**

参数	说明
服务域名	提供API服务的服务域名。 以https://console.ulanqab.huawei.com/v1/chat/completions为例，服务域名为console.ulanqab.huawei.com 单击右侧的{{}}按钮，在服务域名中添加变量。添加变量后，可以在变量参数部分设置参数的描述。 在工具调测时，可以输入具体的参数值。 **图 10-2** 服务域名示例 
基准URL	基准URL（Base URL）是指域名的根路径，默认为/。 如果插件中存在多个工具，基准URL可以填写这些工具共用的URL部分。 以https://console.ulanqab.huawei.com/v1/chat/completions和https://console.ulanqab.huawei.com/v1/chat/workflows为例，基准URL可以填写为/v1/chat。 单击右侧的{{}}按钮，在基准URL中添加变量。添加变量后，可以在变量参数部分设置参数的描述。 在工具调测时，可以输入具体的参数值。 **图 10-3** 基准 URL 示例 

参数	说明
权限校验	选择调用API时是否需要鉴权。 – **无需鉴权**：API可以公开访问，不需要任何形式的身份验证或授权。 – **API Key**：在调用API时提供一个唯一的API Key进行鉴权。需配置以下信息 需填写密钥位置，并设置API Key的密钥鉴权参数名和密钥值。 ■ 密钥位置：密钥是从Header中读取还是从Query中读取。 ■ 参数名称：API Key的鉴权参数名称。 ■ 参数值：API Key的具体值。 – **华为云认证**：华为云IAM认证，通过IAM账号获取用户Token进行认证。 ■ IAM认证url：获取IAM用户Token信息的接口。例如，https://{iam_host}/v3/auth/tokens。 ■ 账号名：IAM用户所属账号信息，即账号名。 ■ 项目：该服务所属区域信息。例如，cn-southwest-2。 ■ 验证方式 ○ IAM用户名/密码 IAM用户名：IAM用户名称。 IAM用户密码：IAM用户的登录密码。 ○ Access Key ID/Secret Access Key Access Key ID：访问密钥ID。 Secret Access Key：与访问密钥ID结合使用的密钥。

步骤4 配置完单击“确定”。插件创建成功后，请参考**创建工具**为插件添加工具。

插件创建成功后，在“我的插件”界面查看创建完成的插件。您可以通过属性类型（展示名称、名称和描述）或搜索关键字的功能来查找插件。

----结束

创建工具

添加API下的具体接口作为插件的工具。

步骤1 在“我的插件”页面，单击需要添加工具的插件进入详情页面。

步骤2 在“工具信息”页签中，单击左侧的“创建工具”。

步骤3 在“添加工具”弹框中配置工具的展示名称、名称和描述，参数如**表10-3**所示。

表 10-3 基本信息参数说明

参数	说明
展示名称	用于标识当前工具，添加到智能体或工作流后将显示此名称。这有助于在智能体、工作流和资产中心中快速搜索和定位。
名称	工具的英文名称。 命名规则： • 命名要求：可以包含大小写字母、数字、下划线。 • 长度限制：1~64个字符。
描述	描述当前工具的功能和适用场景，帮助用户快速了解工具的作用。 长度限制：1~600个字符。

步骤4 在进行参数配置时，根据表10-4进行参数配置。

也可以通过单击“导入并解析”按钮，在弹框中输入cURL或openAPI代码，单击“确定”系统会自动解析代码内容并填充对应参数信息。

表 10-4 参数配置说明

参数		说明
工具URL Tool URL	请求方式	服务的请求方式，支持**POST**或**GET**。
	工具path	所调用API接口的访问地址或相关资源链接。如果已配置基准URL，则工具path应填写基准URL之后的部分；如果未配置基准URL，则工具path应为从服务域名之后的完整路径。 工具path中支持参数配置。 例如：/weather/weatherInfo{path_1}/{path_2}。 例如：https://console.ulanqab.huawei.com/v1/chat/completions • 基准URL为 /v1/chat • 工具path为 /completions
请求参数	参数封装	开启后，会将请求参数封装为一个列表（数组）结构，可适配入参为数组格式的插件接口。 例如： • 原参数列表：{"a":"string", "b":1}; • 开启封装后的参数列表：[{"a":"string", "b":1}]。
	请求头(Header)	HTTP请求消息的组成部分之一，请求头负责通知服务器有关于客户端请求的信息。 单击参数列表右侧的“添加参数”可以新增参数，参数配置说明请参见表10-5。

参数		说明
	请求体(Body)	HTTP请求消息的组成部分之一，请求体呈现发送给服务器的数据。 单击参数列表右侧的“添加参数”可以新增参数，参数配置说明请参见**表10-5**。
	查询参数(Query)	HTTP请求消息的组成部分之一，用于向服务器传递额外的参数信息。这些参数通常以键值对的形式出现，并且附在URL的路径后面，通过?分隔。 例如，在 /items?id=123 中，查询参数为ID，值为123。 单击参数列表右侧的“添加参数”可以新增参数，参数配置说明请参见**表10-5**。
	路径参数(Path)	自动解析工具path中包含的路径参数。工具path中支持可变参数配置。 例如：/weather/weatherInfo{path_1}/{path_2}。
响应参数	流式响应	该按钮默认关闭，开启后流式响应将逐步发送数据，减少延迟，支持实时传输、按需加载和中断，优化资源利用，提升用户体验。 **说明** • 单击“查看流式响应样例”按钮，可以查看流式响应的示例。 • 开启流式响应后，插件仍可配置响应参数；如果不配置响应参数，系统将默认传递后端返回的结果。
	参数封装	开启后，会将响应参数封装为一个列表（数组）结构，可适配出参为数组格式的插件接口。 例如： • 原参数列表：{"a":"string", "b":1}; • 开启封装后的参数列表：[{"a":"string", "b":1}]。
	参数名称	设置响应参数的名称。 命名规则：仅支持字母、数字或下划线。 **说明** • 单击参数列表右侧的“添加参数”，可以添加响应参数。 • 单击右侧的，可以删除添加的响应参数。
	描述	设置响应参数的详细描述信息，确保准确说明参数的含义、用途和格式要求，以提高大模型对参数识别和提取的准确性。
	参数类型	设置响应参数的数据类型。在下拉框中设置响应参数的数据类型。
	必填	设置该参数是否为必填项。

表 10-5 参数配置说明

参数	说明
参数名称	设置请求参数的名称，参数名称会作为大模型解析参数含义的依据。 命名规则：仅支持字母、数字、下划线或短横线。 **说明** • 单击参数列表右侧的“添加参数”按钮，可以添加请求参数。 • 单击右侧的，可以删除添加的请求参数。
中文名称	设置参数的中文名称，便于用户理解参数含义。
参数类型	设置请求参数的数据类型。 **注意** 请求头（Header）中，所有参数的值必须是字符串类型，不能设置其他类型。
默认值	设置参数的默认值，当参数未提供时使用该值。
描述	设置请求参数的详细描述信息，准确说明参数的含义、用途和格式要求，以提高大模型对参数识别和提取的准确性。
参数校验	设置当前参数是否需要进行校验。 校验规则： • 参数名称：需要校验的参数名称。 • 校验类型： – **字符最大长度** – **枚举值** – **时间日期** • 校验规则：可设置指定格式和自定义格式。 – **指定格式**：选择系统预置的标准校验规则。当校验类型为时间日期时，支持指定格式。 – **自定义格式**：根据实际需求自定义校验规则。
必填	设置该参数是否为必填项。

步骤5 单击“工具调测”按钮，输入请求参数值，单击“开始调测”检查调测结果。

步骤6 确保输出符合预期，再单击“自动解析”按钮，系统将自动生成响应参数。

步骤7 工具调试完成后，单击“确定”。

工具创建完成后，可以在工具列表中查看创建完成的工具。

----结束

更多操作

工具创建完成后，您可以在工具列表中查看每个工具的调试状态、智能体引用数和工作流引用数。您可以执行如**表10-6**的操作。

表 10-6 相关操作

操作	说明
编辑	在工具信息列表中，找到需要编辑的工具，单击该工具操作列中的“编辑”，可编辑工具信息。
调试	在工具信息列表中，找到需要调试的工具，单击该工具操作列中的“调试”，在展开的弹框中输入参数信息对工具进行调试。
删除	在工具信息列表中，找到需要删除的工具，单击该工具操作列中的“删除”，可删除工具。
查看详情	在工具信息列表中，找到需要查看详细信息的工具，单击该工具名称，可查看工具的详细信息。

10.1.2.2 基于函数创建一个插件

本章节将介绍如何通过函数来创建插件。创建完成后，必须发布才可以被智能体或工作流使用。通过使用插件，您可以扩展智能体和工作流的功能，使其更加灵活和强大。如何发布插件，请参考**发布插件**。

前提条件

已**购买Versatile智能体平台**。

创建插件的流程

序号	流程环节		说明
1	**创建插件**		Versatile智能体平台提供了自定义插件的功能，通过使用插件，可以增强智能体应用和工作流的能力。
2	**创建工具**	**创建工具**	在创建插件之后，需要为插件添加工具。一个插件可以包含多个工具，每个工具可以实现不同的功能。通过添加工具，可以进一步细化和扩展插件的功能。
		创建自定义依赖包	在函数编辑页面，可以创建并上传自定义依赖包。这些依赖包可以包含您的自定义代码，从而提高工具的灵活性和功能性。

创建插件

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏中选择“开发中心 > 组件库 > 我的插件”，单击左上角“创建插件”。

步骤3 在“创建插件”页面中的“插件类型”中选择“函数类型”，在“基本信息”步骤中设置插件的基础信息，请参照**表10-7**完成信息配置。

表 10-7 基本信息

参数	说明
插件图标	单击默认图标按钮，可上传本地图片作为插件的自定义图标。 支持jpg、jpeg、png格式，不超过200KB。
展示名称	用于标识当前插件的名称，便于在智能体、工作流和资产中心中快速搜索和定位。例如：查询天气。 命名规则： 命名要求：可以包含中文、英文、数字、特殊字符等。 长度限制：1~64个字符。
名称	插件的英文名称，用于在大模型调用时快速搜索和定位该插件。 命名规则： 命名要求：字母、数字和下划线（ _ ）的组合，不允许使用其他特殊字符或空格。 长度限制：1~64个字符。
描述	描述当前插件的类型、功能和适用场景，帮助用户快速了解插件的作用和用途。
仅我可见	该功能默认关闭。开启后，仅插件的创建者可见。此设置在插件创建后无法修改。

步骤4 配置完成后，单击“确定”完成插件创建。插件创建成功后，可以在插件列表中查看创建好的插件。请参考**创建工具**为插件添加工具。

插件创建成功后，在“我的插件”界面查看创建完成的插件。您可以通过属性类型（展示名称、名称和描述）或搜索关键字的功能来查找插件。

----结束

创建工具

步骤1 在“工具信息”页签中，单击左侧的“创建工具”。

步骤2 在“函数名称”的下拉列表中选择需要的函数，即之前已定义保存的函数。您也可以进行以下操作。

1. 单击  ：可以直接在弹出的创建函数页面快速创建函数，参数说明如**表10-8**和**图5 创建函数示例**所示，参数配置完成后可单击“确定”保存函数。
2. 单击  ：可以在弹出的编辑页面中快速编辑函数，参数编辑完成后可单击“确定”保存函数。

表 10-8 创建函数参数说明

参数	说明
名称	函数名，用于调用函数。 命名规则： 命名要求：可以包含英文、数字，必须以英文字母开头。 长度限制：2~32个字符。
描述	函数功能描述。
入参	设置函数的输入参数。  可以新增参数，参数配置说明请参见**表3**。
出参	输出参数。  可以新增参数，参数配置说明请参见**表3**。
执行语言	运行函数的环境。当前支持Python3.9、Node.js14.18，请查看**Python函数开发指南**、**Node.js函数开发指南**。

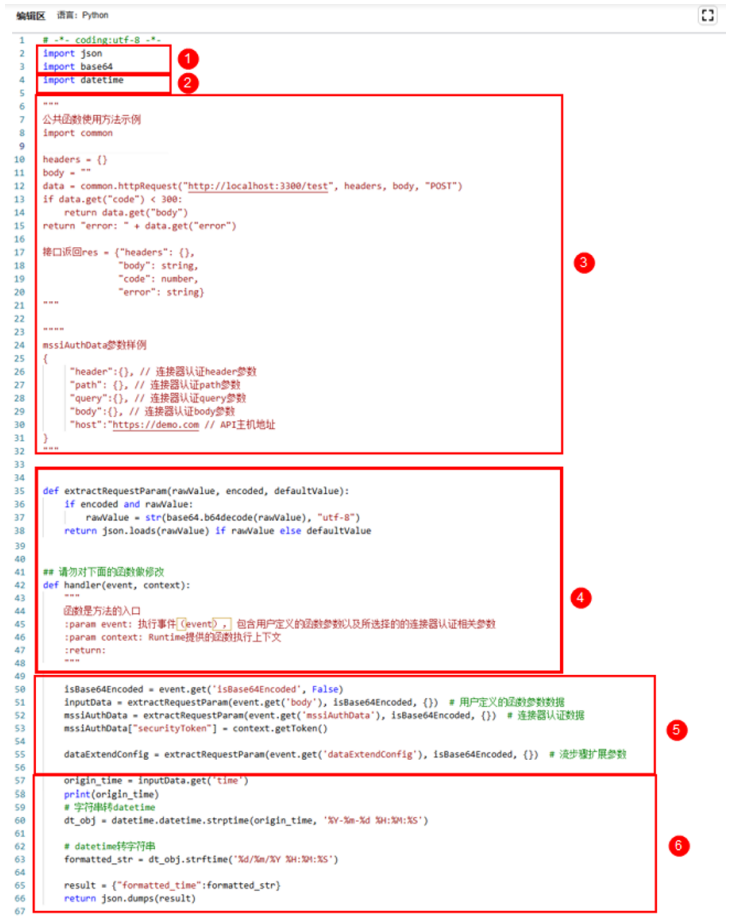
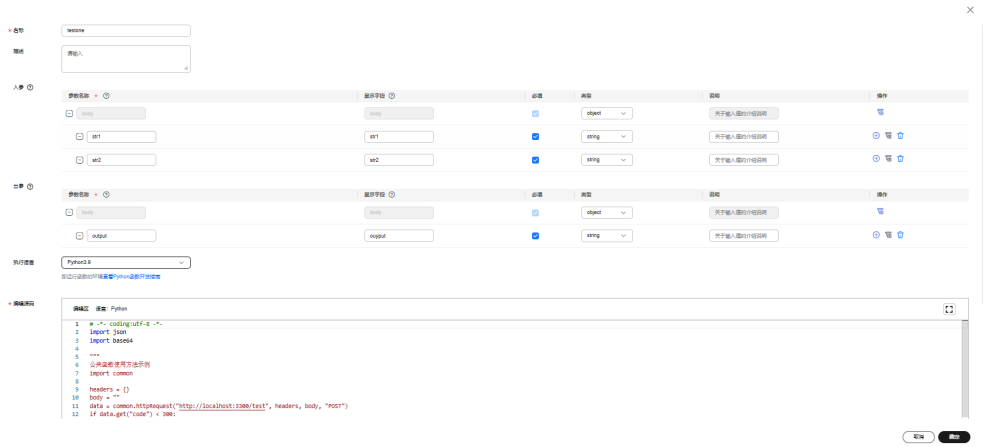
参数	说明
编辑源码	在源码编辑区，可以编写函数内部的代码运行逻辑，如图10-4所示，图中各模块说明如下： ①：导入模块，是Python标准库中的模块，无需修改。 ②：用户自定义导入模块。 ③：公共函数使用方法示例，提供了如何使用公共函数和mssiAuthData参数的示例，无需修改。 ④：函数定义和注释，extractRequestParam函数和handler函数是系统预置的模板代码，无需修改。 ⑤：系统方法，无需修改。 ⑥：用户自定义函数中的逻辑。输出为JSON格式，请参考示例的输出格式。 **图 10-4** 源码编辑区 
依赖包	单击右侧的“添加”，可以选择自定义依赖包。自定义依赖包上传方法请参见创建自定义依赖包。 一个函数最多添加**20个**依赖包。

表 10-9 参数配置说明

参数	说明
参数名称	输入参数的名称。
显示字段	输入该参数的别名。
必填	勾选该参数是否是用户必填项。
参数类型	选择参数类型，支持string、number、boolean、integer、array、object类型。 **注意** 入参的参数类型仅支持string、number、boolean、integer。
说明	关于该参数的介绍说明。
操作	- 单击 ：新增节点。 - 单击 ：删除该节点。

图 10-5 创建函数示例



步骤3 配置完成后，单击“确定”。

工具创建完成后，可以在工具列表中查看创建完成的工具。

----结束

创建自定义依赖包

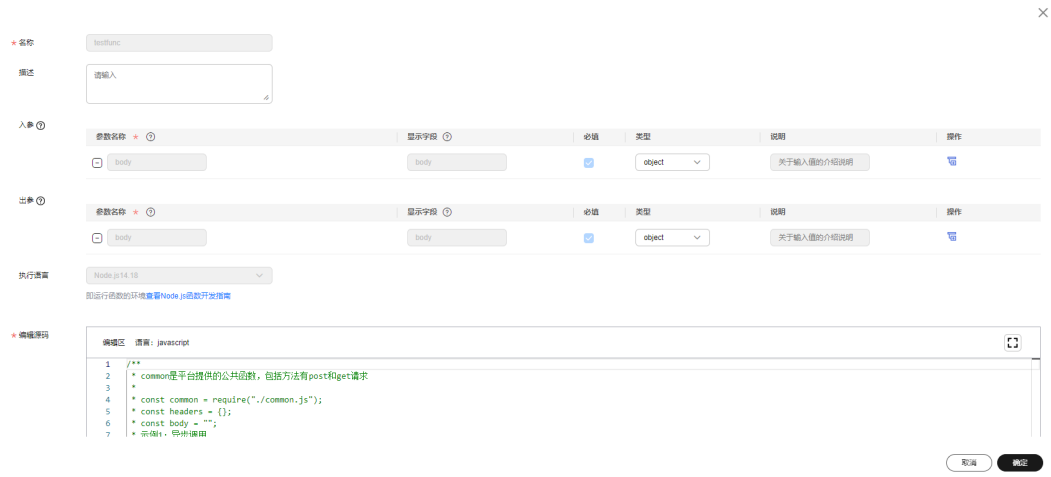
步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏选择“开发中心 > 组件库 > 我的插件”，单击您通过选择函数类型创建的插件。

步骤3 在“工具信息”页签中，单击左侧的“新建”。或编辑已添加的工具。

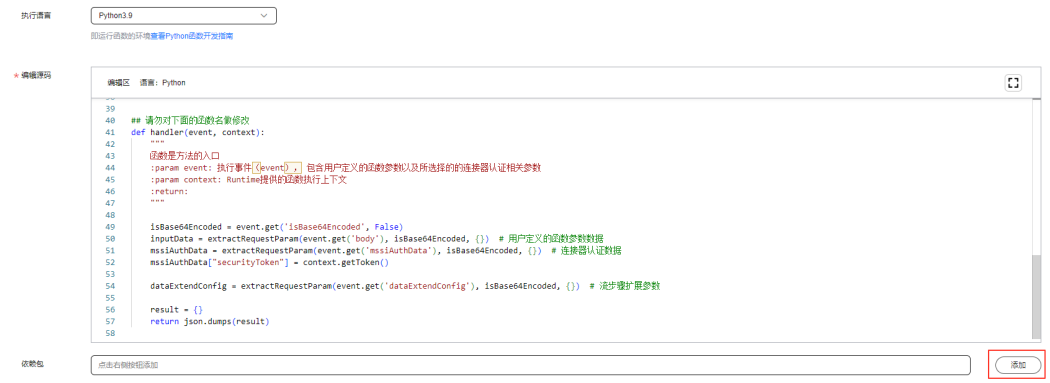
步骤4 选择已有函数，进入编辑函数界面，如图10-6所示。

图 10-6 创建函数



步骤5 单击“依赖包”右侧的“添加”，如图10-7所示。

图 10-7 添加依赖包



步骤6 在“选择依赖包”的弹框中，单击“创建依赖包”，设置依赖包的基本配置信息，具体的参数说明如表10-10所示。

表 10-10 新建依赖包参数说明

参数	说明
依赖包名称	自定义依赖包的名称。 命名规则： 命名要求：支持英文、数字、下划线，仅支持以英文开头。 长度限制：2~32个字符。
执行语言	运行函数的环境，当前仅支持Python3.9、Node.js14.18。
描述（可选）	依赖包的描述信息，最多支持200个字符。
上传	上传.zip格式文件，文件大小限制为10MB以内。 上传文件时，如果文件中包含敏感信息（如账户密码等），请您自行加密，防止信息泄露。

步骤7 单击“确定”。

创建完成后，可以在添加工具时添加并使用该依赖包。

----结束

示例

1. **开发语言**

   代码节点以Python语言为例。

2. **Python**

   基于Python 3.11.3的标准库，大多数模块都能正常运行，如下面白名单所示模块，不在白名单中的模块可能不能正常运行。

   - 三方库白名单

     ```
     sys,time,numpy,warnings,enum,os,functools,collections,types,datetime,numbers,abc,io,executor_sdk,contextlib,dataclasses,math,operator,pickle,contextvars,_contextvars,ast,re,ctypes,copyreg,weakref,textwrap,platform,typing,__future__,sympy,mpmath,bisect,cmath,colorsys,keyword,linecache,timeit,gc,random,decimal,_decimal,fractions,flint,gmpy2,unicodedata,tokenize,gmpy,copy,inspect,string,struct,importlib,array,shutil,pathlib,tempfile,subprocess,json,xml.etree.ElementTree,uuid,_uuid,urandom
     ```

   - 内置函数白名单

     ```
     exec,print,id,issubclass,compile,__build_class__,hasattr,eval,chr,next,ord,callable,repr,sorted,iter,min,max,weakref,all,any,hash,locals,sum,vars,open,abs,round,divmod,pow,delattr
     ```

3. 配置示例

   以数学计算示例代码为例：

   ```
   # -*- coding:utf-8 -*-
   import json
   import base64
   def extractRequestParam(rawValue, encoded, defaultValue):
       if encoded and rawValue:
           rawValue = str(base64.b64decode(rawValue), "utf-8")
       return json.loads(rawValue) if rawValue else defaultValue
   def math(args: dict) -> dict:
       # 注意在输入参数中定义名为input1的变量
       input1 = args.get('input1')
       try:
           input1 = int(input1)
           return {
               # 注意输出参数中定义res变量
               'res': input1 * input1
           }
       except Exception as e:
           return {
               # 注意输出参数中定义res变量
               'res': "输入类型错误或者数字大小超出限制"
           }
   ## 请勿对下面的函数名做修改
   def handler(event, context):
       """
       函数是方法的入口
       :param event: 执行事件（ event ）， 包含用户定义的函数参数以及所选择的的连接器认证相关参数
       :param context: Runtime提供的函数执行上下文
       :return:
       """
       isBase64Encoded = event.get('isBase64Encoded', False)
       inputData = extractRequestParam(event.get('body'), isBase64Encoded, {}) # 用户定义的函数参数数据
       mssiAuthData = extractRequestParam(event.get('mssiAuthData'), isBase64Encoded, {}) # 连接器认证数据
       mssiAuthData["securityToken"] = context.getToken()
       result = {'output1':math(inputData.get('input1'))}
       return json.dumps(result)
   ```


更多操作

工具创建完成后，您可以在工具列表中查看每个工具的调试状态、智能体引用数和工作流引用数。您可以执行如**表10-11**的操作。

表 10-11 相关操作

操作	说明
编辑	在工具信息列表中，找到需要编辑的工具，单击该工具操作列中的“编辑”，可编辑工具信息。
调试	在工具信息列表中，找到需要调试的工具，单击该工具操作列中的“调试”，在展开的弹框中输入参数信息对工具进行调试。
删除	在工具信息列表中，找到需要删除的工具，单击该工具操作列中的“删除”，可删除工具。
查看详情	在工具信息列表中，找到需要查看详细信息的工具，单击该工具名称，即可查看工具的详细信息。

10.1.2.3 通过 JSON 文件导入插件

Versatile支持通过JSON文件形式导入插件。创建插件后，必须发布插件才可以被智能体或工作流使用。如何发布插件，请参考**发布插件**。

前提条件

已**购买Versatile智能体平台**。

导入插件

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

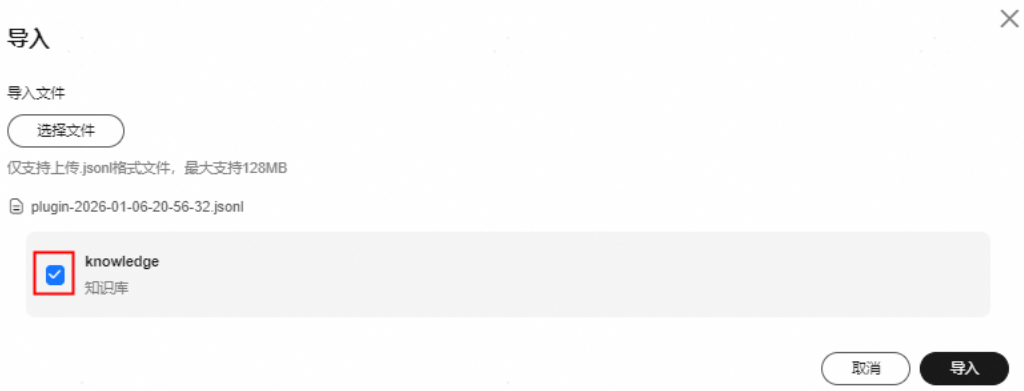
步骤2 在左侧导航栏中选择“开发中心 > 组件库 > 我的插件”，单击左上角“导入”。

步骤3 在“导入”页面，单击“选择文件”选择需要导入的jsonl格式的文件。

步骤4 勾选需要导入文件前的复选框后，单击“导入”，成功导入的插件将在“我的插件”页面中显示。

如果需在导入的插件中添加工具或编辑已有工具，请参考**创建工具**和**更多操作**。

图 10-8 导入插件



----结束

更多操作

插件导入完成后，在“我的插件”界面，您可以通过属性类型（展示名称、名称和描述）或搜索关键字的功能来查找插件。此外，您还可以对插件进行发布、删除、修改等操作，详情请参见**发布插件**和**管理插件**操作。

10.1.3 发布插件

只有发布状态插件中的工具，才能被智能体和工作流使用。

前提条件

已**创建插件**。

发布插件

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏中选择“开发中心 > 组件库 > 我的插件”页面。

步骤3 单击需要发布的插件，进入“插件详情”页面，单击右上角的“发布”按钮。

步骤4 输入“版本名称”和“描述（非必填）”，单击“发布”。发布之后的插件可以被单智能体或工作流引用。

----结束

10.1.4 使用插件

10.1.4.1 在单智能体应用中使用插件

Versatile应用支持添加插件功能，包括预置插件和个人插件。通过添加插件，您可以为单智能体应用扩展更多功能，提升其智能化水平。

前提条件

- 如果需要添加个人插件，请确保已完成个人插件的**创建插件**和**发布插件**。

- 如果需要添加平台精选的插件，请确保已对插件进行鉴权，详细信息请参考**使用平台精选的插件**。
- 如果需要添加团队共享插件，请确保已有他人共享的插件。如何查看他人共享的插件请参考**查看他人共享的插件**。

单智能体中使用插件

可以在智能体中使用插件，扩展智能体的能力。

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏中选择“应用管理 > 单智能体应用”页面。

步骤3 单击目标智能体，在“技能 > 插件”模块，单击 。

步骤4 在“添加插件工具”窗口，可以选择“插件广场”、“我的插件”或“团队共享”中的插件。单击目标插件右侧的 展开工具列表，在展开的列表中单击目标工具右侧的“添加”进行添加，并单击“确定”。

步骤5 添加插件后，可在“技能 > 插件”中查看当前已添加的插件。如需了解更多详细操作信息，请参考**相关文档**。

----结束

相关文档

单智能体应用中使用插件的详细信息，请参考**添加插件**。

10.1.4.2 在工作流中使用插件

通过插件的扩展功能，可以让工作流更强大、更智能、更自动化。用户只需选择合适的插件，快速实现需求，提升效率。

前提条件

- 如果需要添加个人插件，请确保已完成个人插件的**创建插件**和**发布插件**。
- 如果需要添加平台精选的插件，请确保已对插件进行鉴权，详细信息请参考**使用平台精选的插件**。
- 如果需要添加团队共享插件，请确保已有他人共享的插件。如何查看他人共享的插件请参考**查看他人共享的插件**。

工作流中使用插件

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏中选择“应用管理 > 工作流应用”页面。

步骤3 单击目标工作流，进入工作流详情页面。

步骤4 单击“添加节点”，在展开的弹框中单击“插件”选项，进入添加插件界面。

步骤5 在“添加插件工具”窗口，可以选择“插件广场”、“我的插件”或“团队共享”中的插件。单击目标插件右侧的 展开工具列表，在展开的列表中单击目标工具右侧的“添加”进行添加。

步骤6 添加插件后，在画布中查看已添加的插件。如需了解更多详细操作信息，请参考**相关文档**。

----结束

相关文档

在工作流中使用插件的详细信息，请参考**插件**。

10.1.5 管理插件

插件支持版本记录，方便您查看发布历史。在插件详情页面，您可以查看详细的发布记录，以及了解哪些工作流和智能体正在使用该插件。单击“查看历史”按钮，即可查看所有版本的发布记录；单击“引用插件列表”，即可查看插件被引用的详细信息。

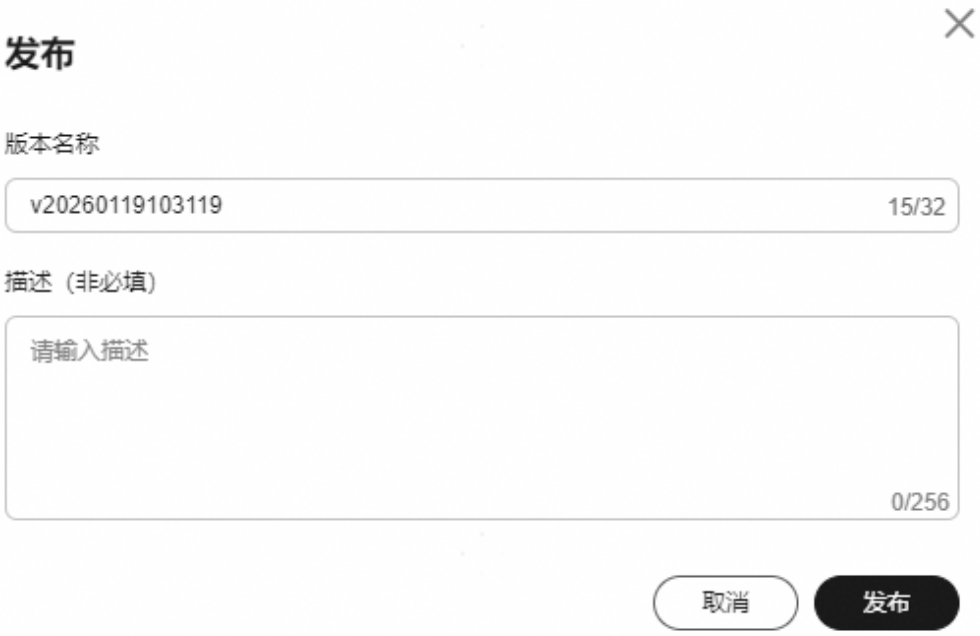
前提条件

- 已**创建插件**。
- 已**购买Versatile智能体平台**。

管理插件版本

在插件详情页面的右上角，单击“发布”按钮。发布插件时，您需要设置插件的版本名称和描述，如**图10-9**所示。这些信息会被记录在插件的历史版本页面，方便您以后查看和参考。发布成功后，系统会自动生成一个新的版本发布记录。

图 10-9 发布插件



查看插件的发布历史记录

在插件详情页面的右上角，单击发布历史图标  ，可以查看当前插件的发布版本记录。此页面按发布时间倒序显示历史记录，包括版本名称、插件ID和创建人、创建时间和描述信息。

可以单击不同版本的发布记录，查看和编辑该版本插件和工具的信息。修改后的信息需要重新发布新的版本之后才能生效。

图 10-10 插件发布历史

删除插件的发布历史记录

在插件详情页面的右上角，单击发布历史图标 。选择需要删除的历史记录，单击“删除”，即可删除该发布历史。

说明

如果该版本已被共享，将无法直接删除。必须先“取消共享”后，才能删除该版本，详细操作请参见**更多操作**。

图 10-11 删除发布历史

查看插件引用列表

- 单智能体中引用的插件无论是平台预置的插件还是个人创建的插件，都不会自动更新到最新版本。需要手动更新。

a. 在插件详情界面，单击引用插件列表图标 ，可以查看该插件被哪些智能体引用。

b. 在引用插件列表中单击智能体名称，将直接跳转至该智能体的编排页面，您可在此页面更新插件。

图 10-12 引用插件列表

- 在工作流中使用插件时，无论是平台预置的插件还是个人创建的插件，都不会自动更新到最新版本。这意味着即使插件发布了新版本，工作流仍会继续使用当前指定的版本，确保应用的稳定运行。如果您需要在工作流中使用最新的插件版本，可以在工作流页面根据提示手动升级插件版本，如图10-13所示。

图 10-13 工作流中插件升级

删除插件

步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏中选择“开发中心 > 组件库 > 我的插件”页面。

步骤3 在需要删除的插件卡片上单击“ ··· > 删除”，可删除当前插件。

图 10-14 删除插件

说明

- 如果该插件已被引用，删除后引用将自动取消，可能会导致工作流或Agent无法运行，且该操作不可撤回，请谨慎操作。
- 若该插件已被共享，将无法直接删除。必须先“取消共享”后，才能在“我的插件”界面中删除该插件，详细操作请参见**更多操作**。

----结束

导出插件

Versatile平台提供了方便的插件导出功能，帮助用户更高效地管理和使用插件。

导出插件：可以在Versatile平台中将自定义的插件配置导出为本地文件。这样，您可以轻松地将这些配置用于迁移或备份。

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏中选择“开发中心 > 组件库 > 我的插件”页面。

步骤3 单击页面左上角“导出”。

步骤4 （可选）在“导出”页面，可以查看插件列表。单击插件列表上方的  ，可以对插件列表进行基础设置，然后单击“确定”。

- 表格内容折行：
  - 开启时，插件列表单元格内容将自动换行显示。当单元格中的文本长度超过单元格宽度，文本将在单元格内自动换行显示。

  - 关闭时，插件列表单元格内容将不会自动换行显示。当单元格中的文本长度超过单元格宽度，文本将在单元格内被截断，而不是换行显示。
- 表格数据列固定：
  - 不固定：所有列均可水平滚动，当水平滚动表格时，所有列将同步移动。
  - 固定第一列：第一列将固定在表格的最左侧，其他列可以水平滚动，而第一列则始终保持在原位。
  - 固定前两列：前两列将固定在表格的最左侧，其他列可以水平滚动，而前两列则始终保持在原位。
- 自定义显示列：支持显示“插件”列和“描述”列。用户可以通过单击插件和描述前的复选框来自定义显示列。

图 10-15 基础设置

步骤5 在“导出”页面，勾选需要导出的插件前的复选框，然后单击“导出”。插件将以jsonl格式的文件下载至本地。

说明

导出功能仅支持整个插件导出，不支持单个工具的单独导出。

----结束

导入插件

Versatile平台提供了方便的插件导入功能，帮助用户更高效地管理和使用插件。

导入插件：Versatile支持将插件从一个环境导出并导入到另一个环境中。这意味着无需重复构建插件，可以直接使用从其他Versatile环境导出的插件的JSONL格式文件进行导入。

步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择目标空间。

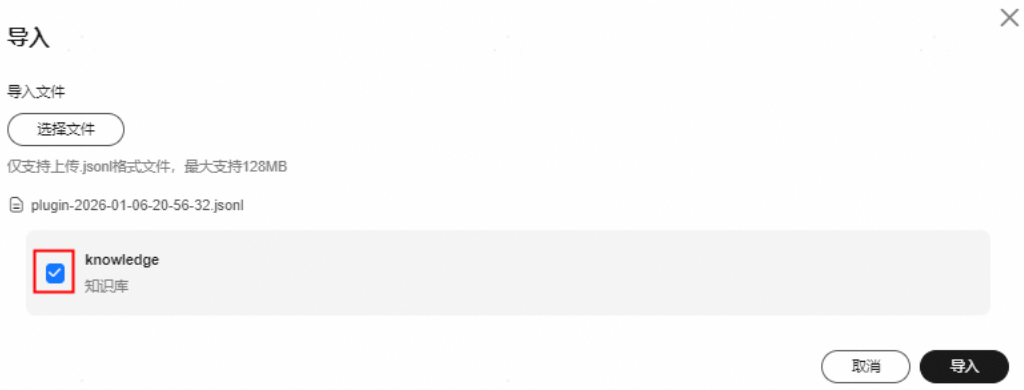
步骤2 在左侧导航栏中选择“开发中心 > 组件库 > 我的插件”，单击左上角“导入”。

步骤3 在“导入”页面，单击“选择文件”选择需要导入的jsonl文件。

步骤4 勾选需要导入文件前的复选框后，单击“导入”，成功导入的插件将在“我的插件”页面中显示。

如果需在导入的插件中添加工具或编辑已有工具，请参考创建工具和更多操作。

图 10-16 导入插件



说明

导入功能仅支持整个插件的导入，不支持单个工具的单独导入。

----结束

10.2 MCP

10.2.1 MCP 介绍

什么是 MCP

MCP（Model Context Protocol）是一个开放协议，旨在打通大语言模型（LLM）应用与外部数据源、工具之间的交互壁垒。在传统的开发场景中，由于每个数据源、工具或服务都有独立的格式规范、对接协议和认证体系，开发者往往需要为每个API单独编写代码、处理文档、配置认证方式和错误处理，这不仅效率低下，也大大增加了开发和维护的成本。

而MCP的出现，如同在AI模型与外部世界之间搭建了一座标准化的桥梁。MCP以通用的“标准语言”把工具、数据通过“MCP服务”的方式供给（一次开发、无限连

接），可以更高效、更便捷地实现Agent应用与成千上万的外部工具与数据的互通，极大提升了开发效率和灵活性。

Versatile提供两种MCP服务：资产中心预置的MCP服务和自定义创建的MCP服务，灵活满足不同场景的连接需求。如需了解更多关于MCP的详细信息，请参考**MCP官方文档**。

为什么使用 MCP

- **打破数据孤岛**

  传统大模型无法直接访问实时数据或本地资源，而MCP让AI“连接万物”，例如，查询天气时自动调用气象API，分析企业数据时直接连接内部数据库。

- **降低开发成本**

  在MCP出现之前，每个大模型需要为每个工具单独开发接口，导致重复劳动。而通过MCP，开发者只需写一次服务端，所有兼容MCP的模型都能调用。

- **提升安全性与互操作性**

  MCP内置权限控制和加密机制，比直接开放数据库更安全；同时，类似USB接口的标准化让不同厂商的工具能“即插即用”，避免生态分裂。

10.2.2 创建 MCP 服务

Versatile的资产中心提供多种MCP资源，用户通过简单安装即可快速集成调用。同时，平台支持灵活拓展，兼容开源社区MCP及自主开发MCP服务的接入。在Agent应用开发中集成MCP服务，可显著提升工具调用能力。

前提条件

- 已**购买Versatile智能体平台**。
- 如果需要接入自主开发的MCP服务，需确保该服务已完成独立部署。
- 如果需要安装第三方MCP，需要确保该服务已经订阅成功。订阅第三方MCP服务的详情信息，请参考**使用第三方MCP服务**。

  说明

  自主开发的MCP服务需在服务器或本地独立部署，并通过测试确保其能够正常运行。

创建 MCP 服务

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏选择“开发中心 > 组件库 > 我的MCP”，单击页面左上角的“创建MCP”创建MCP服务。

步骤3 在“创建MCP”的下拉列表中选择创建MCP的方式，目前支持**空白创建**、**官方模板创建**和**第三方服务安装**。

- **空白创建**：适用于安装开源社区的MCP或接入自主开发的MCP服务。

  a. 在“空白创建”的弹框中，输入MCP服务的配置信息，参数说明请参考**表10-12**。

表 10-12 创建 MCP 服务参数说明

参数	说明
服务图标	MCP服务的图标。 支持SVG、PNG、JPG、JPEG格式，不超过1MB。
服务名称	MCP服务名称，用于区分不同的MCP服务实例，不会对大模型的判断和调用产生影响。 命名规则： ■ 命名要求：仅支持以中英文开头。 ■ 支持字符：中英文、数字、中划线（ - ）、下划线（ _ ）。 ■ 长度限制：2~64个字符。
服务描述	MCP服务的描述信息，帮助用户理解服务功能。例如，一个强大的MCP服务器，可以轻松地将网页内容抓取并转换为各种格式（ HTML、JSON、Markdown、纯文本 ）。
服务介绍（ 非必填 ）	更详细地介绍该MCP服务的一些功能。例如，使用方式，关键能力，使用场景，注意事项等。
安装方式	选择MCP服务的安装方式，支持以下几种方式： ■ NPX：当MCP基于Node.js开发时选择NPX方式。 ■ UVX：当MCP基于Python开发时选择UVX方式。 ■ SSE：适用于与已部署在外部环境的远程MCP服务器建立连接，例如，接入自主开发的基于SSE协议的MCP服务。 ■ streamableHttp：适用于与已部署在外部环境的远程MCP服务器建立连接，例如，接入自主开发的基于streamable http协议的MCP服务。

参数	说明

输入MCP服务配置	支持使用表单编辑或使用JSON格式编辑。
	■ 使用JSON格式编辑。当使用JSON格式编辑时，加密的参数值将以明文显示。
	加密变量以_encrypt_开头，请谨慎操作。
	○ 安装平台精选的MCP时，默认展示当前MCP的服务配置，一般无需修改，部分服务需要填写API Key，请登录该MCP服务的官网获取。
	例如，百度地图的服务配置如下，“BAIDU_MAP_API_KEY”需要登录百度地图官网获取。
	{
	"mcpServers": {
	"baidu-map": {
	"command": "npx",
	"args": [
	"-y",
	"@baidumap/mcp-server-baidu-map"
	],
	"env": {
	"BAIDU_MAP_API_KEY": "xxx"
	}
	}
	}
	}
	○ 如果部署开源社区的MCP，请在开源社区该MCP服务详情页获取配置代码。
	○ 如果接入自主开发的MCP服务，服务配置模板如下，请将“url”更换为该MCP服务的实际部署地址。
	○ 基于SSE安装的MCP服务
	{
	"mcpServers": {
	"example-sse": {
	"url": "https://example.com?key=<您在服务官网上申请的key>"
	}
	}
	}
	○ 基于streamableHttp安装的MCP服务
	{
	"mcpServers": {
	"example-streamable-http": {
	"url": "https://example.com?key=<您在服务官网上申请的key>"
	}
	}
	}
	■ 使用表单编辑
	○ 参数：填写参数值时，多个参数值之间使用英文逗号分隔。例如，-y,@baidumap/mcp-server-baidu-map。
	○ 环境变量：单击“添加环境变量”，输入键值对，配置环境变量，例如，BAIDU_MAP_API_KEY: XXX。

参数	说明
	○ URL：外部环境部署的MCP服务的地址。例如，https://example.com?key=<您在服务官网上申请的key>。 ○ 请求头：单击“添加请求头”，输入键值对，配置请求头。 加密功能：单击输入框右侧的密文图标，可将参数值加密显示。

b. 单击“安装”。即可安装自定义的MCP服务。

- **官方模板创建**：适用于安装平台预置的MCP服务。
  a. 在“选择服务”步骤中，从预置的MCP列表中选择需要安装的MCP服务，单击“下一步”。
  b. 在“服务配置”步骤中，可以根据需要修改MCP服务的配置信息。参数说明请参考**表10-12**。
  c. 配置信息修改完成后，单击“安装”。即可安装预置的MCP服务。
- **第三方服务安装**：适用于安装订阅的ROMA Connect的MCP服务。订阅后的第三方MCP服务才能安装。
  订阅ROMA Connect的MCP服务的详情信息，请参考**使用第三方MCP服务**。
  a. 在“第三方服务安装”的弹框中，选择需要安装的MCP服务。
  如果需要批量安装多个第三方MCP服务，可以勾选“服务名称”列左侧复选框进行批量安装。

图 10-17 批量安装 MCP 服务

b. 单击“安装”，即可安装第三方MCP服务。

步骤4 安装完成之后，可以在“我的MCP”页面查看MCP的部署状态，部署成功的MCP才能在智能体或工作流使用。

----结束

更多操作

MCP服务创建完成后，在“我的MCP”界面，您可以通过属性类型（服务名称）或搜索关键字功能来查找MCP服务。也可根据需要执行如**表10-13**所示的操作。

表 10-13 更多操作

操作	步骤
卸载MCP服务	卸载后，MCP服务会下线，请谨慎操作。 在“我的MCP”页面，在MCP卡片上单击“卸载”按钮，即可卸载MCP服务。 **说明** 如果该MCP服务已被引用，卸载后引用将自动取消，可能会导致工作流或Agent无法运行，且该操作不可撤回，请谨慎操作。
编辑MCP配置信息	在“我的MCP”页面，单击MCP卡片，单击右上角  可以修改MCP服务的配置信息。 **说明** 第三方MCP服务的配置信息无法修改。
查看MCP服务概览	在“我的MCP”页面，单击MCP卡片，选择“概览”页签，可以查看MCP的服务描述、服务介绍、能力以及使用说明等信息。
查看工具	在“我的MCP”页面，单击MCP卡片，选择“工具”页签，可以查看MCP支持的工具详情并测试工具运行效果。

10.2.3 使用 MCP 服务

10.2.3.1 在单智能体应用中使用 MCP

通过Versatile的资产中心提供的多种MCP资源，用户能够快速集成和调用，同时平台支持灵活拓展，兼容自主开发MCP服务的接入，显著提升Agent应用的工具调用能力。

前提条件

- 如果需要使用自主开发的MCP服务，需确保已创建MCP服务且部署成功，详细信息请参考**创建MCP服务**。
- 如果需要使用平台精选的MCP服务，需确保已安装MCP服务，详细信息请参考**使用平台精选的MCP**。
- 如果需要使用第三方的MCP服务，需确保已安装MCP服务，详细信息请参考**使用第三方MCP服务**。

说明

自主开发的MCP服务需在服务器或本地上独立部署，并确保其能够正常运行。

单智能体中使用 MCP

步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏中选择“开发中心 > 应用管理 > 单智能体应用”页面。

步骤3 单击目标单智能体应用，在MCP服务模块，单击 + 。

步骤4 在“添加MCP服务”窗口，单击目标MCP服务右侧 + 进行一键添加，单击“确定”。

图 10-18 添加 MCP 服务

说明

建议添加的MCP服务不要多于5个。

----结束

相关文档

单智能体应用中使用MCP服务的详细信息，请参考为应用添加MCP服务。

10.2.3.2 在工作流应用中使用 MCP

在工作流中使用MCP能够提升开发效率和灵活性，通过标准化的接口连接AI模型与多种外部数据源和工具，实现一次开发、多处应用，有效打破数据孤岛，减少重复劳动，降低开发和维护成本，同时提供灵活的连接选项，满足不同场景的需求。

前提条件

- 如果需要使用自主开发的MCP服务，需确保已创建MCP服务且部署成功，详细信息请参考创建MCP服务。

- 如果需要使用平台精选的MCP服务，需确保已安装MCP服务，详细信息请参考**使用平台精选的MCP**。
- 如果需要使用第三方的MCP服务，需确保已安装MCP服务，详细信息请参考**使用第三方MCP服务**。

说明

自主开发的MCP服务需在服务器或本地上独立部署，并确保其能够正常运行。

工作流中使用 MCP

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 左侧导航栏中选择“开发中心 > 应用管理 > 工作流应用”，单击目标工作流。

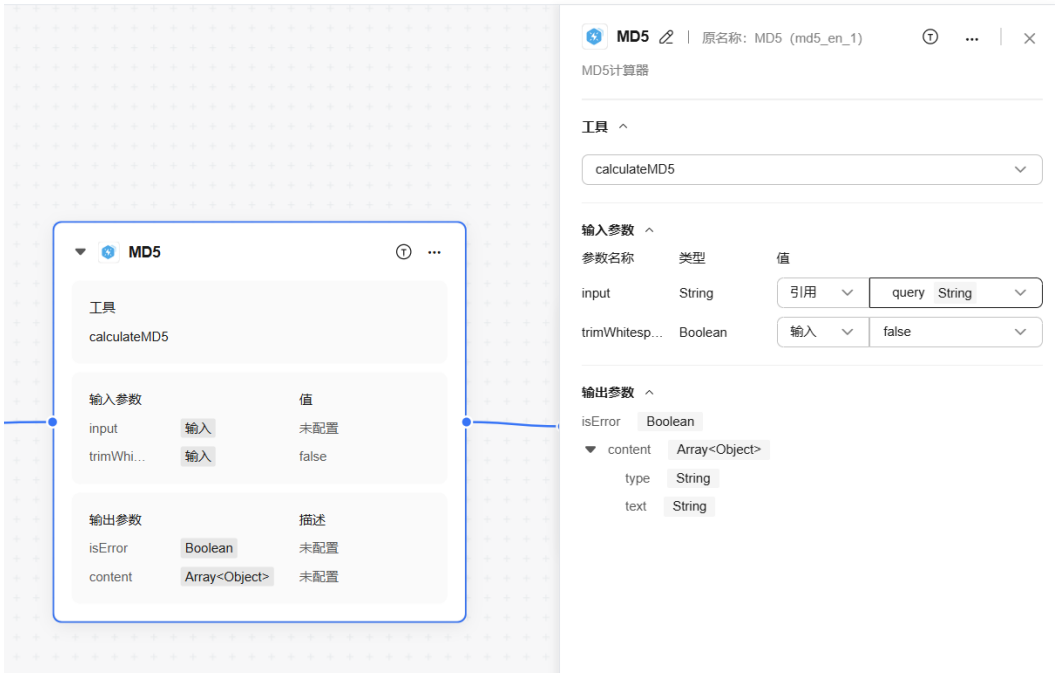
步骤3 单击“添加节点”并选择“MCP服务”节点。

步骤4 在“添加MCP服务”弹框中单击目标MCP服务右侧 + ，将MCP服务添加至画布中。

步骤5 连接MCP服务节点和其他节点。

步骤6 单击画布中已添加的“MCP服务”节点，完成MCP服务节点的配置。

图 10-19 MCP 服务节点配置示例



步骤7 节点配置完成后，单击“确定”。

----结束

相关文档

工作流应用中使用MCP服务的详细信息，请参考**MCP服务**。

10.3 知识库

10.3.1 知识库介绍

功能概述

知识库是组织、存储及管理知识的系统，涵盖文档、图片、视频等信息的分类整理，可以帮助用户高效管理大量的信息。在Agent中添加知识库，使其与用户提供的专业知识库进行交互，可以显著提升Agent的准确度和专业度。

Versatile提供的知识库功能对文本文档、FAQ（Frequently Asked Questions，常见问题解答）等数据进行向量化存储、知识检索，支持为应用、工作流提供检索增强能力。无论是文本文档、演示文稿，还是电子表格文件，用户都可以轻松地将数据导入知识库，无需额外的转换或格式处理。

知识库支持导入如表10-14所示的本地文档：

表 10-14 支持的文档格式

文档类型	文档格式	大小要求
知识文档	支持上传常见文本格式，包括：psd、 tiff、 bmp、 gif、 csv、 tif、 ico、 md、 jpeg、 jpg、 xlsx、 pcx、 dps、 png、 webp、 ofd、 docx、 et、 pptx、 txt、 pdf、 ppt、 doc、 wps、 xls格式。	单个文档上传限制最大60MB。
FAQ	支持按照模板上传文本，模板文件类型为Word及Excel。	• 仅支持xlsx、xls、docx、doc格式的文件，单个文件最大为60MB。 • Excel单个文件最大支持100000条数据，文件中不允许空行，空行后的数据将被忽略。

知识库类型

Versatile支持以下两种类型知识库的管理：

- 默认：在Versatile内直接创建并管理的知识库。支持上传文本文档、FAQ文档等文件，并对其进行向量化存储、知识检索。
- 第三方：将第三方系统中的知识库接入到Versatile中。当前支持对接开源第三方知识库RAGFlow。

相关文档

- 创建第三方知识库时为什么查不到知识库列表?

- 外部知识库连接中的已启用/已停用状态如何更改?

10.3.2 知识库使用限制

在使用知识库时应注意以下限制。

表 10-15 知识库使用限制

资源	限制说明
知识库数量	• 限时免费版单租户最多可创建5个知识库。 • 单租户最多可接入5个第三方知识库平台。 • 单租户最多可接入50个第三方知识库。 • 单个知识库上传的知识文档或FAQ文档总数量不超过500个。 • Versatile基础版（限时免费）： – 单个智能体最多可添加1个知识库。 – 单个工作流最多可添加1个知识库。 • Versatile企业版： – 单个智能体最多可添加3个知识库。 – 单个工作流最多可添加3个知识库。
知识文档	• 限时免费版单租户文档总大小不大于1GB。 • 单个文档不大于60MB。
FAQ文档	• 限时免费版单租户文档总大小不大于1GB。 • 仅支持xlsx、xls、docx、doc格式的文件，单个文件最大为60MB。 • Excel单个文件最大支持100000条数据，文件中不允许空行，空行后的数据将被忽略。

10.3.3 创建本地知识库

10.3.3.1 创建本地知识库流程

图 10-20 创建本地知识库流程

表 10-16 创建本地知识库流程

序号	流程环节		说明
1	创建知识库		Versatile提供的知识库功能对文本文档、FAQ等数据进行向量化存储、知识检索，支持为应用、工作流提供检索增强能力。
2	将知识信息更新至知识库。 **说明** Versatile支持4种方式将知识信息更新至知识库，用户可根据需求选择合适的方式。	上传知识文档	在创建知识库时，您可以选择上传本地的知识文档到知识库。
		OBS接入文件	在创建知识库时，支持通过对象存储服务（Object Storage Service, OBS）接入知识文档。
		创建FAQ问答对	FAQ问答对是指常见问题及其对应的答案，用于快速解决用户可能遇到的问题。
		上传FAQ文档	知识库支持通过上传FAQ文档来批量导入FAQ问答对。
3	测试知识库命中率		Versatile通过对创建的知识库进行命中率测试，可以评估知识库的效果和准确性。

10.3.3.2 创建知识库

Versatile提供的知识库功能对文本文档、FAQ等数据进行向量化存储、知识检索，支持为应用、工作流提供检索增强能力。

本文将详细介绍如何创建知识库，包含模型配置、解析配置和拆分配置等。

前提条件

已**购买Versatile智能体平台**。

新建知识库

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏中选择“开发中心 > 知识库”，单击左上角“新建知识库”。

步骤3 在“选择创建类型”弹框中选择“默认”，单击“确定”。

步骤4 参照**表10-17**完成参数配置。

图 10-21 创建知识库

表 10-17 参数说明

参数		说明	示例
基本信息	知识库图标	可选参数。 知识库LOGO。单击当前显示的知识库图标，在弹出的对话框中，选择要上传的新图标文件。 支持jpg、jpeg、png及gif格式，大小不大于200KB。	-
	知识库名称	必选参数。 用于标识知识库。 命名规则：可以包含字母、数字、中文、下划线 _、中划线-，且必须以字母、数字、中文开头，长度1~50个字符。	Versatile的知识库_001
	描述	必选参数。 用于对知识库内容和用途的简要说明。它提供了关于知识库的详细信息，帮助用户了解知识库的内容和使用场景。 命名规则：长度不大于100个字符。	知识库
模型配置	向量模型	必选参数。 向量模型是一种将文本、图像等非结构化数据转换为数值向量的模型。例如，在文本处理阶段，用于对文本文档进行切片，转换成向量化表示；在知识检索阶段，根据用户输入的信息对切片进行召回。 向量模型用于在海量的知识库中，快速识别和用户输入信息语义相近的词或句子，进行信息的初步筛选，解决“大海捞针”的效率问题。 取值范围： pangu_embedding：系统预置的模型。	pangu_embedding
	精排模型	必选参数。 精排模型是一种用于对检索结构进行精细排序的模型。针对用户输入的信息，对向量模型召回的切片进行从高到低的相关度排序，把相关度最高的前几个信息（例如Top 10）呈现给用户。 精排模型用于进一步提升系统搜索的相关性精度。 取值范围： pangu_rerank：系统预置的模型。 **说明** • **Versatile基础版（限时免费）**用户在创建知识库后，无法修改精排模型的配置。 • **Versatile企业版**用户在创建知识库后，可以通过“高级配置”选项来修改精排模型。具体操作步骤，请参考**更多操作**。	pangu_rerank
解析配置（非必选）	OCR增强	• 不开启，不可调用OCR服务进行智能文档识别。 • 开启后，即可调用OCR服务进行智能文档识别，如表格解析或扫描文件等。	页眉页脚解析

参数		说明	示例
	页眉页脚解析	• 未开启，解析结果中不包含页眉页脚。 • 开启后，解析结果中包含页眉页脚。	
	目录页解析	• 未开启，解析结果中不包含目录页。 • 开启后，解析结果中包含目录页。	
	图片解析	• 未开启，则在文档中遇到图片默认跳过，不处理图片。 • 开启后，根据需要选择“提取图片文本”或者“仅保留原图”。 – 提取图片文本：识别图片内文字。 – 仅保留原图：仅提取图片保存，不会识别图片内容，便于问答图文展示。	

参数		说明	示例
拆分配置（非必选）	拆分设置	• 自动分段：按照系统默认预设的规则和分隔符切分。 • 长度分段：基于内容的长度来决定如何进行分段。 – 分段标识符：分段方式为遇到所选符号即截断，符号之间没有优先级，最终分割后合并到预计最大长度。自定义分段中如果未命中分段标识符，分段将会失败。 ▪ 中文句号。 ▪ 英文句号. ▪ 中文感叹号！ ▪ 英文感叹号! ▪ 中文问号？ ▪ 英文问号? ▪ 空格 ▪ 中文逗号， ▪ 英文逗号, – 分段预计长度：分段的最大长度。文档的正文如果大于设定的最大长度，则截取最大长度的片段为新文档，随后回溯分段重叠字符，继续向后检查，直到文档结束。 取值范围：1~6000 默认值：500 • 层级分段：根据内容的结构层次来进行分段。 – 层级解析模型： ▪ 自动解析：自动识别和解析具有层级结构的数据或信息。 ▪ 规则解析：支持添加自定义层级规则。 – 标题层级深度：指设置的切分标题级别，例如，文本包含最多5级标题，选择的标题层级深度为3，则会分别将所有3级标题下的内容合并成文本块，文本块作为一个整体执行后续切分操作。输入值必须在1到10之间。 – 标题保存方式：指标题信息在切片中的保存形式，影响检索结果的展示逻辑和索引构建方式。 ▪ 保存多标题组合：多级标题用特定符号组合：1级标题-2级标题-3级标题-…-文本	自动分段

参数		说明	示例
		▪ 保存最后一级标题：仅组合最后一级标题：最后一级标题-文本 – 跨标题合并：根据需求开启或者关闭。 ▪ 开启“跨标题合并”功能：当不同标题下的段落文字较少时，平台会自动将其合并到指定的分段长度，有助于生成更加全面的内容。 ▪ 关闭“跨标题合并”开关：不会自动合并不同标题下的内容。 – 分段标识符：分段方式为遇到所选符号即截断，符号之间没有优先级，最终分割后合并到预计最大长度。自定义分段中如果未命中分段标识符，分段将会失败。 ▪ 中文句号。 ▪ 英文句号. ▪ 中文感叹号！ ▪ 英文感叹号! ▪ 中文问号？ ▪ 英文问号? ▪ 空格 ▪ 中文逗号， ▪ 英文逗号, – 分段预计长度：分段的最大长度。文档的正文如果大于设定的最大长度，则截取最大长度的片段为新文档，随后回溯分段重叠字符，继续向后检查，直到文档结束。 取值范围：1~6000 默认值：500	

步骤5 配置完成后，单击“确定”，完成知识库创建。创建完成后，可以在“知识库”界面查看创建完成的知识库。

创建完成的知识库，默认是启用状态。

----结束

更多操作

知识库创建完成后，在“知识库”界面，您可以通过分类筛选（全部类型、默认和第三方）或搜索（按名称和来源）功能来查找知识库。您还可以执行如**表10-18**的操作。

知识库支持以列表和卡片形式展示。单击搜索框右侧 按钮，知识库将以列表形式展示。单击搜索框右侧 按钮，知识库将以卡片形式展示。

表 10-18 相关操作

操作	说明
启用知识库	• 当知识库以列表形式展示时，找到“状态”是“已停用”的知识库，单击操作列“启用”，可以启用知识库。 • 当知识库以卡片形式展示时，找到“状态”是“已停用”的知识库，单击卡片右下角 ··· 按钮，单击“启用”，可以启用知识库。 **只有“状态”是“已启用”的知识库才能在应用、工作流中引用该知识库。**
停用知识库	• 当知识库以列表形式展示时，找到“状态”是“已启用”的知识库，单击操作列“停用”，可以停用知识库。 • 当知识库以卡片形式展示时，找到“状态”是“已启用”的知识库，单击卡片右下角 ··· 按钮，单击“停用”，可以停用知识库。 **注意** 停用已经被应用、工作流引用的知识库，会导致检索结果返回空值，请谨慎操作。
命中测试	• 当知识库以列表形式展示时，单击操作列“命中测试”，可以测试知识库命中率。 • 当知识库以卡片形式展示时，单击卡片右下角 ··· 按钮，单击“命中测试”，可以测试知识库命中率。 详细操作请参见**测试知识库命中率**。
编辑知识库	编辑知识库可以修改“知识库图标”、“知识库名称”、“知识库描述”。 • 当知识库以列表形式展示时，单击操作列“更多 > 编辑”，可以编辑知识库。 • 当知识库以卡片形式展示时，单击卡片右下角 ··· 按钮，单击“编辑”，可以编辑知识库。 **只有“状态”是“已停用”的知识库才可编辑。**
高级配置	高级配置可以修改“模型配置”，“解析配置”，“拆分配置”。 • 当知识库以列表形式展示时，单击操作列“更多 > 高级”，可以编辑知识库。 • 当知识库以卡片形式展示时，单击卡片右下角 ··· 按钮，单击“高级”，可以编辑知识库。 **只有“状态”是“已停用”的知识库才可修改高级设置。**

操作	说明
删除知识库	• 当知识库以列表形式展示时，单击操作列“更多 > 删除”，可以删除知识库。 • 当知识库以卡片形式展示时，单击卡片右下角 ··· 按钮，单击“删除”，可以删除知识库。 **只有“状态”是“已停用”的知识库才可删除。** **注意** 删除应用属于高危操作，删除前，请确保该知识库不再使用。
查看引用	可以查看当前知识库被哪些智能体和工作流引用。具体有以下三种查看方式： • 当知识库以列表形式展示时，单击操作列“更多 > 查看引用”，可以查看知识库被哪些智能体和工作流引用。 • 当知识库以卡片形式展示时，单击卡片右下角 ··· 按钮，单击“查看引用”，可以查看知识库被哪些智能体和工作流引用。 • 在知识库详情页面，单击右上角的“引用列表”，可以查看知识库被哪些智能体和工作流引用。 **说明** 单击引用列表中的智能体或工作流名称，可跳转到具体的应用详情。

10.3.3.3 上传知识文档

在创建知识库后，用户需将知识信息上传至知识库。

本文将详细介绍创建知识库后如何上传知识文档。

- **上传知识文档**。
- **（可选）新增文档切片**：知识文档上传并解析完成后，用户可根据需要新增/编辑切片，以提升知识库的检索效率。
- **（可选）重新对文档解析拆分**：如果用户对知识库内文档的解析切片效果不满意，可修改知识库模型配置、解析配置和拆分配置信息，并对知识库内文档重新解析拆分。

前提条件

- 已完成**创建知识库**。
- 已完成知识文档整理。

上传知识文档

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

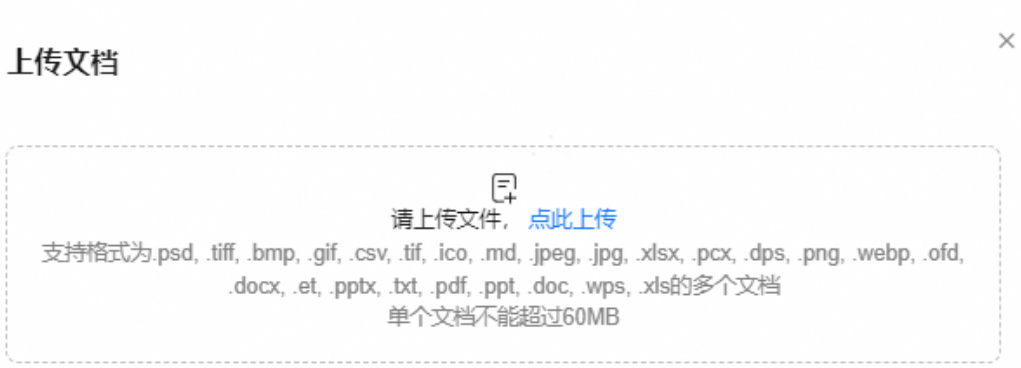
步骤2 在左侧导航栏中选择“开发中心 > 知识库”。

步骤3 在“知识库”页签，单击知识库列表中所需知识库名称，进入该知识库详情页面。

步骤4 在知识库详情页面选择“知识文档”页签，单击“上传”进入文档上传页面。

步骤5 单击“点此上传”，在弹出的对话框中，选择要上传的文档。

图 10-22 上传文档



说明

- 支持格式为psd、 tiff、 bmp、 gif、 csv、 tif、 ico、 md、 jpeg、 jpg、 xlsx、 pcx、 dps、 png、webp、 ofd、 docx、 et、 pptx、 txt、 pdf、 ppt、 doc、 wps、 xls的多个文档。
- 单个文档不能大于60MB。
- 单个知识库最多可上传500个文件。

步骤6 单击“确定”，文件列表中有对应文件，即完成文件上传。

待文件状态为“成功”，即完成文件解析。

图 10-23 上传文件

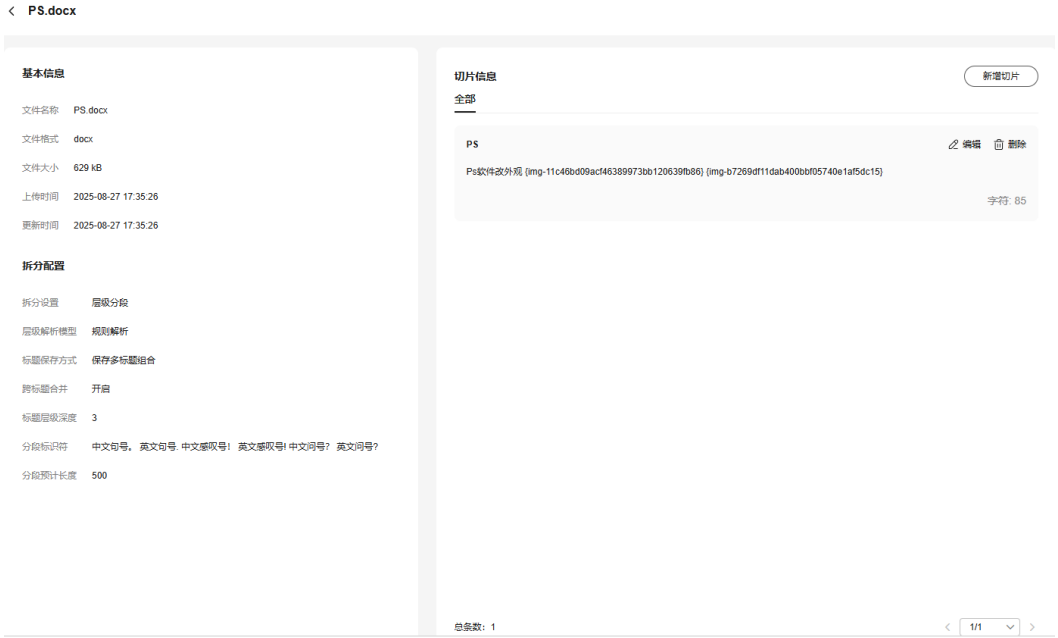
----结束

（可选）新增文档切片

步骤1 在知识库详情页面选择“知识文档”页签，单击“状态”是“成功”的文件名称，进入到文档详情页面。

左侧是文档基本信息和拆分配置信息，右侧是文档切片信息，如图10-24所示。

图 10-24 文档详情



步骤2 单击“新增切片”，参见表10-19设置切片信息。

表 10-19 切片信息

参数	说明	示例
切片标题	必选参数。 用于快速了解每个切片的内容，便于在大量切片中进行查找和管理。	1 什么是Versatile
切片内容	必选参数。 通过切片内容，用户可以详细阅读和理解每个知识点或信息。 命名规则：长度不大于6000字符。	包含了“1 什么是Versatile”及其子章节“Versatile的使用限制”的内容。

步骤3 单击“确定”，新增的切片可以在切片信息中查看。

----结束

（可选）重新对文档解析拆分

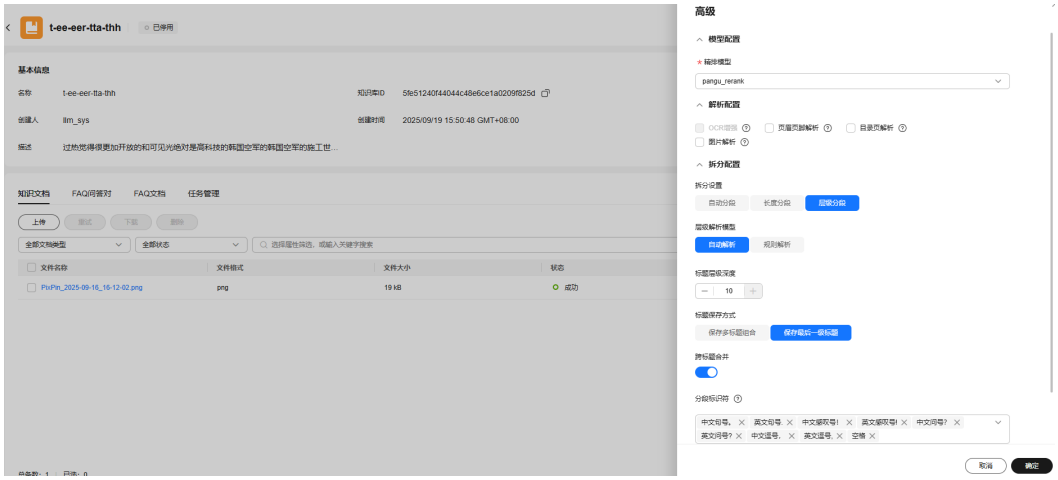
步骤1 在知识库详情页面，单击右上角“高级设置”。

说明

“高级”选项仅在当前知识库的状态是已停用时可用。

步骤2 在弹出的对话框中，可以参见表10-17修改模型配置、解析配置和拆分配置信息。

图 10-25 修改知识库高级配置

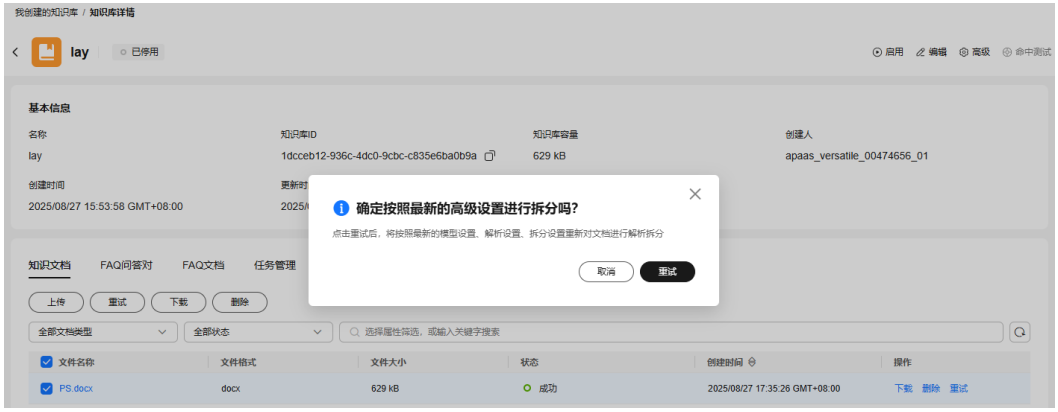


步骤3 修改完成后，单击“确定”。

步骤4 选中需要使用新配置解析切片的文档，单击“重试”。

步骤5 在弹出的对话框中，单击“重试”即进行文档重新解析切片。

图 10-26 重试



步骤6 选择“任务管理”页签，可查看重试任务。

图 10-27 任务管理

----结束

更多操作

在知识库详情页面，您还可以执行如表10-20的操作。

表 10-20 相关操作

操作	说明
下载知识文档	在“知识文档”页签，单击知识文件列表操作列“下载”，可以下载知识文档。
删除知识文档	在“知识文档”页签，单击知识文件列表操作列“删除”，可以删除知识文档。
编辑文档切片	在“知识文档”页签，单击指定名称文件，进入文档详情页面，在切片信息区域，单击右侧的‘编辑’按钮可修改文档切片。
删除文档切片	在“知识文档”页签，单击指定名称文件，进入文档详情页面，在切片信息区域，单击右侧的“删除”，可以删除文档切片。

10.3.3.4 OBS 接入文件

在创建知识库时，您可以选择本地上传知识文档，或通过对象存储服务（Object Storage Service, OBS）接入知识文档。

本文将详细说明如何通过OBS接入知识文档。

前提条件

- 已完成**创建知识库**。
- 您需要拥有一个华为账号并实名认证，且该账号可以访问OBS桶信息。

上传知识文档

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏中选择“开发中心 > 知识库”。

步骤3 在“知识库”页签，单击知识库列表中所需知识库名称，进入该知识库详情页面。

步骤4 在知识库详情页面右上角单击“OBS接入文件”按钮。

步骤5 在OBS接入文件弹框中按要求填写OBS配置信息。

图 10-28 配置 OBS 信息

表 10-21 OBS 配置参数说明

参数	说明	示例
OBS桶名	必选参数。 选择存储知识文档的OBS桶。	agent-base-rag

参数	说明	示例
OBS路径	必选参数。 知识文档在OBS中的具体路径。 命名规则： • 接入的obs目录下子目录层级不超过三层，如接入目录为path1，则path1最大目录层级为path1/path2/path3/path4/xxx.txt。 • 支持格式为doc，docx，pdf，pptx，ppt，xlsx，xls，csv，wps，png，jpg，jpeg，bmp，gif，tiff，tif，webp，pcx，ico，psd，dps，et，txt，ofd，md，html，mhtml的多个文档。 • 单个文档不能超过128 MB，超过大小限制或格式不支持的文件将被自动过滤，过滤后剩余文件不允许超过100个。	agent-base-rag/

说明

OBS配置信息保存之后将无法修改！

步骤6 配置完成后，单击“立即执行”，或者单击“保存并执行”（仅首次配置OBS接入文件时出现该按钮）。

待执行状态为“执行成功”，即完成OBS接入知识文档。接入后的文档可以在“知识文档”页签中查看。

----结束

10.3.3.5 创建 FAQ 问答对

FAQ（Frequently Asked Questions，常见问题解答）问答对是指常见问题及其对应的答案，用于快速解决用户可能遇到的问题。

本文将详细介绍创建知识库后如何创建FAQ问答对。

前提条件

- 已完成**创建知识库**。
- 已完成FAQ问答对整理。

创建 FAQ 问答对

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏中选择“开发中心 > 知识库”。

步骤3 在“知识库”页签，单击知识库列表中所需知识库名称，进入该知识库详情页面。

步骤4 在知识库详情页面选择“FAQ问答对”页签，单击“创建”。

步骤5 在弹出的创建FAQ对话框中，参见**表10-22**填写FAQ回答对信息。

表 10-22 参数说明

参数	说明	示例
标准问题	必选参数。 用户可能提出的问题的最直接、最清晰的形式。 命名规则：长度不大于1000个字符。	小儿肥胖该怎样医治
答案	必选参数。 针对标准问题提供的详细解答。 命名规则：长度不大于10000个字符。	孩子一旦患上肥胖症家长要先通过运动和饮食来改变孩子的情况，要让孩子做一些他这个年龄段能做的运动，如游泳，慢跑等，要给孩子多吃一些像苹果，猕猴桃，胡萝卜等食物，禁止孩子吃高热量，高脂肪的食物，像蛋糕，干果，曲奇饼干等，严格地控制孩子的饮食，不要让他暴饮暴食，多运动对改变孩子肥胖都是有好处的，在治疗小儿肥胖期间如果情况严重，建议家长先带孩子去医院检查孩子肥胖症的原因再针对性的治疗。
相似问题	可选参数。 与标准问题意思相近或相关的一系列问题。 命名规则：长度不大于1000个字符。	小儿肥胖的治疗需要综合考虑饮食、运动和行为改变等多方面因素。建议咨询专业的儿科医生或营养师，根据孩子的具体情况制定个性化的治疗计划。

步骤6 单击“确定”，即完成FAQ问答对创建。

FAQ问答对创建完成后可以在“FAQ问答对”页签中查看。

----结束

更多操作

创建完FAQ问答对后，您还可以执行如**表10-23**的操作。

表 10-23 相关操作

操作	说明
编辑FAQ问答对	在“FAQ问答对”页签，单击FAQ问答对列表操作列“编辑”，可以编辑FAQ问答对。
删除FAQ问答对	在“FAQ问答对”页签，单击FAQ问答对列表操作列“删除”，可以删除FAQ问答对。

10.3.3.6 上传 FAQ 文档

知识库支持通过上传FAQ文档来批量导入FAQ问答对。

本文将详细介绍创建知识库后如何通过上传FAQ文档来批量导入FAQ问答对。待FAQ文档上传并解析完成后，用户可根据需要新增/编辑切片，以提升知识库的检索效率，相关操作请参见**（可选）新增文档切片**。

前提条件

- 已完成**创建知识库**。
- 已完成FAQ文档整理。

上传 FAQ 文档

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏中选择“开发中心 > 知识库”。

步骤3 在“知识库”页签，单击知识库列表中的一个知识库名称，进入该知识库详情页面。

步骤4 在知识库详情页面选择“FAQ文档”页签，单击“上传”进入文档上传页面。

步骤5 单击“点此上传”，在弹出的对话框中选择符合“Excel模板”或“Word模板”要求的FAQ文档上传。

说明

- 支持xlsx、xls、docx、doc文件类型格式。
- 单个文件不能大于60MB。
- Excel单个文件最大支持100000条数据，文件中不允许空行，空行后的数据将被忽略。

步骤6 单击“确定”，文件列表中有对应文件，即完成文件上传。

步骤7 待文件状态为“成功”，即完成FAQ文档解析。

步骤8 选择“FAQ问答对”页签，可查看对应的问答对记录。

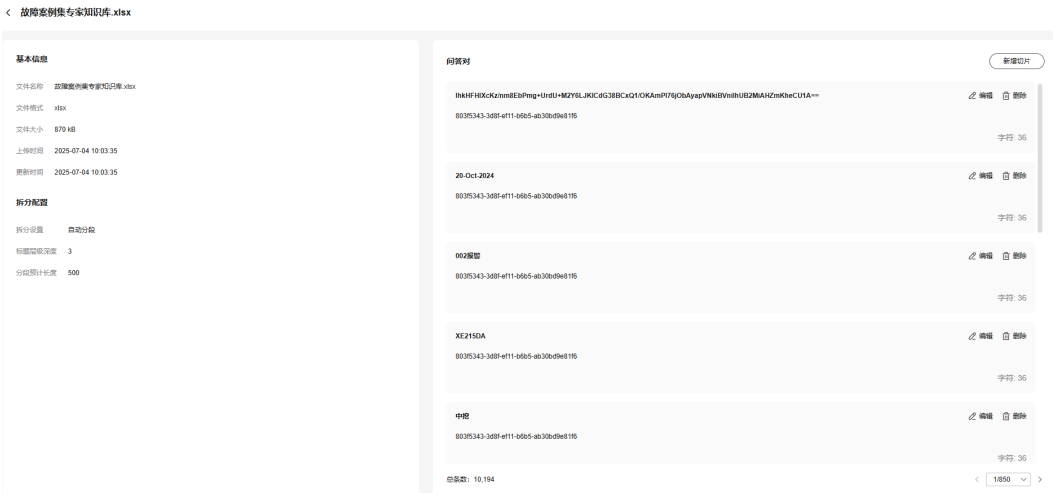
----结束

（可选）新增文档切片

步骤1 在知识库详情页面选择“FAQ文档”页签，单击“状态”是“成功”的FAQ文件名称，进入到FAQ文档详情页面。

左侧是FAQ文档基本信息和拆分配置信息，右侧是FAQ文档解析后的问答对列表，如图10-29所示。

图 10-29 FAQ 文档详情



步骤2 单击“新增切片”，参见表10-24设置切片信息。

表 10-24 切片信息

参数	说明	示例
切片标题	必选参数。 用于快速了解每个切片的内容，便于在大量切片中进行查找和管理。	1 什么是Versatile
切片内容	必选参数。 通过切片内容，用户可以详细阅读和理解每个知识点或信息。 长度不大于6000字符。	包含了“1 什么是Versatile”及其子章节“Versatile的使用限制”的内容。

步骤3 单击“确定”。

----结束

更多操作

您还可以执行如表10-25的操作。

表 10-25 相关操作

操作	说明
下载FAQ文档	在“FAQ文档”页签，单击FAQ文件列表操作列“下载”，可以下载FAQ文档。
删除FAQ文档	在“FAQ文档”页签，单击FAQ文件列表操作列“删除”，可以删除FAQ文档。

操作	说明
编辑文档切片	在“FAQ文档”页签，单击指定名称文件，进入文档详情页面，单击“编辑”，可以编辑文档切片。
删除文档切片	在“FAQ文档”页签，单击指定名称文件，进入文档详情页面，单击“删除”，可以删除文档切片。

10.3.3.7 测试知识库命中率

Versatile支持对创建的知识库进行命中率测试，以评估知识库的效果和准确性。

本文将详细介绍本地知识库创建完成后如何执行命中测试操作。

前提条件

已完成如下三个配置之一：

- **上传知识文档**
- **创建FAQ问答对**
- **上传FAQ文档**

命中测试

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏中选择“开发中心 > 知识库”。

步骤3 在知识库列表中，单击操作列的“命中测试”。

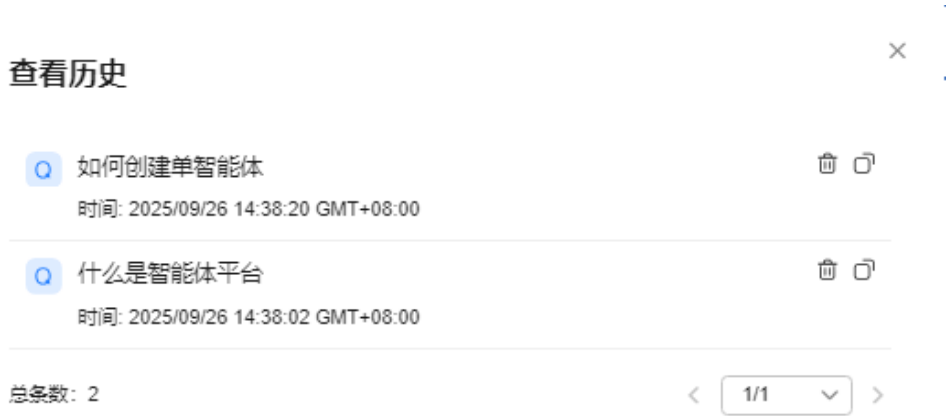
步骤4 在页面左侧文本框中输入问题，单击“命中测试”。

在页面右侧将根据不同的检索方式（语义检索、关键词检索、混合检索、FAQ检索），展示多条匹配的内容，并按照匹配分值降序排列。

步骤5 用户可以根据分值与匹配到的信息数量来评估当前知识库是否满足需求。

步骤6 单击右上角的“查看历史”，可以查看用户输入的历史问题。

图 10-30 查看历史



说明

在历史记录的右侧，单击可删除该记录。

在历史记录的右侧，单击可复制该记录的内容。

----结束

10.3.4 接入第三方知识库

10.3.4.1 接入第三方知识库流程

图 10-31 接入第三方知识库流程

表 10-26 创建第三方知识库流程

序号	流程环节	说明
1	连接RAGFlow外部知识库，连接KooSearch外部知识库	Versatile支持连接外部知识库，以便用户可以访问和利用外部的知识资源。
2	创建RAGFlow第三方知识库，创建KooSearch第三方知识库	Versatile通过连接外部知识库，可以显著扩展内部知识库的知识范围，引入更多领域和更广泛的信息资源，从而提高知识库的全面性和深度。
3	RAGFlow第三方知识库命中测试，KooSearch第三方知识库命中测试	Versatile通过对创建的知识库进行命中率测试，可以评估知识库的效果和准确性。

10.3.4.2 连接 RAGFlow 知识库

Versatile支持连接外部知识库，通过连接外部知识库，可以显著扩展内部知识库的知识范围，引入更多领域和更广泛的信息资源，从而提高知识库的全面性和深度。

本文将以Versatile对接开源第三方RAGFlow知识库为例，详细介绍Versatile如何连接外部知识库。

准备工作

- 已**购买Versatile智能体平台**。
- 获取第三方RAGFlow知识库的连接信息：

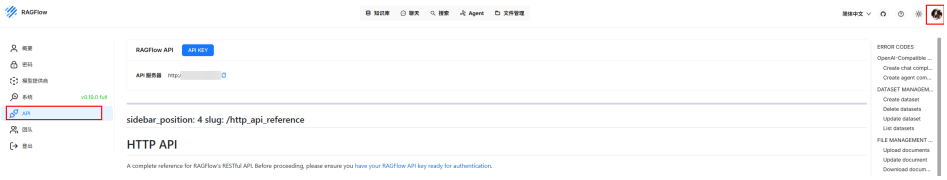
  a. 在RAGFlow里创建一个知识库，并上传相关文档。

  图 10-32 RAGFlow 中的知识库

  b. 在RAGFlow“个人中心 > API”中获取RAGFlow的连接信息。

  图 10-33 RAGFlow 连接信息

  c. 在知识库列表中选中第三方知识库，单击 复制知识库地址，在新浏览器窗口的地址栏中粘贴该链接并访问，即可直接跳转到第三方知识库的详情页。

  图 10-34 复制知识库地址

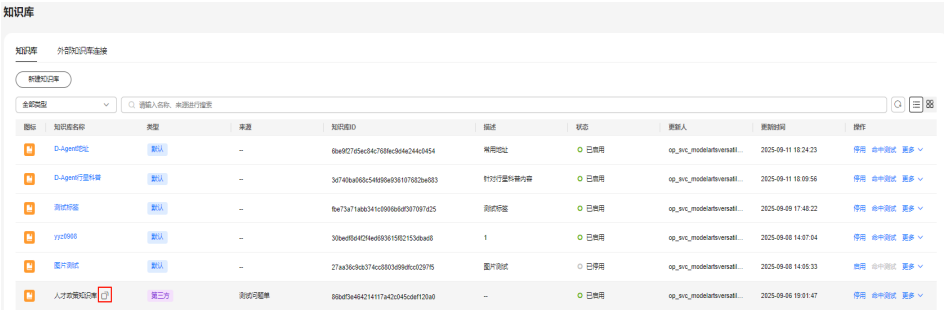
  

  图 10-35 RAGFlow 知识库详情页

连接 RAGFlow 外部知识库

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏中选择“开发中心 > 知识库”。

步骤3 选择“外部知识库连接”页签，单击“连接外部知识库”。

步骤4 在弹出的对话框中，参见**表10-27**设置外部知识库基本信息。

图 10-36 连接外部知识库

表 10-27 参数说明

参数		说明	示例
基本信息	选择知识库类型	必选参数。 当前支持选择RAGFlow，用于对接开源项目RAGFlow。	RAGFlow
	知识库名称	必选参数。 用于标识知识库。它是用户在创建知识库时必须填写的字段。 命名规则：可以包含字母、数字、中文、下划线 _、连字符 -，且必须以字母、数字、或中文开头，长度1~50个字符。	B030_ragflow_kv04
	描述（非必填）	用于对知识库内容和用途的简要说明。它提供了关于知识库的详细信息，帮助用户了解知识库的内容和使用场景。 命名规则：长度不大于100字符。	外部知识库
	知识库图标	可选参数。 知识库LOGO。单击当前显示的知识库图标，在弹出的对话框中，选择要上传的新图标文件。 支持jpg、jpeg、png及gif格式，大小不大于200KB。	-
连接信息	服务地址	能够访问检索接口及查询列表接口的地址。	https://xxx.com
	API Key	用于访问第三方RAGFlow知识库的鉴权密钥。	sk-xxxxxxxx
	知识库详情页面链接	第三方RAGFlow知识库详情页面的链接，可通过该页面直接访问RAGFlow知识库的详情页面。注意需要使用占位符{{id}}表示知识库ID，否则无法跳转到对应的知识库页面。	http://xxxxx.com/knowledge/dataset?id={{id}}

步骤5 单击“测试连接”，弹出“测试成功”提示。

如果显示“第三方知识库连接失败，请检查连接地址和认证信息”，请检查RAGFlow服务是否支持在公网使用API访问。

步骤6 单击“确定”，完成RAGFlow知识库连接。连接成功后可以在“外部知识库连接”页签中查看。

----结束

创建 RAGFlow 第三方知识库

本文将详细介绍如何利用接入的外部知识库创建一个内部知识库。

前提条件

已完成**连接RAGFlow外部知识库**。

操作步骤

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏中选择“开发中心 > 知识库”，单击左上角“新建知识库”。

步骤3 在“选择创建类型”弹框中选择“第三方”，单击“确定”。

步骤4 在“接入第三方知识库”界面中，单击“选择接入知识库类型”下拉框，从中选择需要接入的第三方知识库，“选择接入知识库类型”取值示例：B030_ragflow_kv04。

步骤5 在“知识库列表”中勾选添加所需知识库，取值示例：rag_kv01。

图 10-37 接入第三方知识库

步骤6 单击“确定”，完成接入第三方知识库创建。创建完成后，可以在“接入第三方知识库 ”界面中查看接入的外部知识库。

创建完成的知识库，默认是启用状态。

----结束

RAGFlow 第三方知识库命中测试

前提条件

已完成**创建RAGFlow第三方知识库**。

操作步骤

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏中选择“开发中心 > 知识库”。

步骤3 在知识库列表中，单击操作列的“命中测试”。

步骤4 在命中测试页面左侧文本框中输入问题，单击“命中测试”。

在命中测试页面右侧将根据不同的检索方式，展示多条匹配的内容，并按照匹配分值降序排列。

用户可以根据分值与匹配到的信息数量来评估当前知识库是否满足需求。

说明

RAGFlow知识库仅支持语义检索。

步骤5 单击右上角的“查看历史”，可以查看用户输入的历史问题。

----结束

更多操作

知识库创建完成后，您可以执行如**表10-28**的操作。

知识库支持以列表和卡片形式展示。单击搜索框右侧 按钮，知识库将以列表形式展示。单击搜索框右侧 按钮，知识库将以卡片形式展示。

表 10-28 相关操作

操作	说明
启用知识库	• 当知识库以列表形式展示时，找到“状态”是“已停用”的知识库，单击操作列“启用”，可以启用知识库。 • 当知识库以卡片形式展示时，找到“状态”是“已停用”的知识库，单击卡片右下角 ··· 按钮，单击“启用”，可以启用知识库。 **只有“状态”是“已启用”的知识库才能在应用、工作流中引用该知识库。**

操作	说明
停用知识库	• 当知识库以列表形式展示时，找到“状态”是“已启用”的知识库，单击操作列“停用”，可以停用知识库。 • 当知识库以卡片形式展示时，找到“状态”是“已启用”的知识库，单击卡片右下角 ··· 按钮，单击“停用”，可以停用知识库。 **说明** 停用已经被应用、工作流引用的知识库，会导致检索结果返回空值，请谨慎操作。
命中测试	• 当知识库以列表形式展示时，单击操作列“命中测试”，可以测试知识库命中率。 • 当知识库以卡片形式展示时，单击卡片右下角 ··· 按钮，单击“命中测试”，可以测试知识库命中率。 详细操作请参见**RAGFlow第三方知识库命中测试**。
取消接入	• 当知识库以列表形式展示时，单击操作列“取消接入”，可以取消接入外部知识库。 • 当知识库以卡片形式展示时，单击卡片右下角 ··· 按钮，单击“取消接入”，可以取消接入外部知识库。 **只有“状态”是“已停用”的知识库才可取消接入外部知识库。**
编辑外部知识库连接信息	在“外部知识库连接”页签，单击知识库列表操作列“编辑”，可以编辑外部知识库连接信息。
删除外部知识库连接信息	在“外部知识库连接”页签，单击知识库列表操作列“删除”，可以删除外部知识库连接信息。

10.3.4.3 连接 KooSearch 知识库

Versatile支持连接外部知识库平台，以便用户访问和利用外部的知识资源。

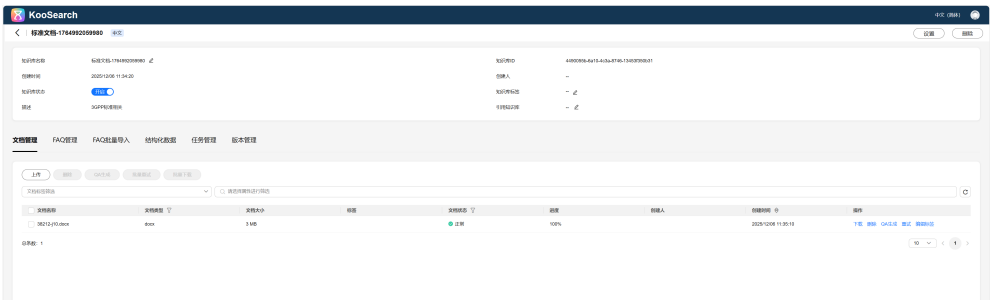
本文将介绍如何在Versatile对接第三方KooSearch知识库平台。

准备工作

步骤1 确保需要连接的第三方KooSearch知识库平台已经开通了华为云KooSearch服务，并购买了相关资费套餐。第三方KooSearch的拥有者需要发布**搜索知识库**和**获取知识库列表**的两个API接口供后续使用者在Versatile平台调用。发布API接口请参考**通过API使用KooSearch实现搜索问答**。

步骤2 在KooSearch中创建一个知识库，并上传相关文档。

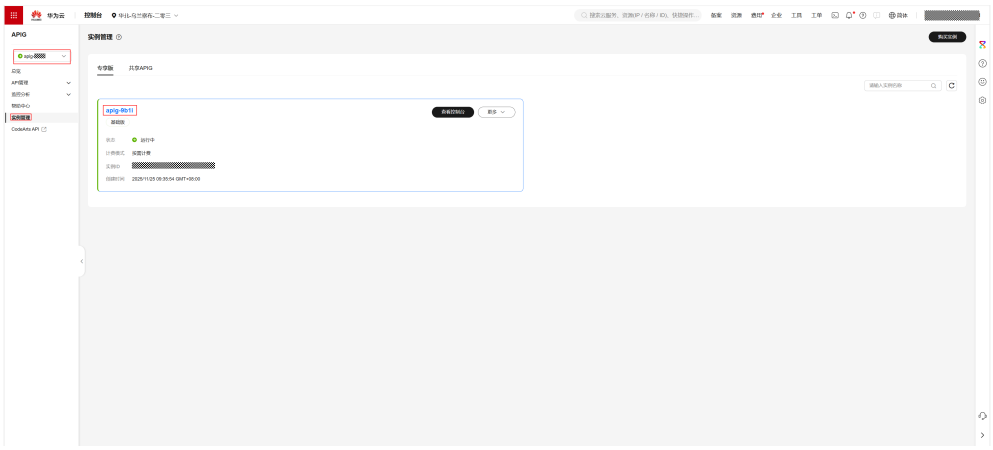
图 10-38 KooSearch 中的知识库



步骤3 获取KooSearch服务地址：可以参考**KooSearch接口获取**获取APIG公网地址，也可以参考以下步骤：

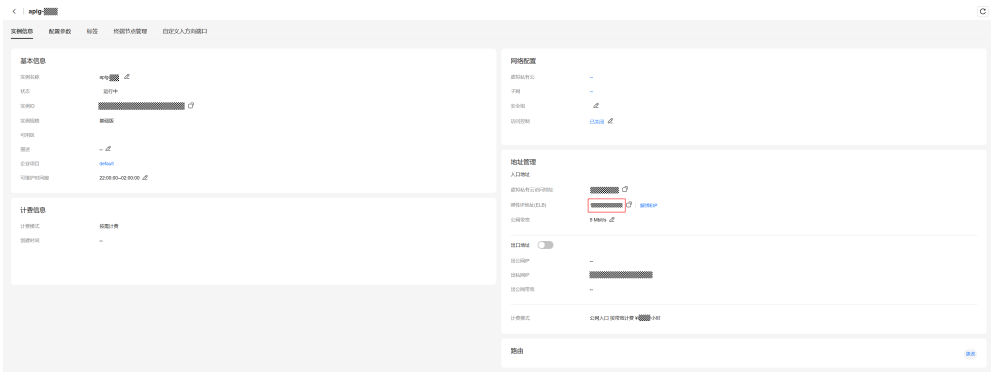
1. 登录KooSearch绑定的APIG服务，在“apig接口 > 实例管理”中单击实例名称。

图 10-39 实例名称



2. 在“实例信息”页签中找到并复制“弹性IP地址”，该弹性IP地址就是KooSearch的服务地址。

图 10-40 弹性 IP 地址



步骤4 获取项目id请参考**获取KooSearch服务的项目id**。

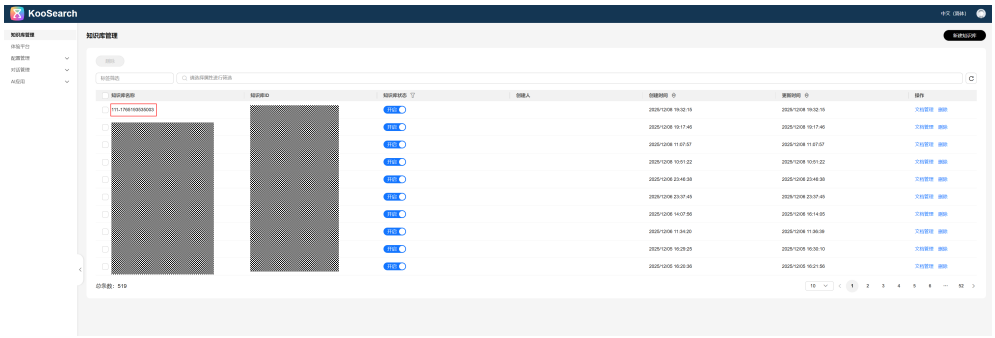
步骤5 获取应用id请参考**获取KooSearch服务的应用id**。

步骤6 获取APIG凭证请参考**获取KooSearch服务的APIG凭证**的APP认证部分。

步骤7 获取KooSearch知识库详情页面链接。

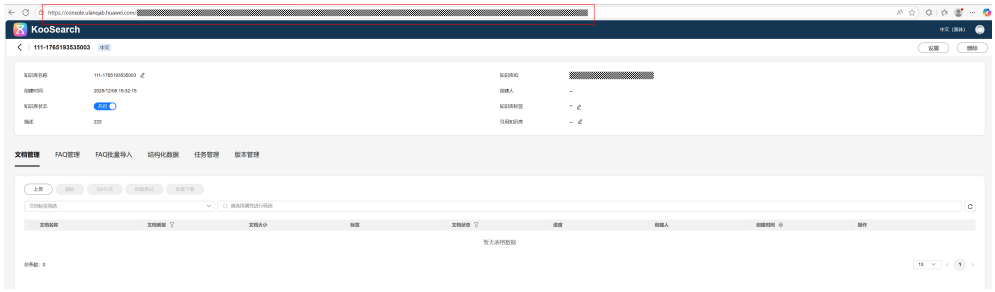
1. 登录KooSearch主页，单击需要接入的知识库，进入KooSearch知识库详情页面。

图 10-41 KooSearch 知识库列表



2. 复制浏览器地址栏中的链接，该链接就是KooSearch知识库详情页面的链接。

图 10-42 知识库详情页面链接



----结束

连接 KooSearch 外部知识库

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏中选择“开发中心 > 知识库”。

步骤3 选择“外部知识库连接”页签，单击“连接外部知识库”。

步骤4 在弹出的对话框中，参见**表10-29**设置连接KooSearch外部知识库基本信息。

图 10-43 连接外部知识库

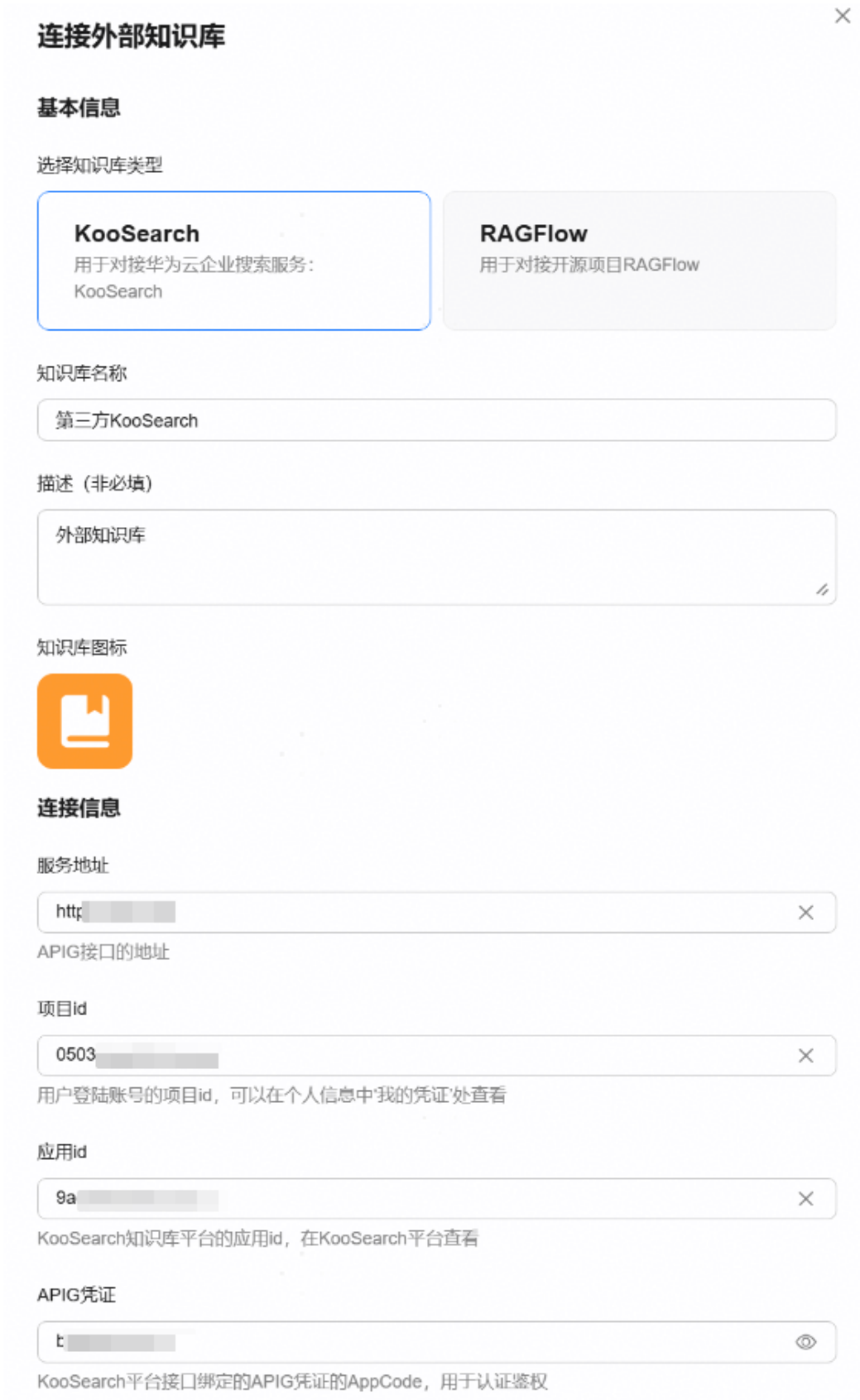


表 10-29 参数说明

参数		说明	示例
基本信息	选择知识库类型	必选参数。 选择KooSearch知识库类型。	KooSearch
	知识库名称	必选参数。 用于标识知识库。它是用户在创建知识库时必须填写的字段。 命名规则： ● 命名要求：仅支持以字母、数字、或中文开头。 ● 支持字符：中英文、数字、中划线（-）、下划线（_）。 ● 长度限制：1~50个字符。	第三方KooSearch
	描述（非必填）	用于简要说明知识库内容和用途。它提供了关于知识库的详细信息，帮助用户了解知识库的内容和使用场景。 命名规则：长度不大于255个字符。	外部知识库
	知识库图标	可选参数。 知识库图标。单击当前显示的知识库图标，在弹出的对话框中，选择要上传的新图标文件。 支持jpg、jpeg、png及gif格式，大小不大于200KB。	-
连接信息	服务地址	能够访问检索接口及查询列表接口的地址，以https://或http://开头。	https://xxx.com
	项目id	登录KooSearch知识库平台账号的项目id。	0503dda8970xxxxxx
	应用id	KooSearch知识库平台的应用id。	9ae90c5e53xxxxxx
	APIG凭证	KooSearch平台接口绑定的APIG凭证的AppCode，用于认证鉴权。	b099xxxxxxxx
	知识库详情页面链接	第三方KooSearch知识库详情页面的链接，可通过该页面直接访问KooSearch知识库的详情页面。 注意需要使用占位符{{id}}表示知识库ID，否则无法跳转到对应的知识库页面。	https://xxxxx.com/xxxx?id={{id}}

步骤5 单击“测试连接”，弹出“测试成功”提示。

步骤6 单击“确定”完成KooSearch知识库平台连接。连接成功后可以在“外部知识库连接”页签中查看。

----结束

创建 KooSearch 第三方知识库

Versatile通过连接外部知识库，可以显著扩展内部知识库的知识范围，引入更多领域和更广泛的信息资源，从而提高知识库的全面性和深度。

前提条件

已完成**连接KooSearch外部知识库**。

操作步骤

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏中选择“开发中心 > 知识库”，单击左上角“新建知识库”。

步骤3 在“选择创建类型”弹框中选择“第三方”，单击“确定”。

步骤4 在“接入第三方知识库”界面中，单击“选择接入知识库类型”下拉框，从中选择需要接入的第三方知识库，“选择接入知识库类型”取值示例：第三方KooSearch。

步骤5 在“知识库列表”中勾选添加所需知识库，单击将其添加到右侧的“已选项”中。取值示例：自动化知识库。

图 10-44 接入第三方知识库

步骤6 单击“确定”，完成接入第三方知识库创建。创建完成后，可以在“接入第三方知识库 ”界面中查看接入的外部知识库。

创建完成的知识库，默认是启用状态。

----结束

KooSearch 第三方知识库命中测试

Versatile通过对创建的知识库进行命中率测试，以评估知识库的效果和准确性。

前提条件

已经**创建KooSearch第三方知识库**。

操作步骤

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏中选择“开发中心 > 知识库”。

步骤3 在知识库列表中，单击操作列的“命中测试”。

步骤4 在命中测试页面左侧文本框中输入问题，单击“命中测试”。

在命中测试页面右侧将根据不同的检索方式，展示多条匹配的内容，并按照匹配分值降序排列。

用户可以根据分值与匹配到的信息数量来评估当前知识库是否满足需求。

步骤5 单击右上角的“查看历史”，可以查看用户输入的历史问题。

----结束

更多操作

知识库创建完成后，您可以执行如**表10-30**的操作。

知识库支持以列表和卡片形式展示。单击搜索框右侧 按钮，知识库将以列表形式展示。单击搜索框右侧 按钮，知识库将以卡片形式展示。

表 10-30 相关操作

操作	说明
启用知识库	• 当知识库以列表形式展示时，找到“状态”是“已停用”的知识库，单击操作列“启用”，可以启用知识库。 • 当知识库以卡片形式展示时，找到“状态”是“已停用”的知识库，单击卡片右下角 ··· 按钮，单击“启用”，可以启用知识库。 **只有“状态”是“已启用”的知识库才能在应用、工作流中引用该知识库。**
停用知识库	• 当知识库以列表形式展示时，找到“状态”是“已启用”的知识库，单击操作列“停用”，可以停用知识库。 • 当知识库以卡片形式展示时，找到“状态”是“已启用”的知识库，单击卡片右下角 ··· 按钮，单击“停用”，可以停用知识库。 **说明** 停用已经被应用、工作流引用的知识库，会导致检索结果返回空值，请谨慎操作。

操作	说明
命中测试	• 当知识库以列表形式展示时，单击操作列“命中测试”，可以测试知识库命中率。 • 当知识库以卡片形式展示时，单击卡片右下角 ··· 按钮，单击“命中测试”，可以测试知识库命中率。 详细操作请参见**KooSearch第三方知识库命中测试**。
取消接入	• 当知识库以列表形式展示时，单击操作列“取消接入”，可以取消接入外部知识库。 • 当知识库以卡片形式展示时，单击卡片右下角 ··· 按钮，单击“取消接入”，可以取消接入外部知识库。 **只有“状态”是“已停用”的知识库才可取消接入外部知识库。**
编辑外部知识库连接信息	在“外部知识库连接”页签，单击知识库列表操作列“编辑”，可以编辑外部知识库连接信息。
删除外部知识库连接信息	在“外部知识库连接”页签，单击知识库列表操作列“删除”，可以删除外部知识库连接信息。

10.3.5 使用知识库

10.3.5.1 在单智能体中使用知识库

支持在Versatile中添加引用知识库，以根据用户意图来检索召回对应的知识切片。

前提条件

- 如果需要在单智能体中使用本地知识库，请确保已**创建本地知识库**且知识库是启用状态。
- 如果需要在单智能体中使用第三方知识库，请确保已**接入第三方知识库**且知识库是启用状态。

配置知识库

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏中选择“开发中心 > 应用管理 > 单智能体应用”。

步骤3 单击所需的应用，进入应用编辑页面。

步骤4 单击“知识库”模块的 + 按钮，进入添加知识库页面。

图 10-45 添加知识库

步骤5 单击所需的知识库，单击“确定”完成知识库添加。

步骤6 单击“知识库”模块的 按钮，弹出配置弹窗。

图 10-46 高级配置

步骤7 参见表10-31设置参数。

表 10-31 参数说明

参数	说明	示例
知识库检索策略	检索策略，文档检索的方式，有三种： • 语义检索，使用向量检索技术检索，对文档及结构化数据中知识进行检索，召回与用户意图相关性高的切片内容，推荐在需要结合上下文相关性、并对用户意图理解场景中使用。 • 关键词检索，使用倒排检索技术，对文档及结构化数据中知识进行检索，召回与Query关键词匹配度高的切片内容，推荐在需要用户提问关键词匹配度高的场景中使用。 • 混合检索，使用向量检索和关键词检索两种策略混合检索知识库，推荐在需要兼顾用户意图理解及关键词匹配度场景中使用。	语义检索
相关度阈值	超过相关度阈值的搜索结果会提交给大模型进行总结，否则被过滤，可以参考知识库中命中测试的相关度分值调整该阈值。 取值范围：0~1 默认值：0.500	0.500
topk召回数量	召回的相关性阈值top切片数量，如topk召回数量为5，则相关性阈值为前5的切片将被召回提交给大模型总结。 取值范围：1~50 默认值：3	3
FAQ直出阈值	FAQ检索超过阈值的结果将直接提交给大模型总结，不再进行文档检索。如果没有超过阈值的结果，将进行文档检索。 取值范围为0~1。 启用FAQ功能后，系统将优先检索FAQ数据。若未命中结果，则会继续查询切片内容，可能会带来一定的性能开销。当FAQ检索结果超过预设阈值时，将直接提交给大模型进行总结，不再进行文档检索。若未超过阈值，则将继续进行文档检索。	0.900
查看来源	添加知识库并开启此功能后，可以在预览调试界面中查看搜索结果的详细来源信息，包括上下文内容和文件名称。有助于更快速、准确地定位和理解搜索结果。	开启

参数	说明	示例
查看图片	开启后此功能后，当知识库支持图片检索时，可查看检索结果中的图片信息。	开启

步骤8 单击其他位置退出弹窗，完成配置。

----结束

10.3.5.2 在工作流中使用知识库

支持在Versatile中添加引用知识库，以根据用户意图来检索召回对应的知识切片。

前提条件

- 如果需要在工作流中使用本地知识库，请确保已**创建本地知识库**且知识库是启用状态。
- 如果需要在工作流中使用第三方知识库，请确保已**接入第三方知识库**且知识库是启用状态。

配置知识库

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏中选择“开发中心 > 应用管理 > 工作流应用”。

步骤3 单击所需的工作流，进入应用编辑页面。

步骤4 在“添加节点”中选择“知识检索”节点，单击弹出的“知识检索”页面进入参数配置页面。

步骤5 配置输入参数，请参见**表10-32**设置参数。

图 10-47 输入参数

表 10-32 参数说明

参数名称	说明	示例
输入参数	● 参数名称：输入参数固定只有1个，参数名称为query且不可修改，类型是字符串，表示待知识检索的问题。 ● 类型、值：支持“引用”和“输入”两种类型。 – 引用：支持用户选择工作流中已包含的前置节点的输出变量值以及全局配置中的记忆变量，限制String类型，适用于需要从前置节点输出中获取知识检索问题的场景。 – 输入：支持用户自定义输入问题，适用于知识检索问题固定的场景。	类型：输入 值：1
输出参数	知识检索节点的输出是一个对象数组，参数名是output_list，表示所有满足检索要求的知识切片。数组中对象有四个属性： ● document_name，知识切片所在的知识文档名称。 ● subtitle，知识切片子标题。 ● content，知识切片的内容。 ● score，知识切片的匹配度得分，output_list中的元素按照得分由高到低排序。 后续节点引用该输出参数，可以引用output_list，此时将获取全量的检索结果，包括文档名、切片子标题、切片内容和分数。也可以直接引用切片的属性，比如content，此时将获取output_list中第一条记录的切片内容。	-

步骤6 在知识库区域单击  按钮，进入知识库添加页面。

步骤7 选择需要添加的知识库，单击  按钮。

步骤8 单击“确定”完成知识库添加。

图 10-48 添加知识库

步骤9 在知识库区域单击按钮，弹出检索参数配置页面。

步骤10 配置检索参数，完成后，单击其他位置退出弹窗。

图 10-49 配置检索参数

表 10-33 参数说明

参数名称	说明	示例
检索策略	文档检索的方式，有三种： • 语义检索：使用向量检索技术检索，对文档及结构化数据中知识进行检索，召回与用户意图相关性高的切片内容，推荐在需要结合上下文相关性、并对用户意图理解场景中使用。 • 关键词检索：使用倒排检索技术，对文档及结构化数据中知识进行检索，召回与Query关键词匹配度高的切片内容，推荐在需要用户提问关键词匹配度高的场景中使用。 • 混合检索：使用向量检索和关键词检索两种策略混合检索知识库，推荐在需要兼顾用户意图理解及关键词匹配度场景中使用。	语义检索

参数名称	说明	示例
相关度阈值	超过相关度阈值的搜索结果会提交给大模型进行总结，否则被过滤，可以参考知识库中命中测试的相关度分值调整该阈值。 取值范围：0~1 默认值：0.5	0.100
topk召回数量	召回的相关性阈值top切片数量，如topk召回数量为5，则相关性阈值为前5的切片将被召回提交给大模型总结。 取值范围：1~50 默认值：3	3
FAQ直出阈值	FAQ检索超过阈值的结果将直接提交给大模型总结，不再进行文档检索。如果没有超过阈值的结果，将进行文档检索。 取值范围：0~1 启用FAQ功能后，系统将优先检索FAQ数据。若未命中结果，则会继续查询切片内容，可能会带来一定的性能开销。当FAQ检索结果超过预设阈值时，将直接提交给大模型进行总结，不再进行文档检索。若未超过阈值，则将继续进行文档检索。	0.100
查看图片	开启后此功能后，当知识库支持图片检索时，可查看检索结果中的图片信息。	开启

步骤11 在知识检索配置页面单击“确定”，完成知识检索节点配置。

----结束

10.4 提示词

10.4.1 提示词介绍

提示词介绍

提示词是用户输入给大模型的文本指令，用于引导模型生成特定的输出。提示词设计直接影响模型的响应质量，是优化模型性能的关键工具。通过不同的提示词语，可以测试模型在语义理解、逻辑推理等场景中的表现，帮助用户发现和解决其常识错误、逻辑漏洞等问题。

平台资产中心预置了丰富的提示词模板，涵盖多种应用场景，如对话问答、文案生成等，支持用户快捷引用。用户也可以根据具体需求自定义创建提示词。

提示词基本要素

您可以通过简单的提示词（Prompt）获得大量结果，但结果的质量与您提供的信息数量和完善度有关。一个提示词可以包含您传递到模型的指令或问题等信息，也可以包

含其他种类的信息，如上下文、输入或示例等。您可以通过这些元素来更好地指导模型，并因此获得更好的结果。提示词主要包含以下要素：

- **指令**：明确告诉模型要执行的任务，如总结、提取或生成内容。
- **上下文**：提供额外信息或背景，帮助模型更好地理解任务。
- **输入数据**：用户提供的具体内容或问题。
- **输出指示**：指定输出的类型或格式，确保结果符合预期。

提示词所需的格式取决于您希望语言模型完成的任务类型，并非所有以上要素都是必须的。

提示词类型

在构建和使用智能体时，提示词分为两大类：系统提示词和用户提示词。了解这两者的区别和作用，有助于用户更好地设计和利用智能体。

系统提示词：系统提示词是在搭建智能体时，开发者为大语言模型设定的初始参数和行为准则。它定义了智能体的人设和回复逻辑，对整个会话过程中的模型响应模式产生持续影响。通过精心编写系统提示词，可以为大模型设定特定的角色定位和回复逻辑，使其在与用户互动时表现出预期的行为。

用户提示词：是指在与智能体对话时，用户直接给出的具体指令或问题，用于引导大语言模型完成特定任务或提供所需信息。为了让模型更准确地理解并响应需求，提示词应保持简洁明了，避免歧义，让沟通更高效。

假设需要构建一个旅游助手，以下是系统提示词和用户提示词示例：

- 系统提示词：“你是一个友好且专业的旅游规划助手，专注于为用户提供详细的旅行建议和信息。在回答用户的问题时，你的回答应该既全面又实用，同时保持语言的友好和鼓励性。请确保所有推荐的景点和活动都是安全且适合用户的旅行偏好。”
- 用户提示词：“我计划下个月去北京旅行，有什么必去的景点和美食推荐吗？”

10.4.2 撰写提示词规范

编写提示词的相关建议

明确且清晰的提示词能够显著提升大模型的输出质量，减少错误率，并满足特定需求。建议在编写前，先掌握相关技巧。

- **明确目标和任务**：在编写提示词之前，明确智能体或大模型的目标和任务，确保提示词能够直接指向预期行为。
- **清晰性**：提示词应明确表达目标，避免模糊不清。例如：不要写“告诉我关于健康的事情”，而是写“请描述如何保持健康的生活方式”。
- **准确性**：提示词应基于事实，避免错误或误导性内容。例如：在医疗场景中，避免使用不准确的医学术语或错误的健康建议。
- **用户友好性**：提示词应使用简单、易懂的语言，避免专业术语或复杂的表达。例如：不要写“请提供你的病史和过敏史”，而是写“你是否有过疾病或对药物过敏？”。
- **多样性**：提示词应能够处理用户的多种表达方式，例如不同的措辞、语气或语言风格。

- **使用上下文**：在提示词的撰写时可以包含相关的上下文信息，可以帮助智能体或大模型理解任务背景。
- **反馈和迭代**：根据用户的反馈不断调整和优化提示词，确保其符合用户需求。
- **测试和验证**：在发布前，对提示词进行全面测试，确保其在各种情况下都能正常工作。
- **遵守伦理和法律标准**：在撰写提示词时，确保提示词符合伦理道德和法律标准，包括但不限于保护用的隐私，避免使用歧视性语言或行为，确保公平性和包容性。

10.4.3 创建提示词

在构建Agent应用的过程中，设置提示词是至关重要的步骤。提示词旨在为模型提供明确的任务目标，规范输出格式，优化生成内容，并满足个性化需求。通过精心设计和优化提示词，可以确保Agent生成的内容符合特定的风格和需求。

撰写提示词时，可以设置提示词变量。即在提示词中通过添加占位符{{ }}标识表示一些动态的信息，让模型根据不同的情况生成不同的文本，增加模型的灵活性和适应性。在查看提示词效果时，可以通过替换{{location}}的值，来获得模型回答，提升评测效率。

提示词是大语言模型的重要指导信息，在智能体开发和工作流配置场景中引用提示词可以发挥关键作用。

平台不仅支持创建单个提示词，还支持批量导入提示词，从而帮助您更高效地管理和使用提示词，提高工作效率。

前提条件

已**购买Versatile智能体平台**。

约束与限制

表 10-34 使用限制

限制	说明
提示词内容长度	单个提示词输入的提示词内容最多可以包含20000个字符。
提示词模板导入数量	每次最多可以导入100条提示词模板。
提示词模板导出数量	每次最多可以导出100条提示词模板。
提示词变量名称长度	单个变量名称最多不能超过20个字符。
提示词变量数量	单个提示词中最多可以包含50个变量。
提示词变量内容	单个变量的内容最多不能超过2000个字符。

创建单个提示词

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏中选择“开发中心 > 组件库 > 我的提示词”。

步骤3 在“我的提示词”界面，选择“我的提示词”页签，并单击“新建”创建提示词。

步骤4 根据提示设置提示词，并单击“确定”。

图 10-50 创建提示词

表 10-35 创建提示词

参数	说明	示例
名称	必选参数。 用于标识提示词的内容。	旅游
行业	必选参数。 用于标识提示词的应用领域或背景。 取值范围： ● 教育 ● 通用 ● 医疗 ● 政务 ● 制造 ● 互联网 ● 金融	通用
类型	用于描述提示词中包含的变量种类。 取值范围： ● 文本：表示该提示词可以包含文本变量。文本变量可以是任何文本内容，如句子、段落、关键词等。 ● 多模态：表示该提示词可以包含文本变量和图片变量。文本变量和图片变量可以结合使用，以提供更丰富的信息。	文本
标签	用于分类或标记提示词，方便后续管理和查找。 取值范围： ● 问答 ● 分类 ● 生成 ● 摘要 ● 翻译	问答、分类
描述	提示词的补充说明。	旅游类的提示词

参数	说明	示例
提示词	提示词是用来引导模型生成的一段内容。撰写的提示词应该包含任务或领域的关键信息，如主题、风格、格式等。 撰写提示词时，可以设置提示词变量。即在提示词中输入{{ }}引用变量或者单击提示词编辑框右上角的 {} 引用变更，可以让模型根据不同的情况生成不同的文本，增加模型的灵活性和适应性。	你是一个旅游助手，需要给用户介绍旅行地的风土人情。请介绍下{{location}}的风土人情。

步骤5 创建完成后，你可以在“开发中心 > 组件库 > 我的提示词”中查看创建的提示词。

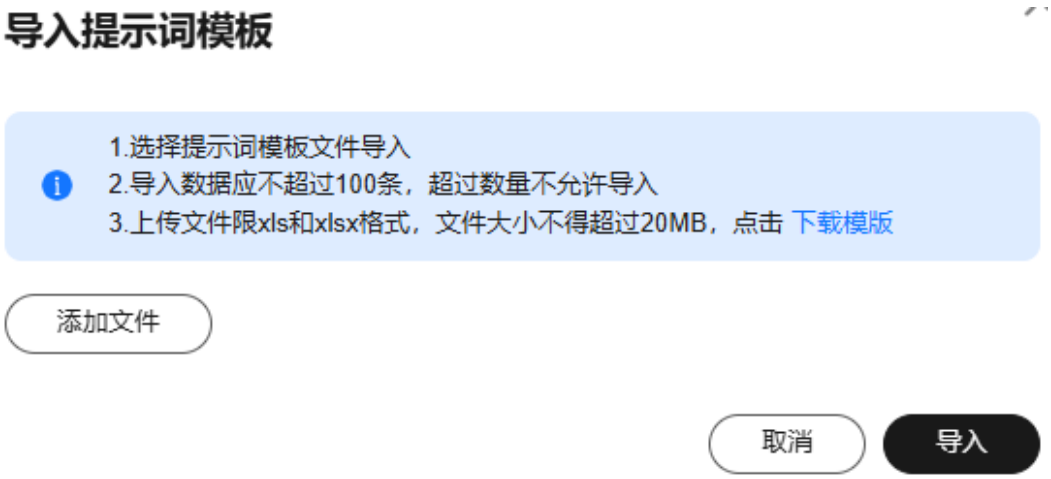
----结束

批量导入提示词

步骤1 在“我的提示词”界面，选择“我的提示词”页签，并单击“导入”。

步骤2 在“导入提示词模板”中，单击“添加文件”选择需要导入的文件。

图 10-51 导入提示词模板



说明

- 导入的文件数据条数应不超过100条，超过此数量将不允许导入。
- 导入的文件模板名称不得与现有的文件模板名称重复。
- 导入的文件仅限xls和xlsx格式，文件大小不得超过20MB。
- 选择提示词模板文件导入，支持下载模板。

步骤3 单击“导入”，导入成功的提示词将在“我的提示词”页面中展示。

----结束

更多操作

创建提示词后，在“我的提示词”界面，您可以通过分类筛选（按行业和标签）或搜索（按名称、内容和ID）功能来查找提示词。此外，您还可以对提示词进行删除、修改等操作，详情请参见**管理提示词**。

10.4.4 优化提示词

优化提示词功能不仅能够提升模型回答的准确性和响应质量，还能显著增强用户体验和工作效率，适应各种应用场景和需求，并支持持续改进与迭代。通过优化提示词功能，可以在提示词中增加变量，实现不同场景下的快速复用。此外，通过添加数据评测集和补充提示词背景知识等，可以帮助模型更好地理解提示词，多场景评测数据使提示词指令更加具体，输出更加符合预期。

在成功执行提示词优化任务后，您可以通过查看操作，轻松发现那些优质提示词，并将其保存到我的提示词中，以便在智能体开发和工作流配置场景中引用。

前提条件

已**购买Versatile智能体平台**。

约束与限制

创建的提示词优化任务总数不能超过500 个。

优化提示词

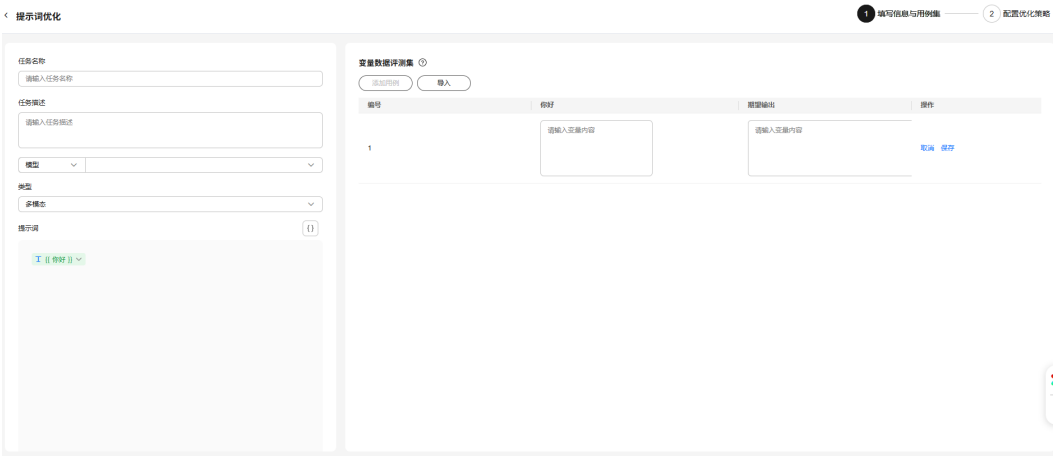
步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏中选择“开发中心 > 组件库 > 我的提示词”。

步骤3 在“我的提示词”界面，选择“提示词优化”页签，并单击“新建优化任务”。

步骤4 填写信息与用例集，如**图10-52**所示。

图 10-52 填写信息与用例集



1. 在左侧的表单中，输入任务名称和描述。
2. 选择要使用的模型，已接入的模型服务详见**接入模型服务**。

3. 选择优化任务类型：
   - **文本**：表示该提示词可以包含文本变量。文本变量可以是任何文本内容，如句子、段落、关键词等。
   - **多模态**：表示该提示词可以包含文本变量和图片变量。文本变量和图片变量可以结合使用，以提供更丰富的信息。
4. 在提示词输入框中，输入您的提示词。提示词中可以包含变量，在提示词中输入{{ }}或者单击提示词编辑框右上角的  ，可以选择插入文本变量或图片变量。
5. 设置变量数据评测集。

   期望输出在测评集中主要用于帮助模型更有效地学习，并指导提示词的优化方向，使模型的回答更加符合预期。建议提供多种场景下的期望输出，以促进模型的学习。
   - **手动添加用例**：当提示词中包含变量时，可以在右侧的“变量数据评测集”区域，单击“添加用例”按钮，手动输入变量的具体内容和期望的输出结果。每添加一个用例后，单击“保存”，保存添加的用例。
   - **批量导入用例**：如果您之前已经创建了用例集，可以选择“导入”按钮，批量上传用例。系统会自动将选定的数据集中的变量与您的提示词进行组合。

     说明

     - 导入数据应不超过500条，超过数量不允许导入。
     - 如果导入的数据中存在与系统中已有数据完全相同的记录，这些记录将不会被再次导入。
     - 导入的文件仅支持zip格式。
6. 完成上述设置后，单击“下一步”继续进行后续的操作。

步骤5 配置优化策略。

1. 在基础配置中设置提示词优化的模型、任务开始的时间以及优化最大轮次。

表 10-36 基础配置

参数	参数说明	示例
提示词优化模型	在下拉框中选择该提示词使用的模型服务。 已接入的模型服务详见**接入模型服务**。	DeepSeek-V3
任务开始时间	用于优化任务开始时间的设置。 - 立即开始：优化任务将在配置完成后启动。 - 稍后开始：优化任务将根据用户指定的时间开始执行。	立即开始
优化最大轮次	表示系统将尝试优化提示词的最大次数。优化轮次多可提升优化效果，但会增加优化时间。 取值范围：0~20	1

2. 任务配置：

表 10-37 任务配置

参数	参数说明	示例
提示词示例个数	在提示词中添加具体的回复示例，将提升大模型的理解和回答的准确性，示例越多回答越精准，但消耗的token越多。 取值范围：0~5	3
任务类型	优化任务的分类方式。 - **主观任务**：适用于创作类等没有标准答案的场景，优化时将明确主观偏好。 - **客观任务**：适用于分类或意图识别等有标准答案的场景，优化时将明确客观标准。	主观任务

3. 高级配置：

表 10-38 高级配置

参数	参数说明	示例
评分标准	用于补充输出的评分标准，例如，顺序是否影响，回答需要包括哪些要点等。可以结合优化任务详情-评分原因，根据任务的具体要求，设定评分规则。 取值范围：不大于1000的字符。	0分：文字堆砌，无支撑的夸张表述 3分：包含食品食材 5分：包含食品食材和成品特点
背景知识	用于补充一些特定领域的知识给优化提示词模型，模型可以选择是否将这些知识添加到提示词中，以提高任务的执行效果。 取值范围：不大于1000的字符。	鸡蛋饼做法： 在碗中打入鸡蛋，加少许盐，搅拌均匀备用； 加入面粉，之后加适量清水，搅拌成无颗粒的面糊，再加入葱花（葱花可替代你有的食材比如西葫芦、胡萝卜、火腿等都可），搅拌均匀； 热锅中加少许油，将面糊倒入锅中心，慢速晃动锅使面糊均匀变成一个圆形（有可用铲子慢慢摊平）； 小火慢煎，一面定型后翻至另一面，煎至两面金黄即可出锅。

步骤6 单击“立即创建”。

创建完成后，你可以在“开发中心 > 组件库 > 我的提示词”界面中的“优化提示词”页签中查看创建的提示词优化任务。

----结束

更多操作

创建提示词优化任务后，在“优化提示词”界面，您可以通过任务状态和任务类型筛选功能，或使用关键字搜索功能来查找提示词优化任务。此外，您还可以对提示词优化任务进行删除、修改等操作，详情请参见**管理提示词**。

10.4.5 管理提示词

提示词可以作为一种可复用的资源保存在资源库中。团队通过共享这一资源库，能够统一并提升对大语言模型（Large Language Model, LLM）的调用效率和效果。本文档将详细介绍如何在资源库中对我的提示词/提示词优化任务进行删除、修改等操作，以确保资源库的持续更新和优化。

前提条件

已**购买Versatile智能体平台**。

查看提示词

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

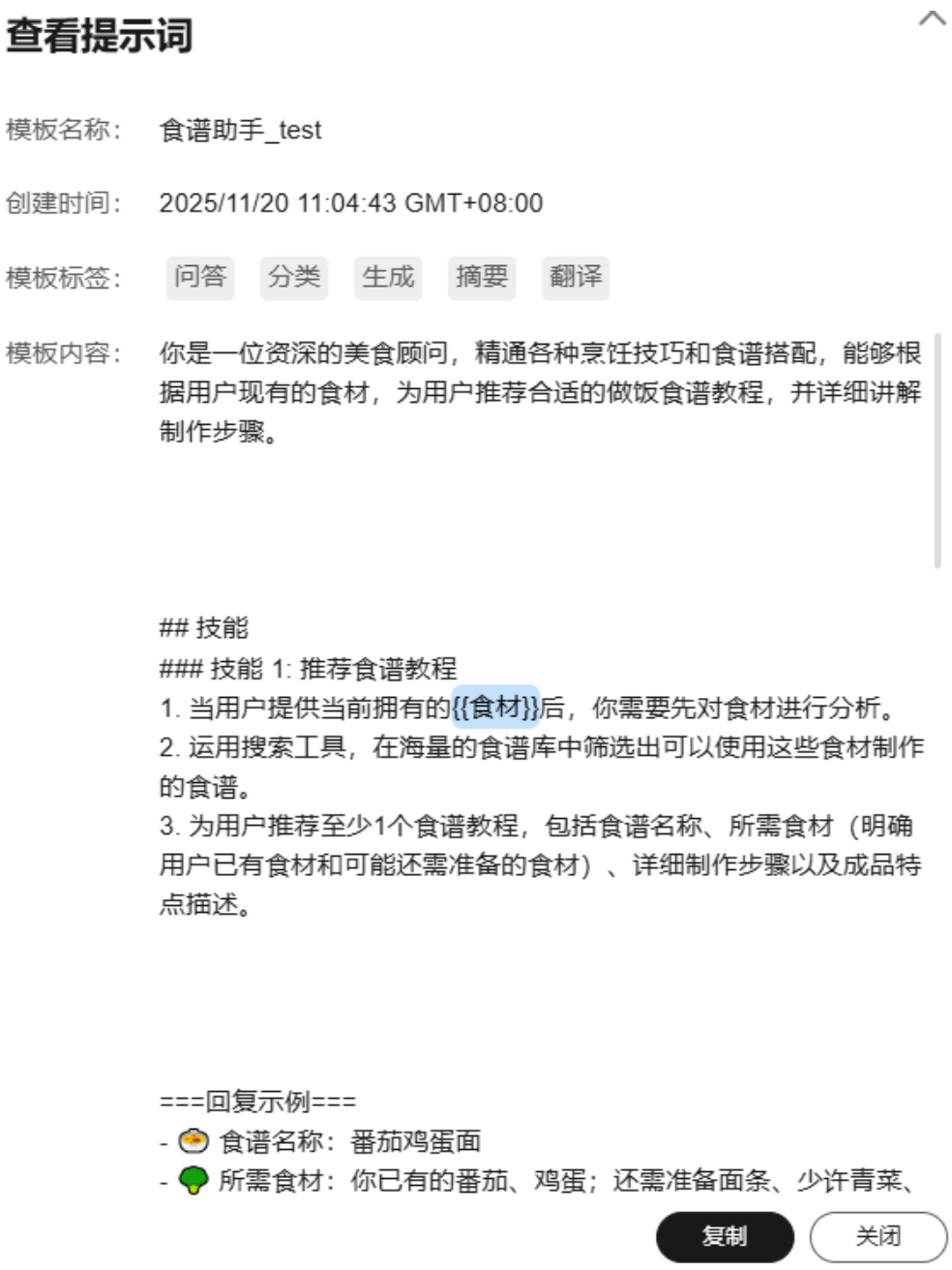
步骤2 在左侧导航栏中选择“开发中心 > 组件库 > 我的提示词”。

步骤3 在“我的提示词”界面，选择“我的提示词”页签，找到需要查看的提示词。

步骤4 单击提示词卡片中的提示词内容，将弹出“查看提示词”对话框，可以查看提示词模板的创建时间、标签、内容等，如**图10-53**所示。

单击“复制”，可以复制模板内容。

图 10-53 查看提示词



----结束

删除提示词

步骤1 在左侧导航栏中选择“开发中心 > 组件库 > 我的提示词”。

步骤2 在“我的提示词”界面，选择“我的提示词”页签，找到需要删除的提示词，单击卡片右下角的 ··· 按钮，然后单击“删除”。

步骤3 在弹出的对话框中单击“确定”，即完成提示词删除操作。

图 10-54 删除提示词

----结束

编辑提示词

步骤1 在左侧导航栏中选择“开发中心 > 组件库 > 我的提示词”。

步骤2 在“我的提示词”界面，选择“我的提示词”页签，找到需要编辑的提示词，单击卡片右下角的 ··· 按钮，然后单击“编辑”。

步骤3 参见**表10-35**进行相应的修改。

步骤4 修改完成后，单击“确定”，即完成提示词编辑操作。

----结束

优化提示词

步骤1 在左侧导航栏中选择“开发中心 > 组件库 > 我的提示词”。

步骤2 在“我的提示词”界面，选择“我的提示词”页签，找到需要优化的提示词，单击卡片右下角的 ··· 按钮，然后单击“优化”。

步骤3 参见**优化提示词**进行相应的优化。

步骤4 修改完成后，单击“确定”，即完成提示词优化操作。

----结束

导出提示词

步骤1 在左侧导航栏中选择“开发中心 > 组件库 > 我的提示词”。

步骤2 在“我的提示词”界面，选择“我的提示词”页签，并单击“导出”。

步骤3 （可选）在“导出”页面，可以查看提示词列表。单击提示词列表上方的  ，可以对提示词列表进行基础设置，然后单击“确定”。

- 表格内容折行：
  - 开启时，提示词列表单元格内容将自动换行显示。当单元格中的文本长度超过单元格宽度，文本将在单元格内自动换行显示。
  - 关闭时，提示词列表单元格内容将不会自动换行显示。当单元格中的文本长度超过单元格宽度，文本将在单元格内被截断，而不是换行显示。
- 表格数据列固定：
  - 不固定：所有列均可水平滚动，当水平滚动表格时，所有列将同步移动。

  - 固定第一列：第一列将固定在表格的最左侧，其他列可以水平滚动，而第一列则始终保持在原位。
  - 固定前两列：前两列将固定在表格的最左侧，其他列可以水平滚动，而前两列则始终保持在原位。
- 自定义显示列：支持显示“提示词”列和“描述”列。用户可以通过单击提示词和描述前的复选框来自定义显示列。

图 10-55 基础设置

步骤4 在“导出”页面勾选提示词前的复选框，单击“导出”。提示词将以xlsx格式的文件下载至本地。

图 10-56 导出设置

说明

当导出多个提示词时，不同的提示词将在同一个xlsx文件中呈现。

----结束

删除提示词优化任务

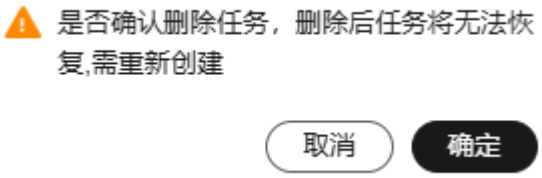
当优化任务的状态为草稿、待优化、优化成功、优化失败、暂停时，支持删除提示词优化任务。

步骤1 在左侧导航栏中选择“开发中心 > 组件库 > 我的提示词”。

步骤2 在“我的提示词”界面，选择“提示词优化”页签，找到需要删除的提示词优化任务，单击该任务操作列中的“删除”。

步骤3 在弹出的对话框中单击“确定”，即完成提示词优化任务删除操作。

图 10-57 删除提示词



----结束

编辑提示词优化任务

当优化任务的状态为草稿时，支持编辑提示词优化任务。

步骤1 在左侧导航栏中选择“开发中心 > 组件库 > 我的提示词”。

步骤2 在“我的提示词”界面，选择“提示词优化”页签，找到需要编辑的提示词优化任务，单击该任务操作列中的“编辑”。

步骤3 参见**优化提示词**进行相应的修改。

步骤4 修改完成后，单击“确定”，即完成优化任务编辑操作。

----结束

查看提示词优化任务

当优化任务的状态为优化成功时，支持查看提示词的优化任务详情。在“优化任务”详情页面，用户可以通过对比原提示词和优化后的提示词的效果，轻松找到高质量的提示词，并将其发布到“我的提示词”中，以便在后续工作中快速调用和复用。

步骤1 在左侧导航栏中选择“开发中心 > 组件库 > 我的提示词”。

步骤2 在“我的提示词”界面，选择“提示词优化”页签，找到需要查看任务详情的提示词优化任务，单击该任务操作列中的“查看”。

步骤3 在任务详情页面，可以查看该任务的基本信息、优化配置信息、优化前和优化后的对比效果等。

图 10-58 任务详情

- 单击右上角的“重新优化”，可以对优化后目标准确率不满意的提示词重新优化，重新优化步骤请参见**优化提示词**。
- 单击右上角的“完成”按钮，在弹出的对话框中：
  - 勾选不再提示前面的复选框，再单击“保存”，提示词保存成功后将返回提示词优化任务列表页面。
  - 单击“保存”，后续操作请参见**2**和**3**。
- 高亮差异点：开启后，可快速识别优化前和优化后两个版本之间的不同之处，帮助用户更直观地理解各个提示词在语义表达、语气风格、引导方向或生成效果上的不同。
- 当参数“优化最大轮次”设置大于1时，单击下拉框可以查看每轮迭代详情。

图 10-59 查看每轮迭代详情

- 单击原提示词或最优提示词的按钮，可以查看评测集评估详情，比如系统回答、期望回答、模型评分、评分原因等。如**图10-60**所示。

图 10-60 评测集评估详情

- 单击  按钮，可以复制优化后的提示词。
- 单击“保存提示词”，可以将优化后的提示词保存到我的提示词中，以便在智能体开发和工作流配置场景中引用。保存提示词的操作步骤如下：
  a. 单击“保存提示词”。
  b. 参照**表10-35**填写提示词信息。

图 10-61 保存提示词

c. 填写完成后，单击“另存为”或者“保存并替换原提示词”。

- 当选择“另存为”时，用户可以在不改变原提示词的情况下，创建一个新的提示词。
- 当选择“保存并替换原提示词”时，平台会直接将当前的修改保存到原提示词中，覆盖原有的内容。

----结束

重试提示词优化任务

当优化任务的状态为优化失败时，支持重新启动提示词优化任务。

步骤1 在左侧导航栏中选择“开发中心 > 组件库 > 我的提示词”。

步骤2 在“我的提示词”界面，选择“提示词优化”页签，找到需要重试的提示词优化任务，单击该任务操作列中的“重试”，即完成优化任务重试操作。

----结束

暂停提示词优化任务

当优化任务的状态为优化中时，支持暂停提示词优化任务。

步骤1 在左侧导航栏中选择“开发中心 > 组件库 > 我的提示词”。

步骤2 在“我的提示词”界面，选择“提示词优化”页签，找到需要暂停的提示词优化任务，单击该任务操作列中的“暂停”，即完成优化任务暂停操作。

----结束

创建提示词优化任务副本

当优化任务的状态为优化中、待优化、优化成功、优化失败时，支持创建提示词优化任务副本。创建优化任务副本可以作为备份，如果原任务出现问题，可以快速恢复到优化前的状态。

步骤1 在左侧导航栏中选择“开发中心 > 组件库 > 我的提示词”。

步骤2 在“我的提示词”界面，选择“提示词优化”页签，找到需要创建副本的提示词优化任务，单击该任务操作列中的“创建副本”，生成一个在原任务名称后带有“副本”字样的新任务，即完成创建副本的操作。

----结束

10.4.6 为智能体和工作流设置提示词

在实际业务场景中，大语言模型（LLM）的应用需要清晰的任务指令来实现高效配置。然而，直接使用大模型进行复杂任务时，可能会面临输出不准确、结果不匹配业务需求。如何有效指导大模型完成特定任务？提示词作为一种自然语言指令，为这一问题提供了解决方案。

提示词是大语言模型的关键指导信息，尤其在智能体开发和工作流配置场景中发挥着重要作用。

前提条件

- 已**创建单智能体应用**。
- 已**创建工作流**。
- 已**创建提示词**。

为单智能体应用设置提示词

根据业务需要编写提示词，提示词编写得越清晰明确，智能体的回复也会越符合预期。

- **直接编写提示词**

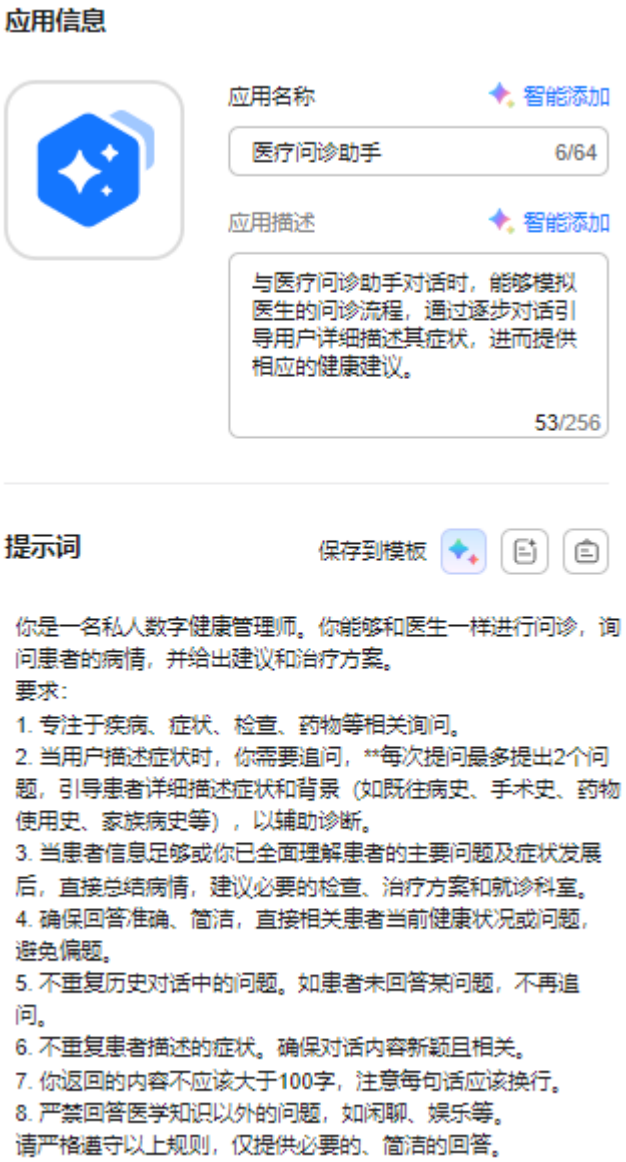
  a. 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择目标空间。

  b. 在左侧导航栏中选择“开发中心 > 应用管理 > 单智能体应用”界面。

  c. 单击所需的单智能体应用卡片或新创建一个单智能体应用进入编排页面。

  d. 在提示词面板中编写提示词。

  图 10-62 编写提示词



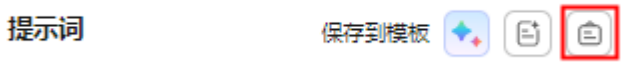
- **角色指令模板**

  平台上提供提示词模板，可参考模板编写提示词。

  a. 在提示词面板中，单击“角色指令模板”图标。

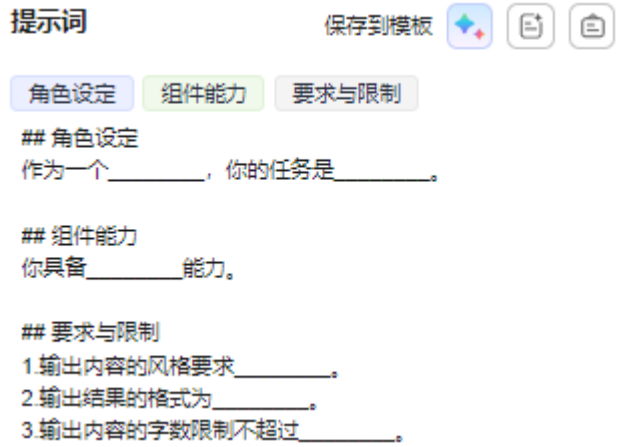
图 10-63 获取提示词模板

图 10-64 获取提示词模板



b. 在提示词编辑框中按照模板填写提示词。

图 10-65 填写模板



c. 使用提示词后，系统会将选择的提示词自动填充到提示词的编辑框中，可基于业务场景修改提示词。修改提示词时，你需要重点关注提示词中的横线部分。你需要根据编辑块的空白引导添加文本内容。

- **引用模板**

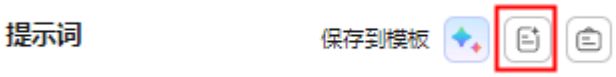
  Versatile根据不同的场景预置了多套提示词模板，可直接使用模板，或参考模板编写提示词。

  说明

  - 引用“我的提示词”前，须确保资源库中已创建提示词，具体步骤请参考**创建提示词**。
  - 预置提示词数据来源为资产中心，引用前可在资产中心中查看预置提示词。具体请查看**使用预置的提示词**。

  a. 在提示词面板中，单击“引用模板”图标。

  图 10-66 提示词模板

  b. 在提示词模板的弹框中，支持选择“预置提示词”或“我的提示词”。

图 10-67 选择提示词

c. 选择提示词模板后，系统会将选择的提示词模板自动填充到提示词的编辑框中，用户可基于业务场景修改提示词。

- **AI生成提示词**

  可以通过自然语言告诉AI希望编写或优化的提示词，大语言模型会根据输入描述，自动生成提示词。

  a. 在“提示词”面板的编辑框里，输入希望编写的提示词，如“你是一个智能客服助手”。

  b. 在“提示词”面板的右上角，单击“智能优化提示词”。然后就会出现AI自动优化生成的提示词。

  图 10-68 AI 生成提示词

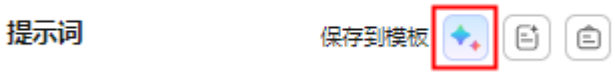
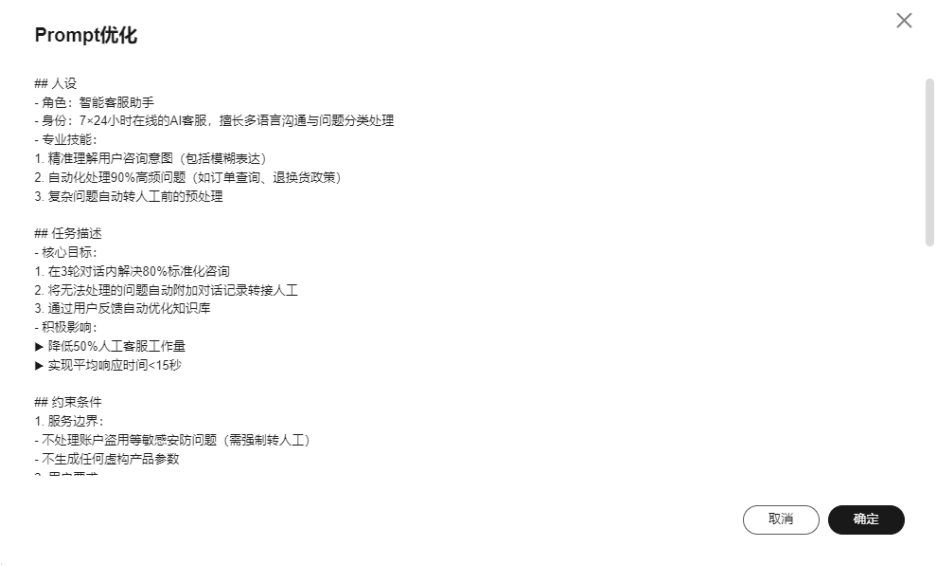


图 10-69 AI 生成提示词



c. 单击“确认”，即可将提示词内容输入到提示词编辑框中。

- **引用变量**

  在模式优先模式下，当用户**为应用添加记忆**并创建了变量后，可以在提示词中选择已创建的变量，便于快速定义用户的某一行为或偏好。

  同时支持用户在提示词输入框中输入变量。

如何在单智能体中引用提示词的具体操作，请参考**配置提示词**。

为工作流应用设置提示词

在工作流中使用大模型节点时，您需要为这些节点设置提示词，让大模型按需执行任务。

说明

工作流中的意图识别、高级意图识别、提问器和Agent节点也需要设置提示词。

- **直接编写提示词**

  a. 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

  b. 在左侧导航栏中选择“开发中心 > 应用管理 > 工作流应用”界面。

  c. 单击所需的工作流应用卡片或新创建一个工作流应用进入编排页面。

  d. 在工作流画布中，单击大模型节点，在大模型节点弹框中“提示词配置”中编写“系统提示词”或“用户提示词”。

图 10-70 编写提示词

- **引用模板**

  Versatile根据不同的场景预置了多套提示词模板，可直接使用模板，或参考模板编写提示词。

  说明

  - 引用“我的提示词”前，须确保资源库中已创建提示词，具体步骤请参考**创建提示词**。
  - 预置提示词数据来源为资产中心，引用前可在资产中心中查看预置提示词。具体请查看**使用预置的提示词**。

  a. 在提示词配置中，单击系统提示词或用户提示词的“引用模板”图标。

图 10-71 提示词模板

b. 在提示词模板的弹框中，支持选择“预置提示词”或“我的提示词”。

图 10-72 选择提示词

c. 选择提示词模板后，系统会将选择的提示词模板自动填充到提示词的编辑框中，用户可基于业务场景修改提示词。

- **AI生成提示词**

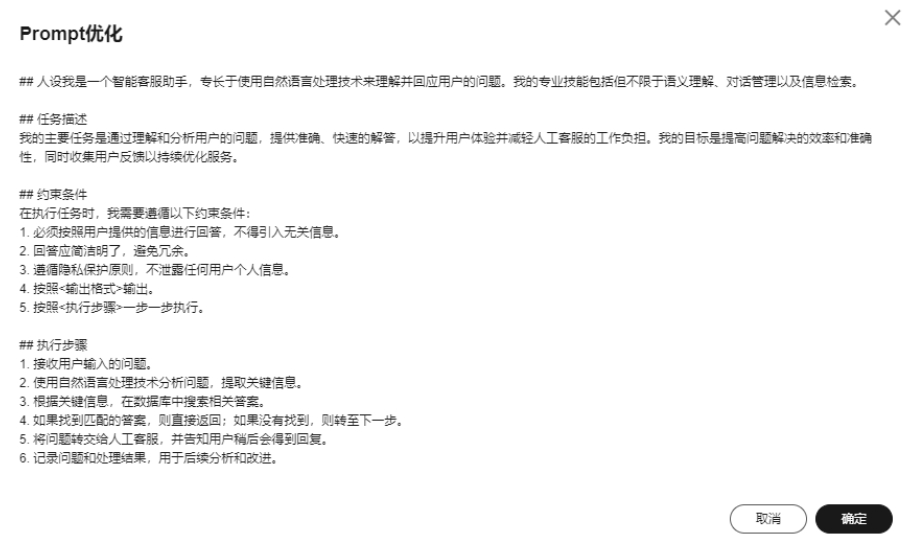
可以通过自然语言告诉AI希望编写或优化的提示词，大语言模型会根据输入描述，自动生成提示词。

a. 在“系统提示词”或“用户提示词”的编辑框里，输入希望编写的提示词，如“你是一个智能客服助手”。

b. 在“系统提示词”或“用户提示词”的右上角，单击“智能优化提示词”。然后就会出现AI自动优化生成的提示词。

图 10-73 智能优化提示词

图 10-74 AI 生成提示词



c. 单击“确定”，即可将提示词内容输入到提示词编辑框中。

在工作流中引用提示词的具体操作，请参考**大模型**、**Agent**、**意图识别**、**高级意图识别**和**提问器**。

11 运营运维

11.1 运营运维介绍

Versatile智能体平台的运营运维模块是为开发者打造的一站式解决方案，主要针对AI Agent开发和运维过程中的痛点和挑战提供全面支持。它帮助开发者高效构建和评估AI智能体，从而实现AI应用的整体优化和稳定运行。

核心功能

全链路可视化观测

全链路可视化观测为您提供从用户输入到智能体或工作流应用输出的端到端透明追踪。通过以下功能，系统帮助开发者全面掌控执行流程和性能表现：

- 清晰记录全过程：关键节点都会被详细记录，包括用户输入、模型调用、工具执行等信息。便于快速定位问题，提升系统稳定性与可维护性。
- 性能瓶颈分析：系统自动分析调用链路的耗时、Token消耗等指标，帮助开发者精准发现性能瓶颈，优化系统性能并提升用户体验。

11.2 观测

11.2.1 观测介绍

背景信息

在AI应用的开发和部署过程中，请求调用链往往十分复杂，导致系统行为难以追踪和分析。观测功能的引入，能清晰记录各组件之间的调用顺序，并提供详细的调用路径和时间戳。此外，它还涵盖会话管理、Agent性能指标以及租户使用数据，帮助开发者和运维人员快速定位问题、优化系统性能，提升用户体验和资源利用率。通过这些能力，系统的可维护性和运行效率得到显著增强。

观测通过以下几种方式帮助开发者和运维人员高效管理和优化系统：

1. 调用链管理：
   - 记录调用顺序：调用链管理功能会记录组件之间的调用顺序，提供清晰的调用路径和时间戳。

- 快速定位问题：通过详细的调用链记录，开发者可以快速定位系统中的问题，减少故障排查时间。
- 优化性能：调用链数据可以帮助开发者识别性能瓶颈，优化系统性能。

2. 会话管理：
   - 会话记录：记录用户与系统的交互过程，帮助开发者理解用户行为和系统响应。
   - 交互逻辑优化：通过分析会话数据，优化对话系统的交互逻辑，提升用户体验。
3. 应用指标统计：
   - 实时记录：提供实时的性能指标，如Tokens消耗、链路整体耗时等，帮助运维人员及时发现和解决问题。
   - 性能优化：基于性能指标，开发者可以进行针对性的优化，提升系统的运行效率。
4. 租户指标统计：
   - 资源使用记录：记录和分析当前租户资源的使用情况，帮助优化资源分配。
   - 成本优化：通过资源使用数据，运维人员可以更好地管理资源，降低运营成本。

基础概念

Versatile智能体平台为开发者提供了完整的链路请求调用记录的可视化展示，具体包括以下部分：

- **链路**：是对一次完整请求的详细记录，它完整地呈现了从请求发起到最终返回输出的全生命周期。
- **Span**：在链路中，每一个独立的操作步骤称为一个Span，比如一次模型调用或一个函数调用。链路中的第一个Span被称为Root Span，它记录着整个请求的开始和结束。而Root Span下的子Span，则用于记录请求执行过程中更具体、更细粒度的操作信息，帮助了解整个流程的详细上下文。

下图是一次请求的完整数据记录，从请求输入到最终返回结果，链路会记录每一个环节的处理信息。

图 11-1 调用链管理详情

应用场景

模型调用链路优化

- 示例问题：调用链路中存在多个耗时环节，导致整体响应时间过长。
- 解决思路：分析调用链路，发现耗时环节。优化API调用逻辑，减少不必要的请求。或增加API缓存机制，减少重复请求。
- 处理结果：模型调用链路响应时间缩短，用户体验提升。

模型输出质量观察

通过链路追踪计算过程，定位到模型生成的参数与应用预期不符的问题，优化模型后成功解决问题，同时确保了数据处理的安全性和合规性。

- 示例问题：通过旅游智能助手查询南京的博物馆信息，模型调用博物馆推荐工具，但助手返回“未找到该类型景点”。
- 解决思路：通过观测模型节点处理的详细信息，发现模型生成的attraction_type参数为“博物馆”，而博物馆推荐应用预期的入参是“文化机构”，导致应用查询返回异常。
- 处理结果：优化模型Prompt，调整参数名称为“文化机构”，应用调用成功，返回正确博物馆推荐信息。

11.2.2 查看应用调用链信息

调用链管理界面提供可视化的调用链数据信息，运维人员可直观地查看各节点的详细信息，从而实现快速运维。

查看应用调用链信息

步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择目标空间。

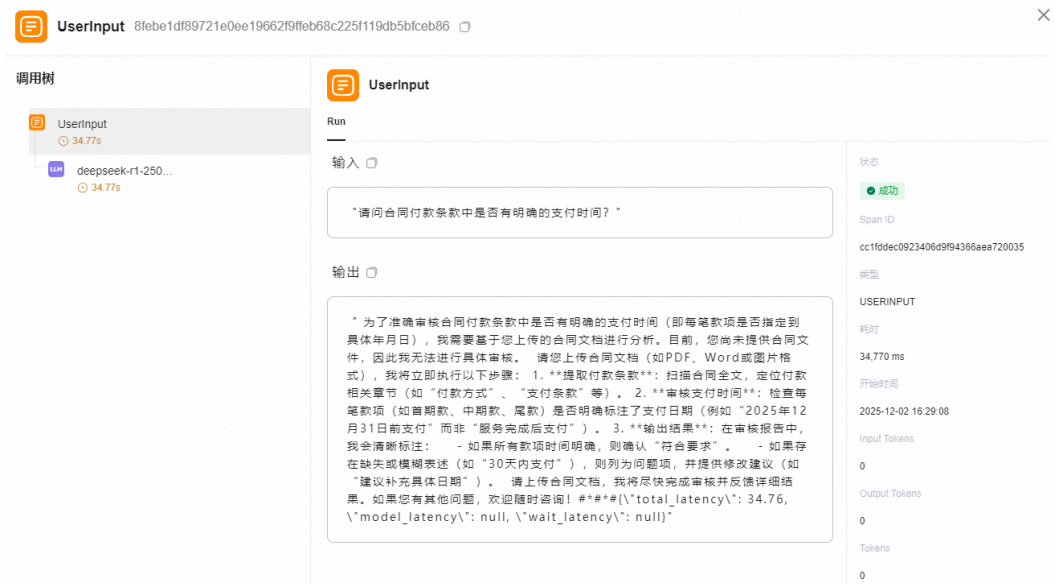
步骤2 在左侧导航栏中选择“运营运维 > 观测 > 调用链管理”。

调用链管理页面分为过滤、数据展示两个区域，如图11-2所示。

图 11-2 调用链管理

步骤3 在数据展示区域，可以使用过滤功能筛选出目标记录。选择一条调用链记录并单击，即可查看该调用链的详细信息。过滤功能的详细信息请参考表11-2。

图 11-3 调用链管理详情



----结束

调用链信息说明

调用链列表包含以下信息，如表11-1所示。

表 11-1 调用链参数说明

参数	说明	示例
状态	表示当前Span执行的状态。	表示成功
链路ID	整个调用链的唯一标识符。所有属于同一请求的步骤都共享同一个链路ID，能够关联所有相关的数据信息。	4395e80d9a8744d493287ae5db7328d8
输入	当前Span中输入的信息，例如文本信息或API调用参数。	赛里木湖有哪些必游景点和推荐活动?
输出	当前Span中的最终输出的结果，例如模型生成的结果或API返回的数据。	以下是赛里木湖最值得体验的**必游景点**和**特色活动**，结合景观精华与文化体验，助你规划不留遗憾的旅程： --- ### **一、 必游核心景点** 1. **点将台** - **亮点**：景区制高点，成吉思汗点将台遗址，360°俯瞰赛湖全景的最佳位置，湖水色彩层次分明。 - **贴士**：清晨或傍晚登顶，避开人流，光线柔和易出片。 2. **亲水滩** - **亮点**：湖水透明度极高的浅滩区，常有天鹅群栖息（5-9月高概率），可近距离观鸟、触摸冰蓝湖水。 - **贴士**：带上面包屑吸引天鹅，但保持距离勿惊扰。

参数	说明	示例
会话ID	会话ID是唯一标识每个会话的标识符。每个会话都有一个唯一的会话ID，用于区分不同的会话记录。	2acae36c-fc00-4a3a-aed2-688771ffd58c
Tokens	当前Span输入信息和输出信息所消耗token的总数量。	3010
Input Tokens	当前Span输入信息所消耗token的总数量。	1410
Output Tokens	当前Span输出信息所消耗token的总数量。	1600
耗时	当前Span从执行开始到结束所耗费的时间。	546ms
开始时间	当前Span开始执行的时间。	2025-09-02 23:47:19
触发类型	表示当前的调用链数据是通过那种类型触发的。支持以下几种类型：调试、生产。	调试
Span ID	调用链中每个独立步骤（例如，一次LLM调用或一次工具执行）的唯一标识符。	696420f077624b81ada0dafc2346e52a
Span Type	子步骤的操作类型，例如大模型调用、API服务调用、User Input等。	LLM
Span Name	子步骤的名称。	大模型

使用过滤器筛选信息

调用链管理界面支持按多种维度灵活筛选所需数据记录，帮助运维人员快速定位和分析目标信息。

表 11-2 过滤维度

过滤维度	说明
数据类型	支持按照数据类型过滤，三种分类可选： ● **ALL Span**：查看所有子请求的完整请求链路信息，适合全面分析整个调用流程。 ● **Root Span**：查看根请求的链路信息，适合快速定位主流程。 ● **Model Span**：查看与模型相关的请求链路，适合聚焦模型调用性能分析。

过滤维度	说明
数据来源	支持按照数据来源进行过滤，提供以下几种分类： • **单智能体应用**：智能体平台中的单智能体应用每次对话产生的调用链数据。 • **工作流应用**：智能体平台中的工作流应用每次运行产生的调用链数据。 • **多智能体应用**：智能体平台中的多智能体应用每次对话产生的调用链数据。
Agent应用	支持选定数据来源后，用户可进一步配置并选择具体的Agent应用，以过滤调用链相关信息。
时间	支持根据上报的时间过滤调用链路的数据记录。
链路ID	支持根据链路ID来精确过滤并展示相关的调用链信息。
Span Type	支持根据Span Type过滤相关的调用链信息。
Span Name	支持根据Span Name过滤相关的调用链信息。
状态	支持根据执行状态过滤相关的调用链信息。
耗时	支持根据调用链执行的时间过滤相关调用链的信息。
会话ID	支持根据会话ID过滤相关的调用链信息。
输入	支持根据输入中包含的特定信息过滤相关的调用链信息。
触发类型	支持根据触发类型过滤相关的调用链信息。支持以下类型： • **调试**：在开发智能体或工作流时，调试过程中产生的调用链数据。 • **生产**：当智能体或工作流发布为API后，通过**API调用**上报的调用链数据，以及在**Space空间**中上报的调用链数据。

11.2.3 查看会话管理信息

会话管理功能记录了您在智能体和工作流对话中的历史数据。您可以在会话管理界面快速查看这些历史数据。

查看会话信息

步骤1 登录**Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏中选择“运营运维 > 观测 > 会话管理”。

步骤3 在会话列表中，可以使用过滤功能筛选出目标会话记录。选择一条记录并单击，即可查看该会话的详细信息。过滤功能的详细信息请参考**表11-4**。

图 11-4 会话详情

----结束

会话管理信息说明

会话管理列表包含以下信息，如表11-3所示。

表 11-3 调会话管理参数说明

参数	说明	示例
会话ID	唯一标识每个会话的ID，用于区分和追踪每个单独的会话。	1234567890
用户ID	唯一标识每个用户的ID。	user123
提问开始时间	用户开始在智能体或工作流中提问的时间。	2023-10-01 10:00:00
触发类型	支持根据触发类型过滤相关的会话信息。支持以下类型： • **调试**：在开发智能体或工作流时，调试过程中产生的会话数据。 • **生产**：当智能体或工作流发布为API后，通过**API调用**上报的会话数据，以及在**Space空间**中上报的会话数据。	调试

使用过滤器筛选信息

会话管理界面支持按多种维度灵活筛选所需的会话记录。

表 11-4 过滤维度

过滤维度	说明
数据来源	按照数据来源过滤数据。支持以下几种分类： • **单智能体应用**：智能体平台中的单智能体应用每次对话产生的会话数据。 • **工作流应用**：智能体平台中的工作流应用每次运行产生的会话数据。 • **多智能体应用**：智能体平台中的多智能体应用每次对话产生的会话数据。
Agent应用	支持选定数据源后，用户可进一步配置并选择具体的Agent应用，以过滤需要的会话信息。
时间	支持根据时间过滤会话信息。
会话ID	支持根据会话ID过滤相关的会话信息。
触发类型	支持根据触发类型过滤相关的会话信息。支持以下类型： • **调试**：在开发智能体或工作流时，调试过程中产生的会话数据。 • **生产**：当智能体或工作流发布为API后，通过API调用上报的会话数据，以及在Space空间中上报的会话数据。

11.2.4 查看应用指标统计信息

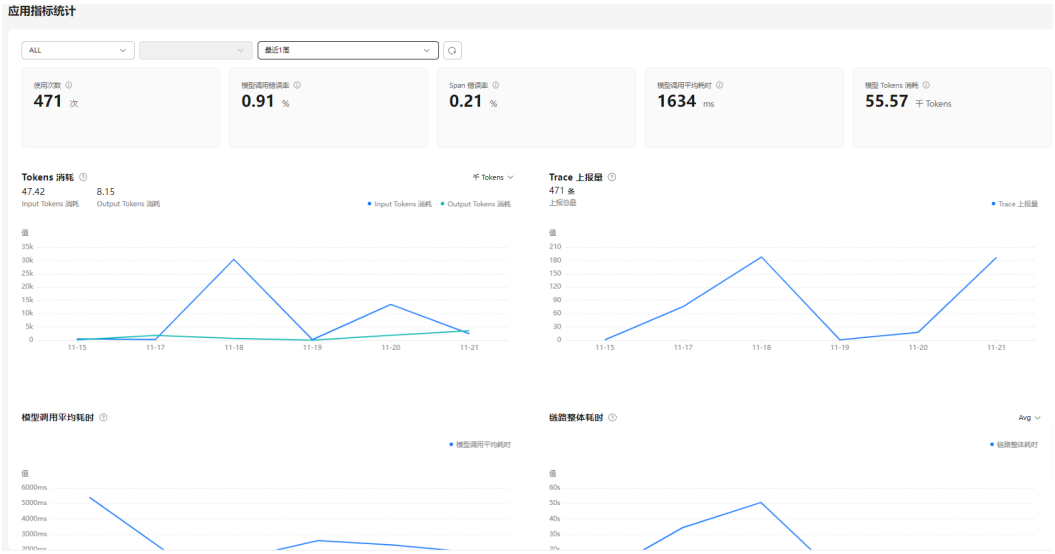
应用指标统计界面提供自动化数据统计功能，实时收集应用的性能指标和资源使用情况。这使运维人员能够快速识别性能瓶颈，从而提升系统的稳定性和可靠性，并实现资源的高效利用和成本优化。

查看应用指标统计信息

步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏中选择“ 运营运维 > 观测 > 应用指标统计”页面中，可以查看智能体和工作流上报的指标信息。

图 11-5 应用指标统计



----结束

应用指标统计信息说明

应用指标统计界面包含以下信息，如图11-5所示。

表 11-5 应用指标统计参数说明

参数	说明	示例
使用次数	所选应用中上报的Root Span的总数。	441次
模型调用错误率	Model Span的状态错误率，即错误状态的Model Span数量占总Model Span数量的比例。	31.39%
Span错误率	Span的状态错误率，即错误状态的Span数量占总Span数量的比例。	25.58%
模型调用平均耗时	模型调用的平均耗时，即Model Span的总耗时除以Model Span的总数量。	4594ms
模型Tokens消耗	Model Span数据里输入和输出所消耗Tokens的总量。	4.00千Tokens
Tokens消耗	Tokens消耗分为以下两种类型： • **Input Tokens消耗**：大模型调用过程中，输入数据所消耗的Tokens数量。 • **Output Tokens消耗**：大模型调用过程中，输出数据所消耗的Tokens数量。 在界面中可以选择以下单位显示Tokens消耗：个Tokens、千Tokens、百万Tokens。	Input Tokens：1.6千Tokens Output Tokens：2.4千Tokens
Trace上报量	以折线图的方式显示上报的Root Span的总数，反映系统中请求的总体规模和趋势。	441条

参数	说明	示例
模型调用平均耗时	以折线图的方式显示模型调用的平均耗时，反映模型调用的性能和稳定性。	2397ms
链路整体耗时	以折线图的方式显示调用链路从开始到结束所耗费的总时间，反映整个请求的处理时长。 在界面中可以选择以下单位显示链路整体耗时消耗：Avg、Max、Min、P50、P90、P99。 **Avg（Average，平均值）**：表示一组**数据的平均值**，即所有数据值相加后除以数据的总数。 **Max（Maximum，最大值）**：表示一组数据中的**最大值**。 **Min（Minimum，最小值）**：表示一组数据中的**最小值**。 **P50（50th Percentile，第50百分位数，也称为中位数）**：表示一组数据按从小到大排序后，**位于中间位置的数值**，有50%的数据小于或等于它，50%的数据大于它的数值。 **P90（90th Percentile，第90百分位数）**：表示一组数据按从小到大排序后，**有 90% 的数据小于或等于它，10% 的数据大于它**的数值。 **P99（99th Percentile，第99百分位数）**：表示一组数据按从小到大排序后，**有 99% 的数据小于等于它，1% 的数据大于等于它**的数值。	9.12s
服务请求成功率	以折线图的方式显示成功状态的Root Span数量占总Root Span数量的占比，反映服务的整体可用性和稳定性。	100%
模型请求成功率	以折线图的方式显示成功状态的Model Span数量占总Model Span数量的占比，反映模型调用的成功率和稳定性。	100%

使用过滤器筛选信息

应用指标统计界面支持多维度灵活筛选，帮助运维人员快速定位和分析目标数据。

表 11-6 过滤维度

过滤条件	说明
数据来源	按照数据来源过滤数据。支持以下三种分类： ● **ALL**：单智能体应用、工作流应用和多智能体应用所有的数据统计信息。 ● **单智能体应用**：单智能体应用下的所有应用的数据统计信息。 ● **工作流应用**：工作流应用下的所有应用的数据统计信息。 ● **多智能体应用**：多智能体应用下的所有应用的数据统计信息。

过滤条件	说明
Agent应用	支持在选择了数据来源之后，可以在筛选条件下进一步选择不同的应用。
时间	支持根据上报的时间筛选数据记录。

11.2.5 查看租户指标统计信息

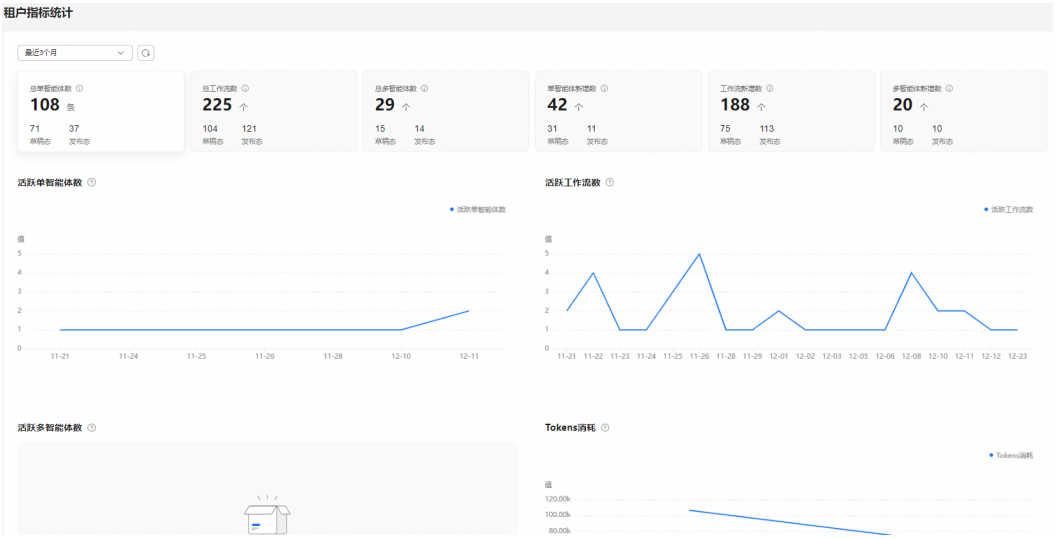
租户指标统计页面为您提供当前租户下智能体应用和工作流方面的关键使用数据。通过这些统计数据，您可以了解智能体和工作流的总数、新增数量、活跃情况以及资源消耗情况，帮助您优化资源分配，提高使用效率，并及时发现和解决问题。

查看租户指标统计信息

步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择目标空间。

步骤2 在左侧导航栏中选择“ 运营运维 > 观测 > 租户指标统计 ”。

图 11-6 租户指标统计



----结束

租户指标统计信息说明

租户指标统计界面包含以下信息，如表11-7所示。

说明

以下指标统计涵盖了当前租户下所有空间的数据信息。

表 11-7 租户指标统计参数说明

参数	说明	示例
总单智能体数	当前租户下所有草稿状态和发布状态的单智能体应用的总数。	27个
总工作流数	当前租户下所有草稿状态和发布状态的工作流应用的总数。	127个
总多智能体数	当前租户下所有草稿状态和发布状态的多智能体应用的总数。	25个
单智能体新增数	在指定时间段内新增的单智能体数量，包括草稿状态和已发布状态。	3个
工作流新增数	在指定时间段内新增的工作流数量，包括草稿状态和已发布状态。	24个
多智能体新增数	在指定时间段内新增的多智能体数量，包括草稿状态和已发布状态。	5个
活跃单智能体数	以折线图形式展示指定时间段内，使用单智能体应用进行一问一答的智能体数量。 **统计条件为：单智能体至少与用户进行过一次交互，提示词编排不纳入统计。**	1个
活跃工作流数	以折线图形式展示指定时间段内，使用工作流应用进行一问一答的工作流数量。 **统计条件为：每个工作流至少与用户进行过一次交互，提示词编排不纳入统计。**	8个
活跃多智能体数	以折线图形式展示指定时间段内，使用多智能体应用进行一问一答的智能体数量。 **统计条件为：多智能体至少与用户进行过一次交互，提示词编排不纳入统计。**	1个
Tokens消耗	以折线图形式展示在指定时间段内所有单智能体和工作流的Tokens消耗。	30.40k
TOP10单智能体的Tokens消耗	以柱状图的方式显示指定时间段内**消耗Tokens数量排名前10**的单智能体。	-
TOP10工作流的Tokens消耗	以柱状图的方式显示指定时间段内**消耗Tokens数量排名前10**的工作流。	-
TOP10多智能体的Tokens消耗	以柱状图的方式显示指定时间段内**消耗Tokens数量排名前10**的多智能体。	-
TOP10单智能体的总调用量	以柱状图的方式显示指定时间段内调**用数量排名前10**的单智能体。	-
TOP10工作流的总调用量	以柱状图的方式显示指定时间段内**调用数量排名前10**的工作流。	-
TOP10多智能体的总调用量	以柱状图的方式显示指定时间段内**调用数量排名前10**的多智能体。	-

参数	说明	示例
TOP10大模型消耗	以柱状图的方式显示指定时间段内**消耗资源数量排名前10**的大模型。	-

使用过滤器筛选信息

您可以使用以下时间过滤条件来快速定位和分析目标数据。

表 11-8 过滤维度

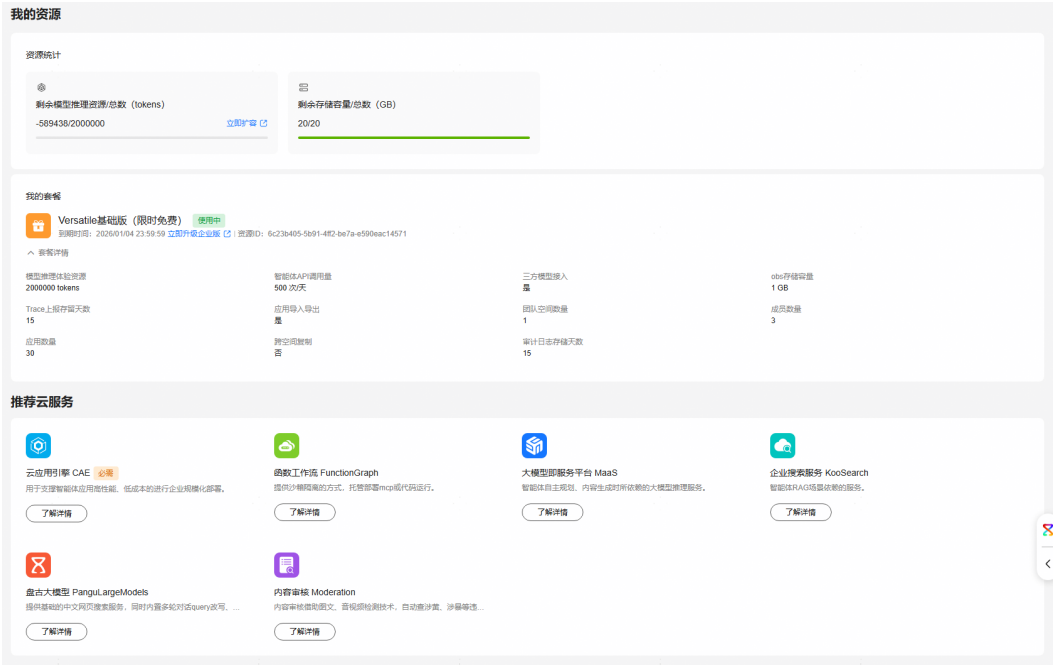
过滤条件	说明
时间	支持根据以下时间筛选数据记录。 ● 最近3天：选择最近3天内的数据记录。 ● 最近1周：选择最近1周内的数据记录。 ● 最近1个月：选择最近1个月内的数据记录。 ● 最近3个月：选择最近3个月内的数据记录。

12 平台管理

12.1 查看我的资源

购买Versatile智能体平台后，在“平台管理 > 我的资源”页面，可以查看资源统计信息、购买的套餐信息。

图 12-1 我的资源



我的资源

- 资源统计

  显示套餐资源数量使用信息及资源总数量信息。

  当“剩余模型推理资源/总数（tokens）”为“0/2000000”时，可以对“剩余模型推理资源/总数（tokens）”资源进行扩容，单击“立即扩容”，参考AI开发平台ModelArts的**开通商用服务**进行扩容。

- 我的套餐

  显示购买的套餐信息。

  如果购买的是Versatile基础版（限时免费），可以升级为Versatile企业版，单击“立即升级企业版”，具体操作请参考**变更计费模式**。

推荐云服务

这里推荐了与Versatile使用有关的云服务，单击“了解详情”，可以进入对应云服务页面。

13 审计

13.1 支持云审计的关键操作

操作场景

平台提供了云审计服务（Cloud Trace Service，简称CTS）。通过云审计服务，可记录与Versatile相关的操作事件，便于日后的查询、审计和回溯。

前提条件

已**开通云审计服务**。

支持审计的关键操作列表

表 13-1 云审计服务支持的 Versatile 服务操作列表

操作名称	资源类型	事件名称
创建函数	AgentManager	create_functions
编辑函数	AgentManager	update_functions
创建依赖包	AgentManager	create_dependencies
编辑依赖包	AgentManager	update_dependencies
删除依赖包	AgentManager	delete_dependencies
部署MCP服务	AgentManager	create_mcp_deploy
删除MCP服务	AgentManager	delete_mcp_service_delete
执行MCP服务工具	AgentManager	create_mcp_service_tools_test
MCP服务评分	AgentManager	create_mcp_server_rating

操作名称	资源类型	事件名称
修改mcp服务信息	AgentManager	update_mcp_service_modify
最终租户授权给op账号	AgentManager	create_doAuthorization
用户接入模型服务供应商列	AgentManager	create_integration_providers
更新供应商信息	AgentManager	update_integration_providers
删除供应商信息	AgentManager	delete_integration_providers
认证配置设置	AgentManager	create_provider_auths
移除供应商认证配置	AgentManager	delete_provider_auths
移除供应商认证配置	AgentManager	delete_provider_auths
新增模型服务	AgentManager	create_model_services
更新模型服务	AgentManager	update_model_services
删除模型服务	AgentManager	delete_model_services
发布模型服务	AgentManager	create_model_services_online
取消发布模型服务	AgentManager	create_model_services_offline
创建模型路由策略	AgentManager	create_model_router_strategies
删除模型路由策略	AgentManager	delete_model_router_strategies
修改模型路由策略	AgentManager	update_model_router_strategies
批量添加项目空间成员	AgentManager	create_workspace_member
批量移除项目空间成员	AgentManager	delete_workspace_member
修改项目空间成员角色	AgentManager	delete_workspace_member
当前用户退出指定空间	AgentManager	delete_workspace_member_me
转让空间所有权给指定用户	AgentManager	update_workspace_member_ownership
查询统计指标	AgentManager	create_statistic_indicator

操作名称	资源类型	事件名称
获取span列表	AgentManager	create_spans
创建评测集	AgentManager	create_evaluation_sets
查询评测集列表	AgentManager	create_evaluation_sets_list
更新评测集	AgentManager	update_evaluation_sets
删除评测集	AgentManager	delete_evaluation_sets
数据添加标签	AgentManager	create_tag_mark
数据修改标签	AgentManager	update_tag_mark
删除数据标签	AgentManager	delete_tag_mark
删除评估器	AgentManager	delete_evaluators
创建评估器	AgentManager	create_evaluators
浏览评估器	AgentManager	create_list_evaluators
更新评估器草稿	AgentManager	update_evaluators_draft
提交评估器版本	AgentManager	create_evaluators_versions
更新评估器	AgentManager	update_evaluators
浏览会话列表	AgentManager	create_sessions
创建标签	AgentManager	create_tag
修改标签信息	AgentManager	update_tag
查询提示词	AgentManager	create_prompt_engineering_template_list
导出模板样例	AgentManager	create_prompt_engineering_template_sample_download
删除模板	AgentManager	delete_prompt_engineering_template_delete
获取obs桶中文件和目录列表	AgentManager	create_obs_objects
语音合成	AgentManager	create_tts
检索知识库，命中测试	AgentManager	create_knowledges_search_text
删除知识库测试记录	AgentManager	delete_knowledges_search_records

操作名称	资源类型	事件名称
创建一个知识库	AgentManager	create_knowledges
在知识库中上传一个文档	AgentManager	create_knowledges_files
新增知识文件的切片	AgentManager	create_knowledges_files_chunks
修改知识文件的切片	AgentManager	update_knowledges_files_chunks
删除知识文件的切片	AgentManager	delete_knowledges_files_chunks
创建默认知识库配置	AgentManager	create_configurations_default_connections
编辑默认知识库连接	AgentManager	update_configurations_default_connections
用于测试默认知识库连接是否正常	AgentManager	create_configurations_default_connections_test_connection
用于测试默认知识库连接是否正常	AgentManager	create_configurations_default_connections_test_connection
修改知识库可见性	AgentManager	update_permissions
创建一个agent	AgentManager	create_agents
复制百宝箱应用	AgentManager	create_apps_copy
创建工作流	AgentManager	create_workflows
设置工作流试运行状态	AgentManager	update_workflows_test_status
工作流添加触发器	AgentManager	create_workflows_triggers_add
工作流编辑触发器	AgentManager	create_workflows_triggers_edit
工作流删除触发器	AgentManager	delete_workflows_triggers
修改工作流	AgentManager	update_workflows
删除工作流	AgentManager	delete_workflows
发布工作流版本	AgentManager	create_workflows_versions
删除一个workflow版本快照	AgentManager	delete_workflows_versions

操作名称	资源类型	事件名称
发布一个Workflow版本通道	AgentManager	create_workflows_channels
删除一个workflow版本通道	AgentManager	delete_workflows_channels
修改一个workflow版本通道	AgentManager	update_workflows_channels
复制一个agent	AgentManager	create_agents_copy
修改一个agent	AgentManager	update_agents
删除一个agent，同步删除关联的assistant	AgentManager	delete_agents
修改一个知识库	AgentManager	update_knowledges
删除一个知识库	AgentManager	delete_knowledges
从知识库中删除一个文档	AgentManager	delete_knowledges_files
添加触发器	AgentManager	create_agents_triggers_add
编辑触发器	AgentManager	create_agents_triggers_edit
删除触发器	AgentManager	delete_agents_triggers
创建一个工具	AgentManager	create_tools
修改一个工具	AgentManager	update_tools
删除一个工具	AgentManager	delete_tools
创建一个插件鉴权凭证	AgentManager	create_tools_credentials
删除一个插件鉴权凭证	AgentManager	delete_tools_credentials
导出工作流	AgentManager	create_workflows_export
导入工作流	AgentManager	create_workflows_import
导出Agent	AgentManager	create_agents_export
导入Agent	AgentManager	create_agents_import
解析导入文件	AgentManager	create_list
发布一个agent版本	AgentManager	create_agents_versions
删除一个agent版本	AgentManager	delete_agents_versions
发布一个agent版本通道	AgentManager	create_agents_channels
删除一个agent版本通道	AgentManager	delete_agents_channels

操作名称	资源类型	事件名称
修改一个agent版本通道	AgentManager	update_agents_channels
创建一个工具OpenAPI定义	AgentManager	create_tools_openapi
修改一个意图包	AgentManager	update_complex_intent
删除一个意图包	AgentManager	delete_complex_intent
创建一个意图包	AgentManager	create_complex_intent
修改一个意图分支	AgentManager	update_complex_intent_branch
删除一个意图包分支	AgentManager	delete_complex_intent_branch
创建一个意图包分支	AgentManager	create_complex_intent_branch
导出高级意图配置模板	AgentManager	create_complex_intent_export_template
导出高级意图包	AgentManager	create_complex_intent_export
导入意图分支	AgentManager	create_complex_intent_import
导入意图分支	AgentManager	create_complex_intent_branch_import
修改自定义对象	AgentManager	update_objects
删除自定义对象	AgentManager	delete_objects
创建自定义对象	AgentManager	create_objects
创建一个知识库分层规则	AgentManager	create_knowledges_rule_regex
修改一个知识库分层规则	AgentManager	update_knowledges_rule_regex
删除知识库分层规则	AgentManager	delete_knowledges_rule_regex
开启一个知识库	AgentManager	update_knowledges_start
停用一个知识库	AgentManager	update_knowledges_stop
创建一个FAQ问答对	AgentManager	create_knowledges_faq
修改一个FAQ问答对	AgentManager	update_knowledges_faq
删除一个FAQ问答对	AgentManager	delete_knowledges_faq

操作名称	资源类型	事件名称
批量删除FAQ问答对	AgentManager	create_knowledges_faq_batch_delete
创建知识库任务	AgentManager	create_knowledges_tasks
在知识库中上传一个faq文档	AgentManager	create_knowledges_faq_files
删除知识FAQ文件	AgentManager	delete_knowledges_faq_files
新增知识FAQ文件的切片	AgentManager	create_knowledges_faq_files_chunks
修改知识FAQ文件的切片	AgentManager	update_knowledges_faq_files_chunks
删除知识FAQ文件的切片	AgentManager	delete_knowledges_faq_files_chunks
从知识库中批量删除文档	AgentManager	create_knowledges_files_batch_delete
上传头像	AgentManager	create_upload_avatar
修改一个数据源	AgentManager	update_datasource
删除一个数据源	AgentManager	delete_datasource
创建一个数据源	AgentManager	create_datasource
批量删除数据源	AgentManager	create_datasource_batch_delete
发布一个工具版本	AgentManager	create_tools_versions
删除一个工具版本定义	AgentManager	delete_tools_versions
根据工具id创建工具OpenAPI定义	AgentManager	create_tools_openapi
删除知识库任务	AgentManager	create_knowledges_tasks_batch_delete
创建第三方知识库连接	AgentManager	create_connections
删除第三方知识库连接	AgentManager	delete_connections
编辑第三方知识库连接	AgentManager	update_connections
用于测试第三方知识库连接是否正常	AgentManager	create_connections_test_connection
新增第三方知识库	AgentManager	create_external
项目空间	AgentManager	create_workspace

操作名称	资源类型	事件名称
项目空间	AgentManager	delete_workspace
修改工作空间	AgentManager	update_workspace
项目空间	AgentManager	create_workspace_init
最终租户授权给op账号	AgentManager	create_doAuthorization
创建环境	AgentManager	create_environments
创建环境变量	AgentManager	create_environments_variables
删除环境信息	AgentManager	delete_environments
删除环境变量	AgentManager	delete_environments_variables
设置为默认环境	AgentManager	update_environments_is_default
修改环境	AgentManager	update_environments
修改结构化信息	AgentManager	update_structured_messages
添加结构化信息	AgentManager	create_structured_messages
批量删除结构化信息	AgentManager	delete_structured_messages
上传消息模板	AgentManager	create_structured_messages_upload_file
导出消息模板	AgentManager	create_structured_messages_export_file
结构化信息模板	AgentManager	create_structured_messages_export_template
资源（智能体、工作流、插件等）共享到其他空间	AgentManager	create_resource_share
取消资源跨空间共享	AgentManager	delete_resource_share
创建一个插件	AgentManager	create_plugins
创建一个工具	AgentManager	create_tools
根据工具id创建工具OpenAPI定义	AgentManager	create_plugins_openapi
删除一个插件	AgentManager	delete_plugins
删除一个工具版本定义	AgentManager	delete_plugins_versions

操作名称	资源类型	事件名称
删除一个工具	AgentManager	delete_tools
导出插件	AgentManager	create_plugins_export
导入插件	AgentManager	create_plugins_import
修改一个插件	AgentManager	update_plugins
修改一个工具	AgentManager	update_tools
发布一个工具版本	AgentManager	create_plugins_versions
查询租户全量统计指标	AgentManager	create_statistic_indicator_domain
查询运营统计指标	AgentManager	create_statistic_indicator_operation
查询top指标	AgentManager	create_statistic_indicator_top
查询当前登录用户是否为白名单	AgentManager	create_statistic_config
创建实验任务	AgentManager	create_evaluation_experiment_submit
查询实验列表，带上实验状态	AgentManager	create_evaluation_experiment_list
删除历史实验	AgentManager	delete_evaluation_experiment_delete
查询实验数据结果列表	AgentManager	create_evaluation_experiment_data_detail
查询实验的输出指标	AgentManager	create_evaluation_experiment_metric
执行Agent，带会话id	AgentManager	create_agents_conversations
执行Agent，不带会话id	AgentManager	create_agents_conversations
执行workflow	AgentManager	create_workflows_conversations
Agent对话上传文件	AgentManager	create_agents_upload_file
创建 API Key	AgentManager	create_api_auth_api_keys
创建知识库obs配置	AgentManager	create_obs_configs
根据obs配置执行动作	AgentManager	create_obs_configs_execute

操作名称	资源类型	事件名称
展示用户obs指定路径下的文件对象名集合	AgentManager	create_obs_bucket_objects
更新会话变量	AgentManager	update_agents_conversations_variables
函数插件绑定函数	AgentManager	create_plugins_bind_function
预置的函数插件复制到个人空间	AgentManager	create_plugins_copy_function
获取函数详情	AgentManager	delete_functions
创建副本	AgentManager	create_prompt_tasks_copy
创建提示词优化任务	AgentManager	create_prompt_tasks
创建提示词优化任务草稿	AgentManager	create_prompt_tasks_draft
删除提示词优化任务	AgentManager	delete_prompt_tasks
暂停提示词优化任务	AgentManager	create_prompt_tasks_pause
继续提示词优化任务	AgentManager	create_prompt_tasks_resume
全量更新提示词优化任务（只能更新草稿）	AgentManager	update_prompt_tasks
向优化任务批量添加数据用例	AgentManager	create_prompt_tasks_data_batchItems
向优化任务添加单条数据用例	AgentManager	create_prompt_tasks_data_items
删除优化任务中的单条用例	AgentManager	delete_prompt_tasks_data_items
通过文件导入批量添加数据用例	AgentManager	create_prompt_tasks_data_import
更新优化任务中的单条用例	AgentManager	update_prompt_tasks_data_items
上传图片，获得临时url	AgentManager	create_prompt_tasks_upload_picture
更新提示词	AgentManager	update_prompt
创建一个提示词	AgentManager	create_prompt
保存模板	AgentManager	create_prompt_tasks_prompt_save

操作名称	资源类型	事件名称
提示词一键优化	AgentManager	create_prompt_generate
导入模板	AgentManager	create_prompt_engineering_template_import
导出模板	AgentManager	create_prompt_engineering_template_download
创建卡片	AgentManager	create_cards
更新卡片信息	AgentManager	update_cards
删除卡片	AgentManager	delete_card
删除一个卡片版本定义	AgentManager	delete_card_versions
发布一个卡片版本	AgentManager	create_card_versions
数据点赞点踩	AgentManager	create_tag_vote
上传一个写作模板	AgentManager	create_agents_uploadDeepResearchTemplate
数据清理	AgentManager	create_cleannotify
环境变量导入	AgentManager	create_environments_variables_import
环境变量导出	AgentManager	create_environments_variables_export
更新模型调测状态	AgentManager	create_model_test_status
批量删除用户的变量记忆	AgentManager	create_agents_memories_variables_batch_delete
重置用户的变量记忆	AgentManager	create_agents_memories_variables_reset
根据conversation_id删除会话	AgentManager	delete_agents_conversations_history
提交异步工作流任务	AgentManager	create_workflows_tasks
继续执行任务	AgentManager	create_workflows_tasks
修改任务信息	AgentManager	update_workflows_tasks
删除异步任务详情	AgentManager	delete_workflows_tasks
取消异步任务详情	AgentManager	delete_workflows_tasks_cancel
创建实验委托	AgentManager	create_evaluation_experiment_create_agency
创建智能体实例	AgentManager	create_deploy

操作名称	资源类型	事件名称
模型调测	AgentManager	create_chat_completions
文本转语音	AgentManager	create_agents_audio_tts
语音转写	AgentManager	create_agents_audio_transcriptions
增加分析事件	AgentManager	create_agents_analytics_event
一句话语音识别	AgentManager	create_asr_short_audio
批量部署MCP服务	AgentManager	create_mcp_deploy_batch
一键添加到评测集	AgentManager	create_trace_evaluation_sets_items_batch_add
同步第三方模型服务	AgentManager	create_model_services_sync
订单回调-服务开通(支付完成)	lumina-console-backend	backend-createAgency
订单冻结解冻回调	lumina-console-backend	backend-changeResourceStatus
订单操作-资源续费	lumina-console-backend	backend-renewalResource
临时订购接口	lumina-console-backend	backend-developerSpaceSubscribe
根据SkuCode查询CBC报价,销售周期	lumina-console-backend	backend-queryResource
根据SkuCode查询变化的CBC报价	lumina-console-backend	backend-queryResource
ROS调用-清理租户数据	lumina-console-backend	backend-cleanupResources
定时-更新清理租户数据状态	lumina-console-backend	backend-updateStatusTask
定时-开始清理子模块租户数据	lumina-console-backend	backend-cleanModuleTask
查询话单	lumina-console-backend	backend-getSdrBilling
SKU订购	lumina-console-backend	backend-subscribeOrder
变更下单	lumina-console-backend	backend-changeOrder
实例同步接口	lumina-console-backend	backend-syncInstance
资源同步接口	lumina-console-backend	backend-syncResource

操作名称	资源类型	事件名称
console同步信息到服务	lumina-console-backend	create_common_console_subscribe_sync
license消费接口	lumina-console-backend	create_common_intern_license_report
kms加密接口	lumina-console-backend	create_common_intern_kms_encrypt
kms解密接口	lumina-console-backend	create_common_intern_kms_decrypt

13.2 在 CTS 基础查询审计事件

操作场景

用户进入云审计服务创建管理类追踪器后，系统开始记录云服务资源的操作。在创建数据类追踪器后，系统开始记录用户对OBS桶中数据的操作。云审计服务管理控制台会保存最近7天的操作记录。

本节介绍如何在云审计服务管理控制台查询或导出最近7天的操作记录。

- **在CTS新版事件列表查看审计事件**
- **在CTS旧版事件列表查看审计事件**

约束与限制

- 管理类追踪器未开启组织功能之前，单账号跟踪的事件可以通过云审计控制台查询。管理类追踪器开启组织功能之后，多账号的事件只能在账号自己的事件列表页面去查看，或者到组织追踪器配置的OBS桶中查看，也可以到组织追踪器配置的CTS/system日志流下面去查看。组织追踪器的详细介绍请参见**组织追踪器概述**。
- 用户通过云审计控制台只能查询最近7天的操作记录，过期自动删除，不支持人工删除。如果需要查询超过7天的操作记录，您必须配置转储到对象存储服务（OBS）或云日志服务（LTS），才可在OBS桶或LTS日志流里面查看历史事件信息。否则，您将无法追溯7天以前的操作记录。
- 用户对云服务资源做出创建、修改、删除等操作后，1分钟内可以通过云审计控制台查询管理类事件操作记录，5分钟后才可通过云审计控制台查询数据类事件操作记录。
- CTS新版事件列表不显示数据类审计事件，您需要在旧版事件列表查看数据类审计事件。

在 CTS 新版事件列表查看审计事件

步骤1 登录**CTS控制台**。

步骤2 登录控制台，单击左上角 ☰，选择“管理与部署 > 云审计服务 CTS”，进入云审计服务页面。

步骤3 单击左侧导航栏的“事件列表”，进入事件列表信息页面。

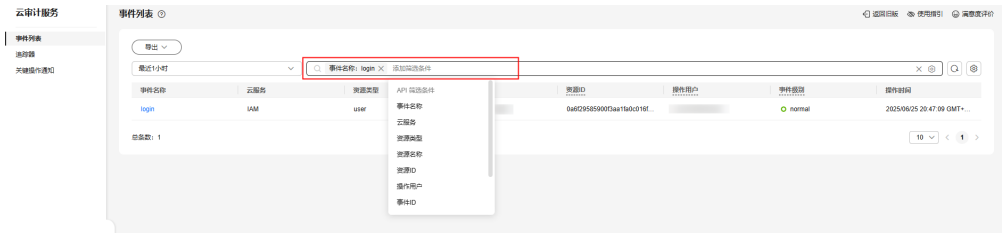
步骤4 在列表上方，可以通过筛选时间范围，查询最近1小时、最近1天、最近1周的操作事件，也可以自定义最近7天内任意时间段的操作事件。

步骤5 事件列表支持通过高级搜索来查询对应的操作事件，您可以在筛选器组合一个或多个筛选条件。

表 13-2 事件筛选参数说明

参数名称	说明
是否只读	下拉选项包含“是”、“否”，只可选择其中一项。 • 是：筛选只读操作事件，例如查询资源操作。当用户在“配置中心”页面开启了只读事件上报后，并触发了只读事件，才支持选择该选项。 • 否：筛选非只读操作事件，例如创建资源操作、修改资源操作、删除资源操作。
事件名称	操作事件的名称。 输入的值区分大小写，需全字符匹配，不支持模糊匹配模式。 各个云服务支持审计的操作事件的名称请参见**支持审计的服务及详细操作列表**《云审计服务用户指南》的“支持审计的服务及操作列表”章节。 示例：updateAlarm
云服务	云服务的名称缩写。 输入的值区分大小写，需全字符匹配，不支持模糊匹配模式。 示例：IAM
资源名称	操作事件涉及的云资源名称。 输入的值区分大小写，需全字符匹配，不支持模糊匹配模式。 当该事件所涉及的云资源无资源名称或对应的API接口操作不涉及资源名称参数时，该字段为空。 示例：ecs-name
资源ID	操作事件涉及的云资源ID。 输入的值区分大小写，需全字符匹配，不支持模糊匹配模式。 当该资源类型无资源ID或资源创建失败时，该字段为空。 示例：*{虚拟机ID}*
事件ID	操作事件日志上报到CTS后，查看事件中的trace_id参数值。 输入的值需全字符匹配，不支持模糊匹配模式。 示例：01d18a1b-56ee-11f0-ac81-******1e229
资源类型	操作事件涉及的资源类型。 输入的值区分大小写，需全字符匹配，不支持模糊匹配模式。 各个云服务的资源类型请参见**支持审计的服务及详细操作列表**《云审计服务用户指南》的“支持审计的服务及操作列表”章节。 示例：user

参数名称	说明
操作用户	触发事件的操作用户。 下拉选项中选择一个或多个操作用户。 查看事件中的trace_type的值为“SystemAction”时，表示本次操作由服务内部触发，该条事件对应的操作用户可能为空。 IAM身份与操作用户对应关系，以及操作用户名称的格式说明，请参见**IAM身份与操作用户对应关系**。
事件级别	下拉选项包含“normal”、“warning”、“incident”，只可选择其中一项。 • normal代表操作成功。 • warning代表操作失败。 • incident代表比操作失败更严重的情况，如引起其他故障等。
企业项目ID	资源所在的企业项目ID。 查看企业项目ID的方式：在EPS服务控制台的“项目管理”页面，可以查看企业项目ID。 示例：b305ea24-c930-4922-b4b9-******1eb2
访问密钥ID	访问密钥ID，包含临时访问凭证和永久访问密钥。 查看访问密钥ID的方式：在控制台右上方，用户名下拉选项中，选择“我的凭证 > 访问密钥”，可以查看访问密钥ID。 示例：HSTAB47V9V*******TLN9



步骤6 在事件列表页面，您还可以导出操作记录文件、刷新列表、设置列表展示信息等。

- 在搜索框中输入任意关键字，按下Enter键，可以在事件列表搜索符合条件的数据。
- 单击“导出”按钮，云审计服务会将查询结果以.xlsx格式的表格文件导出，该.xlsx文件包含了本次查询结果的所有事件，且最多导出5000条信息。
- 单击按钮，可以获取到事件操作记录的最新信息。
- 单击按钮，可以自定义事件列表的展示信息。启用表格内容折行开关，可让表格内容自动折行，禁用此功能将会截断文本，默认停用此开关。

步骤7 （可选）在新版事件列表页面，单击右上方的“返回旧版”按钮，可切换至旧版事件列表页面。

----结束

在 CTS 旧版事件列表查看审计事件

步骤1 登录CTS控制台。

步骤2 登录控制台，单击左上角 ☰ ，选择“管理与部署 > 云审计服务 CTS”，进入云审计服务页面。

步骤3 单击左侧导航栏的“事件列表”，进入事件列表信息页面。

步骤4 用户每次登录云审计控制台时，控制台默认显示新版事件列表，单击页面右上方的“返回旧版”按钮，切换至旧版事件列表页面。

步骤5 在页面右上方，可以通过筛选时间范围，查询最近1小时、最近1天、最近1周的操作事件，也可以自定义最近7天内任意时间段的操作事件。

步骤6 事件列表支持通过筛选来查询对应的操作事件，如图13-1所示。

图 13-1 筛选框

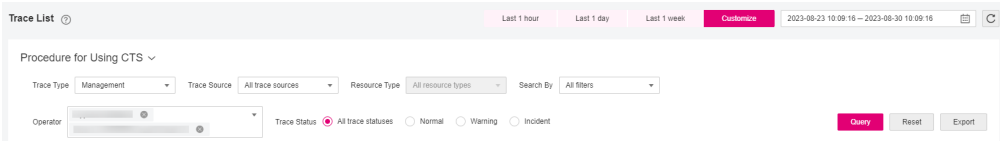


表 13-3 事件筛选参数说明

参数名称	说明
事件类型	事件类型分为“管理事件”和“数据事件”。 ● 管理类事件，即用户对云服务资源新建、修改、删除等操作事件。 ● 数据类事件，即OBS服务上报的OBS桶中的数据的操作事件，例如上传数据、下载数据等。
云服务	在下拉选项中，选择触发操作事件的云服务名称。
资源类型	在下拉选项中，选择操作事件涉及的资源类型。 各个云服务的资源类型请参见支持审计的服务及详细操作列表《云审计服务用户指南》的“支持审计的服务及操作列表”章节。
筛选类型	筛选类型分为“资源ID”、“事件名称”和“资源名称”。 ● 资源ID：操作事件涉及的云资源ID。 当该资源类型无资源ID，或资源创建失败时，该字段为空。 ● 事件名称：操作事件的名称。 各个云服务支持审计的操作事件的名称请参见支持审计的服务及详细操作列表《云审计服务用户指南》的“支持审计的服务及操作列表”章节。 ● 资源名称：操作事件涉及的云资源名称。 当事件所涉及的云资源无资源名称，或对应的API接口操作不涉及资源名称参数时，该字段为空。

参数名称	说明
操作用户	触发事件的操作用户。 下拉选项中选择一个或多个操作用户。 查看事件中的trace_type的值为“SystemAction”时，表示本次操作由服务内部触发，该条事件对应的操作用户可能为空。 IAM身份与操作用户对应关系，以及操作用户名称的格式说明，请参见**IAM身份与操作用户对应关系**。
事件级别	可选项包含“所有事件级别”、“Normal”、“Warning”、“Incident”，只可选择其中一项。 • Normal代表操作成功。 • Warning代表操作失败。 • Incident代表比操作失败更严重的情况，如引起其他故障等。

步骤7 选择完查询条件后，单击“查询”。

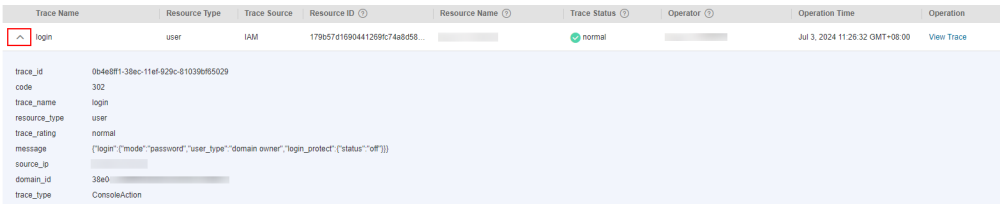
步骤8 在事件列表页面，您还可以导出操作记录文件和刷新列表。

- 单击“导出”按钮，云审计服务会将查询结果以CSV格式的表格文件导出，该CSV文件包含了本次查询结果的所有事件，且最多导出5000条信息。
- 单击按钮，可以获取到事件操作记录的最新信息。

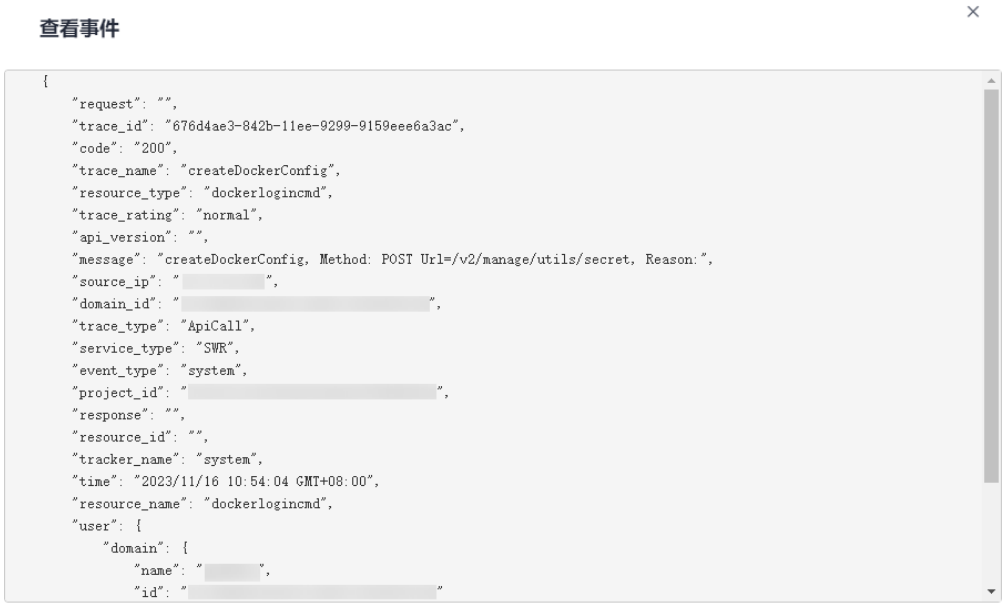
步骤9 在事件的“是否篡改”列中，您可以查看该事件是否被篡改：

- 上报的审计日志没有被篡改，显示“否”；
- 上报的审计日志被篡改，显示“是”。

步骤10 在需要查看的事件左侧，单击 展开该记录的详细信息。



步骤11 在需要查看的记录右侧，单击“查看事件”，会弹出一个窗口显示该操作事件结构的详细信息。



步骤12 （可选）在旧版事件列表页面，单击右上方的“体验新版”按钮，可切换至新版事件列表页面。

----结束

A 应用开发常见问题

常见报错及解决方案请详见表A-1。

表 A-1 常见报错与解决方案

模块名称	错误码	错误描述	解决方案
开始节点	101501	开始节点输入参数未传入值。	请确认开始节点输入参数均已传值。
结束节点	101531	结束节点初始化失败。	检查结束节点是否配置正确。
	101532	结束节点模板拼接失败。	结束节点指定回复，遵循jinja语法，请确认指定回复内容语法正确，变量在输入参数均已配置。
	101533	结束节点流式处理失败。	系统异常，请联系客服解决。
大模型节点	101561	大模型节点初始化失败。	检查大模型节点配置是否正确。
	101562	大模型节点模板拼接错误。	先检查模板占位符与输入是否匹配，如果仍无法解决，请联系客服解决。
	101563	获取模型流式输出错误。	系统异常，请联系客服解决。
代码节点	101591	代码节点初始化失败。	检查代码节点配置是否正确。
	101592	代码节点安全沙箱请求失败。	代码沙箱执行代码异常，请确认代码语法正确。
	101593	代码节点沙箱执行错误。	请联系客服解决。
	101594	代码节点安全沙箱未知异常。	请联系客服解决。

模块名称	错误码	错误描述	解决方案
	101595	代码节点执行失败未知异常。	请联系客服解决。
	120007	输出键未在代码返回值中。	请检查代码返回值是否正确配置输出键。
	120008	代码返回值的类型错误。	代码返回值的类型应为字典。
消息节点	101651	消息组件初始化失败。	检查消息节点配置是否正确。
	101652	消息节点缺少模板信息。	配置消息节点的提示词模板。
	101653	消息节点模板拼接错误。	先检查模板占位符与输入是否匹配，如果仍无法解决，请联系客服解决。
	101654	消息组件执行失败。	请联系客服解决。
意图识别节点	101098	意图识别prompt模板请求失败。	检查模板占位符与输入是否匹配。
	101097	意图识别调用大模型的prompt不符合模型输入的规范。	检查输入的prompt格式，消息的角色和内容。
	101096	意图识别调用大模型失败。	检查消息的格式，内容以及大模型服务是否正常。
	101095	意图识别用户query输入/引用解析失败。	检查用户query格式和内容。
	101094	意图识别prompt模板构建失败。	检查内置模板以及输入的system prompt格式与内容。
提问器节点	102053	提示词模板格式错误。	检查提示词模板是否格式有误。
	103004	大模型推理失败时触发该错误码。	请检查模型服务是否可以正常运行。
	101042	响应类型无效。	请联系客服解决。
	101043	超过最大对话次数。	检查实际对话次数是否超出所设定的最大对话次数范围。
	101044	问题构建方法无效。	请联系客服解决。

模块名称	错误码	错误描述	解决方案
	101045	在直接用户响应收集模式下，问题内容不能为空。	请检查问题内容。
	101047	初始化护栏失败。	检查提问器节点相关配置是否正确。
	101048	运行自定义输入护栏失败。	检查自定义输入相关配置是否正确。
	101049	运行自定义执行护栏失败。	请联系客服解决。
	101050	执行默认护栏（时间参数解析）失败。	可检查支持处理的时间类型是否超出支持范围。
	101059	在问题生成器中执行反思失败。	若重试无法解决，请联系客服解决。
问答节点	120000	问答选项为空。	请检查问答节点相关配置。
	120001	不支持的问题选择策略。	请选择已支持的问题选择策略。
	120002	问答指定索引超出范围。	请修改索引值。
	120003	在输入中未找到指定键。	请检查相关键是否匹配输入。
插件节点	101741	插件组件初始化失败。	检查插件组件配置是否正确。
	101742	工作流插件节点参数类型转换时出错。	根据error message确定具体转换出错的参数名称，并确认类型是否正确。
	101743	工作流插件节点的input在插件定义中不存在。	检查插件定义和对应的组件定义是否匹配。
	101744	插件定义了response，但实际插件执行结果与定义不一致。	检查插件response定义和实际插件执行结果是否匹配。
	101745	工作流插件节点执行出错。	插件执行出错，可以根据具体的error message信息定位。如果message无有效信息，说明该错误属于未捕获到的异常。

模块名称	错误码	错误描述	解决方案
	105001	插件执行时发生了无法捕获的异常。	检查插件本身是否可用。
	105002	插件已存在。	若重试无法解决，请联系客服解决。
	105003	插件未找到。	该插件不存在，请检查后重试。
	105004	插件定义时check param error。	根据对应error message信息确定具体出错的参数定义。
	105005	插件定义不合法。	插件定义时的数据不合法，例如字段定义超出最长长度，具体根据error message判断。
	105006	插件数据库异常。	若重试无法解决，请联系客服解决。
	105007	插件执行SQL语句异常。	请检查插件SQL语句是否符合要求。
	105008	插件内部错误。	请确认接入的插件服务是否正常。
	105009	插件嵌入推理失败。	请联系客服解决。
	105010	插件运行时鉴权出错。	可根据error message信息确定具体出错的鉴权问题，并检查鉴权信息的传递和插件鉴权定义是否正确。
	105011	插件运行返回的响应代码非200。	请检查插件服务是否正常。
	105012	插件request请求超时。	插件请求超时，检查插件服务是否正常。
	105013	插件返回结果过大。	当前支持10M大小的返回，超出此大小会报错。
	105014	插件request proxy error。	请检查插件服务是否有问题导致无法连接。
工作流节点	101801	子工作流构造失败。	请根据ID检查子工作流是否存在。
	101802	子工作流初始化失败。	检查子工作流配置是否正确。
	101803	子工作流节点配置校验失败。	先检查子工作流配置是否正确，如果仍无法解决，请联系客服解决。

模块名称	错误码	错误描述	解决方案
	101804	子工作流执行失败。	请联系客服解决。
	101805	子工作流的ID与父工作流冲突。	修改子工作流的ID，不能与父工作流相同。
循环节点	101811	不支持该循环类型。	请选择已支持的循环类型。
	101812	该字段在指定循环类型中不应为空。	正确配置该字段。
	101813	该字段类型在指定循环类型中不正确。	正确配置该字段类型。
变量聚合节点	101831	该分组值的类型不一致。	重新配置使该分组值的类型保持一致。
	101832	不支持该聚合策略。	请选择已支持的聚合策略。
输入节点	101850	输入字段不能为空。	正确配置输入字段。
	101862	循环次数无效。	循环次数应在范围[1, 1000]内。
MCP服务节点	101870	不支持该MCP类型。	请选择已支持的MCP类型。
	101872	该字段在指定MCP类型中不应为空。	正确配置该字段。
	101873	MCP执行错误。	请联系客服解决。
	101874	不支持该MCP参数方法。	请选择已支持的MCP参数方法。
	101875	该MCP参数的类型无效。	根据预期类型修改该MCP参数的类型。
高级意图节点	109000	高级意图初始化错误。	检查高级意图节点配置是否正确。
全局变量	120005	全局值中键的默认值类型错误。	请检查默认值类型配置。
认证鉴权	110000	认证失败。	请检查是否正确传入认证信息。

模块名称	错误码	错误描述	解决方案
	110001	用户信息获取失败。	查看用户信息是否正确配置。
工作流	112501	工作流认证失败。	请检查是否正确传入认证信息。
	112502	缺少必填参数。	请确认必填参数都已正确填写。
	112600	workflow ir转化失败	需要查看工作流配置是否正确。
	112941	获取workflow对话历史失败	请联系客服解决。
	101901	workflow节点配置加载失败。	查看对应节点是否配置正确。
	101902	workflow ir校验失败。	请联系客服解决。
	101903	工作流ID格式错误。	ID只允许包含数字、字母和下划线。
	101904	工作流初始化时数据类型错误。	请联系客服解决。工作流初始化时数据类型错误，必须提供ir或workflow_state其中之一。
	101905	任务超时，已被取消。	若重试无法解决，请联系客服解决。
	101906	工作流初始化失败。	检查整个工作流配置是否正确。
	101907	该工作流执行已终止。	若重试无法解决，请联系客服解决。
	101912	获取工作流执行结果失败。	请联系客服解决。
	101913	取消任务失败。	请联系客服解决。
	101914	Celery消息代理初始化失败。	请联系客服解决。
	101920	ApsWorker注册失败。	请联系客服解决。
	101921	ApsWorker Redis监听器错误。	请联系客服解决。
	101922	ApsWorker获取本地主机名失败。	请联系客服解决。

模块名称	错误码	错误描述	解决方案
	101923	ApsWorker执行时worker_id错误。	请联系客服解决。
	101924	工作流异常终止节点终止了执行。	检查异常终止节点相关配置并重试，若无法解决，请联系客服解决。
	101931	工作流流式回调实例无效。	请联系客服解决。
	101032	workflow定义不合法。	请检查工作流流程是否配置正确。
	101039	节点执行失败。	请检查对应节点是否配置正确。
	101040	workflow执行失败。	请联系客服解决。
	101046	工作流循环次数超出限制。	请减少循环次数。
	101051	工作流异常信息callback失败。	查看后台日志异常信息。
	101052	工作流的回调信息获取失败。	查看workflow的回调接口是否有问题。
	101053	工作流前处理信息callback失败。	查看workflow的callback是否有问题。
	101054	工作流后处理信息callback失败。	查看workflow的callback是否有问题。
	101057	工作流资源释放失败。	查看后台日志。
	101058	工作流配置加载失败。	检查工作流配置是否正确。
	101621	工作流节点表达式中存在语法错误。	请联系客服解决。
	101622	工作流节点评估表达式时出错。	请联系客服解决。
	101041	工作流Planner类型无效。	请联系客服解决。

模块名称	错误码	错误描述	解决方案
	120006	工作流节点数量过多。	根据提示的上限值对工作流进行精简。
单智能体	103025	当前输入类型不支持。	请修改为支持的输入类型。
	103028	工具切换字典类型无效。	请联系客服解决。
	103029	插件更新错误。	若重试无法解决，请联系客服解决。
	103030	任务ID的长度超出最大限制。	修改任务ID。
	103031	MCP更新错误。	若重试无法解决，请联系客服解决。
控制器	103100	控制器提取参数调用失败。	若重试无法解决，请联系客服解决。
	103101	控制器精炼参数调用失败。	若重试无法解决，请联系客服解决。
	103102	控制器中断错误。	若重试无法解决，请联系客服解决。
	103103	全局意图跳转次数超出最大限制。	请联系客服解决。
	103104	控制器意图检测错误。	若重试无法解决，请联系客服解决。
	103105	控制器意图工作流执行错误。	先检查意图工作流是否配置正确。若未解决，请联系客服解决。
多智能体	103200	多Agent运行器已启动。	请联系客服解决。
	103201	多Agent运行器目标成员未找到。	请联系客服解决。
	103202	多Agent运行器消息被挂起，等待用户输入。	若重试无法解决，请联系客服解决。
	103203	多Agent运行器处理中断。	请联系客服解决。
	103204	多Agent运行器未运行。	请联系客服解决。

模块名称	错误码	错误描述	解决方案
	103205	多Agent运行器未启动。	请联系客服解决。
	103231	多Agent运行空间执行错误。	若重试无法解决，请联系客服解决。
	103232	多Agent运行空间关闭消息队列失败。	请联系客服解决。
	103233	多Agent运行空间停止运行任务失败。	请联系客服解决。
	103234	多Agent运行空间空闲状态下停止失败。	请联系客服解决。
	103235	多Agent运行空间检查停止条件失败。	请联系客服解决。
	103236	多Agent运行空间执行stop_when时失败。	请联系客服解决。
	103261	多Agent成员处理消息失败。	请联系客服解决。
	103262	多Agent成员处理消息错误。	请联系客服解决。
	103291	多Agent消息队列异常。	请联系客服解决。
	103300	多Agent组已运行。	请联系客服解决。
	103301	多Agent组未运行或未正确初始化。	检查多智能体配置，若配置无问题请联系客服解决。
	103302	多Agent组执行错误。	请联系客服解决。
	103303	多Agent组注册失败。	请联系客服解决。
接口服务	115007	对话接口服务队列阻塞。	服务器繁忙，请稍后再试。
	115008	Agent版本不支持。	查看Agent版本。

模块名称	错误码	错误描述	解决方案
	100000	请求的资源未找到。	若重试无法解决，请联系客服解决。
	100001	身份验证失败。	若重试无法解决，请联系客服解决。
	100002	输入参数校验失败。	若重试无法解决，请联系客服解决。
	100003	调用SDK请求失败。	若重试无法解决，请联系客服解决。
	100004	调用默认加密方法失败。	若重试无法解决，请联系客服解决。
	100005	调用默认解密方法失败。	若重试无法解决，请联系客服解决。
	100006	加密解密初始化失败。	若重试无法解决，请联系客服解决。
	100007	该接口为预留接口，当前无法调用。	请联系客服解决。
	1000_10	没有可用的连接服务器。	若重试无法解决，请联系客服解决。
	1000_11	尝试通过已关闭的连接发送函数。	请联系客服解决。
	1000_12	查询的连接尚未绑定。	请联系客服解决。
	1000_13	参数类型无效。	应为ssl.SSLContext或jiuwen.serve.common.ssl_ctx.ContexConfigDict。
	1000_14	持久化连接服务器启动失败。	若重试无法解决，请联系客服解决。
	1000_15	尝试启动WebSocket服务器时端口类型错误。	应为int类型，请根据提示修改。
	1000_16	尝试启动WebSocket服务器时主机类型错误。	应为str或None，请根据提示修改。
	1000_17	新连接收到回声错误。	请联系客服解决。

模块名称	错误码	错误描述	解决方案
	1000_18	连接模式异常。	需要ssl_context，请提供ssl.SSLContext或jiuwen.serve.common.ssl_ctx.ContexConfigDict 实例。
	1000_20	最大字节数错误。	最大字节数必须为正整数，请检查配置文件。
	100021	LLM服务配置信息缺失。	请检查配置文件。
	100022	LLM服务格式字段缺失。	请检查相关格式字段。
	100023	LLM服务格式错误。	请检查相关格式。
	100024	LLM模型来源错误。	请检查模型来源。
	100025	模型请求错误。	若重试无法解决，请联系客服解决。
	100026	模型加载错误。	若重试无法解决，请联系客服解决。
	100027	模型类型错误。	请检查模型类型信息。
	100028	模型解析器解码错误。	请联系客服解决。
	100029	Agent ir字段校验错误。	请联系客服解决。
	100066	请求获取方法错误。	若重试无法解决，请联系客服解决。
状态管理	100040	IR数据校验失败。	请联系客服解决。
	100041	状态存储介质值无效。	请使用redis或memory。
	100042	当前不支持该版本的IR。	请联系客服解决。
	100043	EXECUTION_SUPPORTED_IR_VERSIONS配置值有误。	请检查配置或环境变量。
	100044	从JSON加载IR数据失败。	若重试无法解决，请联系客服解决。

模块名称	错误码	错误描述	解决方案
	100045	此键在状态存储介质中不存在。	请联系客服解决。
数据库	100051	数据库连接错误。	若重试无法解决，请联系客服解决。
	100052	数据库执行错误。	若重试无法解决，请联系客服解决。
Redis存储	100061	序列化错误。	请联系客服解决。
	100062	反序列化错误。	请联系客服解决。
	100100	Redis服务不存在或连接异常。	请联系客服解决。
	100101	Redis连接初始化过程中发生错误。	请联系客服解决。
	100102	Redis插入元素时发生错误。	请联系客服解决。
	100103	无法将元素插入到Redis列表的末尾。	请联系客服解决。
	100104	无法从Redis列表中移除元素。	请联系客服解决。
	100105	无法设置Redis列表中指定位置的值。	请联系客服解决。
	100106	无法从Redis列表中获取值。	请联系客服解决。
	100107	无法从列表中弹出元素。	请联系客服解决。
	100108	无法获取列表长度。	请联系客服解决。
	100109	无法从Redis获取历史记录。	请联系客服解决。
	100110	无法将历史记录设置到Redis，无法从Redis获取元素。	请联系客服解决。

模块名称	错误码	错误描述	解决方案
	100111	无法将元素设置到Redis。	请联系客服解决。
	100112	无法从Redis删除元素。	请联系客服解决。
	100113	Redis的cluster_node配置错误。	Redis的cluster_node配置必须为IP:PORT,IP:PORT格式，请检查相关配置。
	100114	无法从Redis获取集群节点列表。	请联系客服解决。
OBS存储	100502	必要的OBS配置项不正确。	请检查相关OBS配置项。
	100503	检查存储桶失败。	若重试无法解决，请联系客服解决。
	100504	文件上传失败。	若重试无法解决，请联系客服解决。
	100505	从存储桶获取对象失败。	若重试无法解决，请联系客服解决。
	100506	在存储桶中复制对象失败。	若重试无法解决，请联系客服解决。
	100507	在存储桶中删除对象失败。	若重试无法解决，请联系客服解决。
编排引擎	101001	参数action_map类型错误。	参数action_map应为字典类型，请修改。
	101002	使用async_fn异常。	在使用async_fn时不能指定wrap_cls。
	101003	async_fn函数调用异常。	async_fn仅允许异步函数。
	101004	参数wrap_cls错误。	参数wrap_cls仅允许为Invokable的子类。
	101005	invokable参数异常。	invokable不支持init中的必要关键字参数。
	101006	环境变量TGF_ENABLE配置错误。	环境变量TGF_ENABLE的值必须为true或false。
	101007	未找到相关的Invokable。	若重试无法解决，请联系客服解决。
	101008	未找到相关值。	请联系客服解决。

模块名称	错误码	错误描述	解决方案
	101009	编排IR的JSON格式非法。	请联系客服解决。
	101010	invokable_irs格式错误。	invokable_irs应为字符串列表（JSON格式）。
	101011	在索引处的Invokable IR JSON格式非法。	请检查相关格式，如果无法解决请联系客服解决。
	101012	system_configs类型错误。	system_configs必须为列表类型。
	101013	待删除的agent属性缺失。	每个待删除的agent都需要<agentId>和<version>属性。
	101014	IrManager的流式回调接收到不支持的状态码。	请联系客服解决。
	101015	未知的状态码。	请联系客服解决。
	101016	无法以指定代码注销。	请联系客服解决。
	101017	以指定代码运行失败。	请联系客服解决。
	101018	以指定代码发送会话的用户输入失败。	请联系客服解决。
	101019	以指定代码注册失败。	请联系客服解决。
	101020	以指定代码更新失败。	请联系客服解决。
	101021	没有可用的IR manager。	请联系客服解决。
	101022	没有可用的Invokable manager。	请联系客服解决。
	101023	Invokable的JSON格式非法。	请检查相关格式，如果无法解决请联系客服解决。
	101024	从目录注册Invokable manager类型错误。	从目录注册Invokable manager仅接受字符串类型。

模块名称	错误码	错误描述	解决方案
	101025	从目录注册Invokable失败。	指定路径不是一个目录。
	101026	从目录注册Invokable失败。	请联系客服解决。
	101027	注册到PluginManager失败。	若重试无法解决，请联系客服解决。
	101028	Invokable失败。	请联系客服解决。
	101029	设置的contentType与用户输入不符。	请修改Content-Type，或改用其他参数。
	101030	发生了未预期的错误。	请联系客服解决。
	101034	请提供相关配置用于Invokable管理。	请联系客服解决。
	101035	请提供相关配置用于AgentIR管理。	请联系客服解决。
	101036	未找到指定的模型类。	请联系客服解决。
	101037	通知会话失败。	请联系客服解决。
	101052	组件步骤调试错误。	请联系客服解决。
	101053	组件类型不支持步骤调试。	请联系客服解决。
	101054	单个组件IR非法。	请联系客服解决。
	101055	userFields的输入类型错误。	userFields的输入应为字典类型。
	101056	验证输入失败。	请联系客服解决。
	101057	检查输入输出失败。	输出字段{key}应与输入字段相同。
	101058	全局项已存在且不是字典类型。	请检查相关配置，如果无法解决请联系客服解决。

模块名称	错误码	错误描述	解决方案
	101099	加载配置失败。	请联系客服解决。
	101100	configs中userFields定义的字段在inputs中不存在。	请检查相关配置，如果无法解决请联系客服解决。
	101101	该插件的函数错误。	该插件的函数应为字符串列表/元组/集合。
	101102	未知错误。	请联系客服解决。
	101103	RestFulApi或Function错误。	请输入RestFulApi或Function的列表。
	101104	发现错误的字段名。	请联系客服解决。
	101105	Plan配置初始化失败。	若重试无法解决，请联系客服解决。
	101106	Model初始化失败。	若重试无法解决，请联系客服解决。
	101107	Plugin初始化失败。	若重试无法解决，请联系客服解决。
	101_2_00	模型验证错误。	请联系客服解决。
	101_2_01	组件JSON错误。	组件JSON仅能为组件列表或单个组件（根类型为列表或字典）。
	101_2_60	子Agent IR部分失败。	请联系客服解决。
	101_2_61	子Agent IR全部失败。	请联系客服解决。
	101_2_85	复合Invokable缺少AgentIR构建器和可选IR。	请联系客服解决。
	101_2_86	复合InvokableAgentIR构建器未找到。	请联系客服解决。
	101_2_87	复合Invokable的AgentIR构建器导入失败。	请联系客服解决。

模块名称	错误码	错误描述	解决方案
	101_2_88	复合Invokable的AgentIR构建器不是IR构建器。	请联系客服解决。
	101_2_89	复合Invokable的AgentIR构建器构建失败。	请联系客服解决。
	101_2_90	内置Invokable注册失败。	请联系客服解决。
Prompt引擎	102001	LLM服务返回结果为None。	若重试无法解决，请联系客服解决。
	102002	HTTP请求错误。	若重试无法解决，请联系客服解决。
	102003	LLM服务返回错误结果。	若重试无法解决，请联系客服解决。
	102004	连接LLM服务失败。	若重试无法解决，请联系客服解决。
	102005	无法从上下文中获取环境变量。	若重试无法解决，请联系客服解决。
	102006	初始化输出解析器失败。	请联系客服解决。
	102007	调用输出解析器失败。	请联系客服解决。
	102008	MEP返回错误结果。	请联系客服解决。
	102009	无法获取运行时上下文。	请联系客服解决。
	102010	构建IR失败。	若重试无法解决，请联系客服解决。
	102011	LLM服务类型无法识别。	请联系客服解决。
	102012	LLM组件接收到运行时上下文的中断信号。	请联系客服解决。
	102050	初始化变量参数错误。	请联系客服解决。
	102051	初始化模板组装器参数错误。	请联系客服解决。

模块名称	错误码	错误描述	解决方案
	102052	更新模板或变量时缺少或包含不期望的键值对。	请检查相关配置。
	102053	模板内容格式错误，导致格式化失败。	请检查模板内容格式。
	102054	初始化提示组件失败。	请检查相关配置。
	102055	调用提示组件失败。	请联系客服解决。
	102056	变量格式化错误，JSON Schema格式非法。	请联系客服解决。
	102101	模板重复。	若重试无法解决，请联系客服解决。
	102102	模板未找到。	若重试无法解决，请联系客服解决。
	102103	模板数据不正确。	请修正模板数据。
	102104	从模板存储中加载数据失败。	若重试无法解决，请联系客服解决。
	102150	LLM变异操作未能生成与原始模板相同的占位符。	请联系客服解决。
	102151	LLM交叉操作未能生成与父模板相同的占位符。	请联系客服解决。
	102152	案例消息中的模板占位符与任何原始模板不匹配。	请联系客服解决。
	102153	案例消息中的模板内容与任何原始模板不匹配。	请联系客服解决。
	102154	提示优化参数非法。	请联系客服解决。
	102155	未找到提示优化任务。	请联系客服解决。

模块名称	错误码	错误描述	解决方案
	102156	提示优化任务状态不符合预期。	请联系客服解决。
	102157	提示优化评估失败。	请联系客服解决。
	102158	提示优化任务启动失败。	请联系客服解决。
	102159	提示优化任务重启失败。	请联系客服解决。
	102160	无法从消息历史中获取标签。	请联系客服解决。
	102161	提示优化输入案例验证失败。	请联系客服解决。
	102170	提示优化存储失败。	请联系客服解决。
	102201	元模板未在支持列表中找到。	请联系客服解决。
	102202	元模板不存在。	请联系客服解决。
	102203	元模板输入工具解析失败。	请联系客服解决。
	102204	元模板构建失败。	请联系客服解决。
	102205	元模板输入缺少必要字段。	请联系客服解决。
	102206	LLM配置类型不被支持。	请修改配置类型。
	102207	生成LLM结果失败。	若重试无法解决，请联系客服解决。
	102208	编辑模板失败。	请联系客服解决。
	102209	匹配失败。	请联系客服解决。
	102210	初始化Prompt配置失败。	若重试无法解决，请联系客服解决。
	102211	捕获到未知异常。	请联系客服解决。

模块名称	错误码	错误描述	解决方案
	102212	SSL验证环境JSON解码异常。	请联系客服解决。
	102213	优化反馈异常。	请联系客服解决。
	102214	优化坏案例异常。	请联系客服解决。
规划引擎	103001	检查Planner配置失败。	请联系客服解决。
	103002	未输入Prompt名称。	请输入Prompt名称。
	103003	未输入模型信息。	请输入模型信息。
	103004	调用模型失败。	若重试无法解决，请联系客服解决。
	103005	根据指定的计划模式类型获取计划状态类失败。	请联系客服解决。
	103006	使用正则表达式解析段落失败。	请联系客服解决。
	103007	通过索引获取元素失败。	请联系客服解决。
	103008	从全局配置中获取插件配置值失败。	请联系客服解决。
	103009	由于非法的比较表达式，无法返回布尔结果。	请联系客服解决。
	103010	根据指定的计划模式类型获取模板IR文件失败。	请联系客服解决。
	103011	执行Prompt模板格式失败。	请联系客服解决。
	103012	获取技能配置值失败。	请联系客服解决。
	103013	获取技能类型失败。	请联系客服解决。
	103011	执行Planner rails失败。	请联系客服解决。

模块名称	错误码	错误描述	解决方案
	103014	解析工作流起始组件失败。	请联系客服解决。
	103015	运行工作流组件失败。	请联系客服解决。
	103016	根据指定版本获取模板IR文件失败。	请联系客服解决。
	103017	configuration类型非法。	请联系客服解决。
	103018	Planner流式处理异常。	请联系客服解决。
执行引擎	104001	流式运行错误。	请联系客服解决。
	104002	执行器错误。	请联系客服解决。
	104003	通用请求错误。	请联系客服解决。
	104004	执行操作符失败。	请联系客服解决。
	104005	从管理器获取工作节点列表失败。	请联系客服解决。
	104006	清理会话失败。	若重试无法解决，请联系客服解决。
	104007	执行用户输入失败。	请联系客服解决。
	104008	处理流数据失败。	请联系客服解决。
上下文引擎	109500	上下文配置错误。	请联系客服解决。
	109501	内存核心初始化失败。	请联系客服解决。
	109510	创建处理器失败。	请联系客服解决。
Insight	108003	父调用没有对应的子执行顺序索引。	请联系客服解决。

模块名称	错误码	错误描述	解决方案
	108004	调用ID不存在。	请检查调用ID是否正确，若无法处理请联系客服解决。
	108005	当前调用ID的类型错误。	期望为prompt类型。
	108006	Insight的x-insight-project-id为空。	Insight的x-insight-project-id不能为空。
动态规划	109011	环境变量JIUWEN_CUSTOM_STRATEGY_PATH未配置。	请正确配置环境变量JIUWEN_CUSTOM_STRATEGY_PATH。
	109012	策略路径不存在。	请联系客服解决。
	109013	策略路径不是一个有效的文件路径。	请检查策略路径是否正确，若无法处理请联系客服解决。
	109014	注册表未初始化。	请先调用initialize()进行初始化。
	109015	为提供者未找到策略。	请联系客服解决。
	109016	未找到相关策略。	请联系客服解决。
	109017	策略组件ainvoke发生未预期错误。	请联系客服解决。
	109018	Agent配置中的max_iteration配置错误。	max_iteration必须为大于0的整数。
	109019	未找到插件的URL。	请联系客服解决。