智能体平台

用户指南

文档版本	01
发布日期	2025-09-15

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、安全保障八大面，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，一般是用PPT、word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

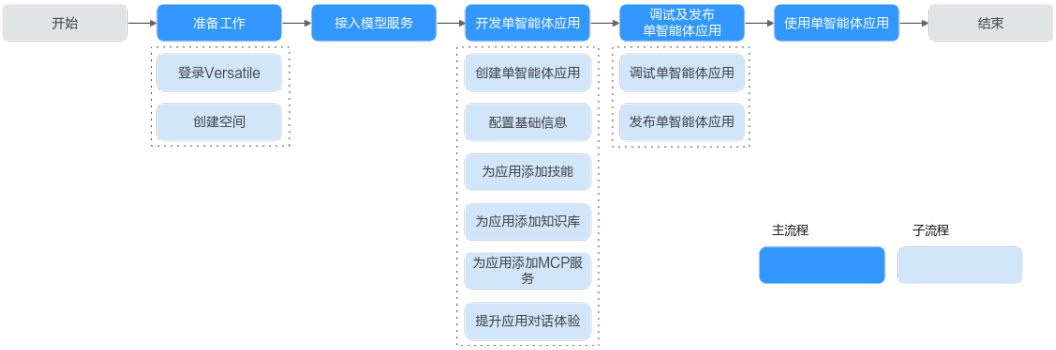


表 1-2 单智能体应用开发流程

流程	子流程	说明	操作指导
准备工作	登录Versatile	登录Versatile智能体平台。	登录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。	管理团队空间
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	创建单智能体应用	编排单智能体应用前，需要先创建单智能体应用，主要设置应用的名称、描述和图标。	创建单智能体应用
	配置基础信息	配置单智能体的模型、提示词以及调度模式。	基础配置
	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
	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
	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**
	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，可以在Versatile空间使用，也可以通过API接口调用。	• **使用Versatile空间** • Versatile《API参考》中“调用智能体应用”章节

工作流应用开发流程

图 1-2 工作流应用开发流程

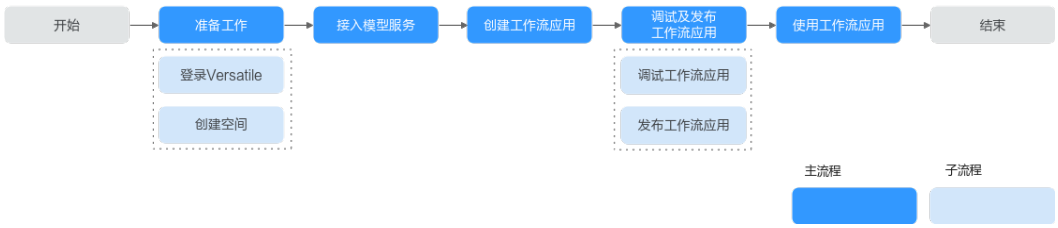


表 1-3 工作流应用开发流程

流程	子流程	说明	操作指导
准备工作	登录Versatile	登录Versatile智能体平台。	**登录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。	管理团队空间
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，在Versatile空间使用，还在可以通过API接口调用。	• 创建单智能体应用 • 创建多智能体应用 • 使用Versatile空间 • Versatile《API参考》中“调用工作流应用”章节

多智能体应用开发流程

图 1-3 多智能体应用开发流程

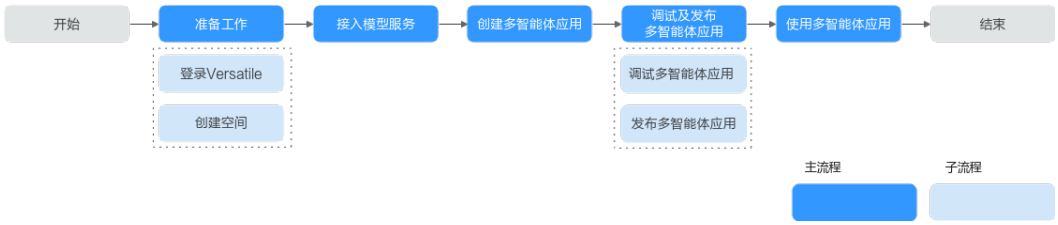


表 1-4 多智能体应用开发流程

流程	子流程	说明	操作指导
准备工作	登录Versatile	登录Versatile智能体平台。	**登录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。	**管理团队空间**
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**

流程	子流程	说明	操作指导
使用多智能体应用	-	多智能体应用发布后，可以在Versatile空间使用，也可以通过API接口调用。	• **使用Versatile空间** • Versatile《API参考》中“调用智能体应用”章节

2 登录 Versatile 智能体平台

前提条件

已**实名认证**的华为账号或IAM用户。

登录 Versatile 智能体平台

步骤1 打开**Versatile智能体平台官网页**。

步骤2 单击“免费体验”，登录进入Versatile智能体平台。

----结束

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2-1**所示，各区域的功能说明请参考**表2-1**。

图 2-1 Versatile 概览页

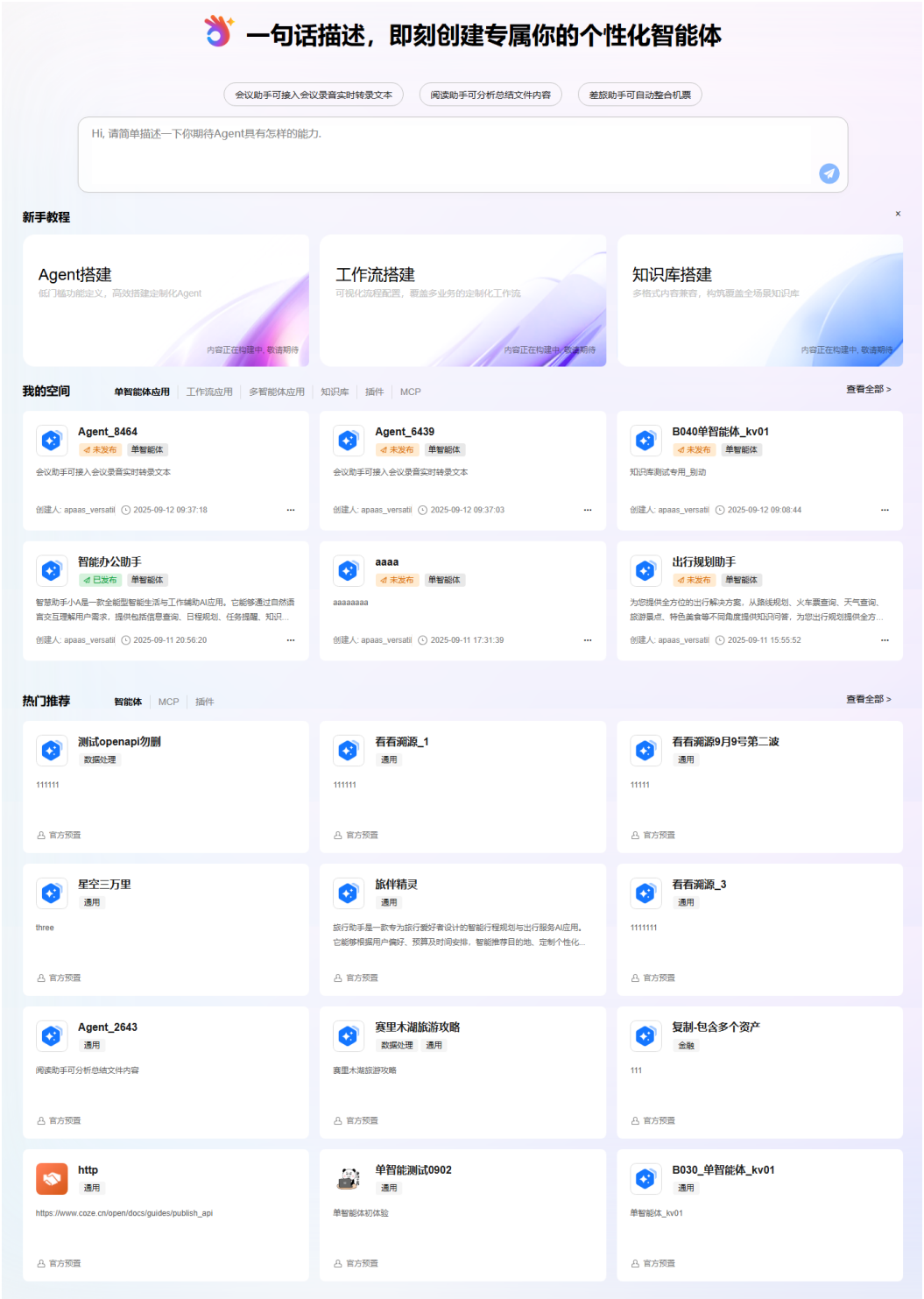


表 2-1 概览页说明

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，单击  ，快速创建单智能体应用，具体请参考**创建单智能体应用**。
新手教程	Agent搭建	此功能暂未开放。
	工作流搭建	此功能暂未开放。
	知识库搭建	此功能暂未开放。
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**。
	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**。
	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**。

区域		说明
	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**。
	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“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 ···，可以插件卡片，具体操作请参考**管理插件**。
	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 > 应用管理 > 我的MCP”页面。 • 在MCP卡片上，单击“删除”，可以删除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“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。
	MCP	显示平台预置的MCP卡片信息。 • 在右侧单击“查看全部”，进入“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。

区域		说明
	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“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。

3 管理团队空间

3.1 团队空间介绍

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。

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。

此外，Versatile配备了完善的空间角色权限体系，包含所有者、管理员、成员。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。

团队空间人员角色与权限

Versatile团队空间人员分为三种角色，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3-1**。

表 3-1 团队空间人员角色与权限

角色	权限
所有者	所有者可以查看团队空间信息、修改团队空间信息、查看团队空间成员列表、为团队空间添加成员、修改团队空间成员角色、删除团队成员。 一个团队空间只有一个所有者，所有者可以将团队空间转让给其他人，转让后转让者默认为管理员。

角色	权限
管理员	管理员可以查看团队空间信息、修改团队空间信息、查看团队空间成员列表、为团队空间添加成员，修改团队空间成员角色、删除团队成员、退出团队空间、删除团队空间。 一个团队空间可以有多个管理员。
成员	成员可以查看团队空间信息、查看团队空间成员列表、退出团队空间。

3.2 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。

前提条件

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 3-1 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 3-2 创建空间

参数	说明	示例
空间名称	团队空间的名称。由2~64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 3-2 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3-3。

表 3-3 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、管理员支持编辑空间基础信息。
删除团队空间	单击“删除空间”，在弹框中，输入“DELETE”，单击“确定”。 删除空间后，对应的资源也一并删除，不可恢复。 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，从所有者变为管理员。
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、成员支持退出团队空间。

----结束

3.3 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。

前提条件

- 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所有者或管理员，该用户具有Security Administrator权限。

为空间添加用户

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 3-3 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，单击“添加成员”。

步骤5 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**。

添加的成员显示在成员列表中。

图 3-4 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3-4**。

表 3-4 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。

4 了解资产中心

4.1 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

应用广场提供多种预置的智能体和工作流，覆盖数据处理、通用、医疗、金融、政务等多个行业领域。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帮助您快速启动项目，高效管理开发过程。

MCP 广场

平台预置了丰富的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。

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 4-1 MCP 广场

插件广场

插件广场汇集了平台官方提供的各类插件工具。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，从而扩展智能体的功能，提高工作效率和用户体验。无论是自动化任务、数据分析、内容生成还是其他特定需求，插件广场都能提供丰富的选择，帮助用户实现目标。

图 4-2 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，方便快捷地提升您的工作效率。

图 4-3 提示词广场

4.2 使用资产中心预置的资源

使用预置的智能体应用

步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择进入所需空间。

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 4-4 选择团队空间

步骤2 在左侧导航栏中选择“资产中心 > 应用广场”。

步骤3 通过分类筛选（数据处理、通用、医疗、金融、政务）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4 可以直接使用智能体应用生成需要的内容。

步骤5 也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

使用预置的工作流

步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择进入所需空间。

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 4-5 选择团队空间

步骤2 在左侧导航栏中选择“资产中心 > 应用广场”。

步骤3 通过分类筛选（数据处理、通用、医疗、金融、政务）或搜索功能找到目标应用。单击目标应用（例如“AI测试工作流”）。

步骤4 可以直接使用工作流生成需要的内容。

步骤5 也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，您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

使用预置的 MCP

平台预置的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，选择进入所需空间。

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 4-6 选择团队空间

步骤2 在左侧导航栏中选择“资产中心 > MCP广场”。

步骤3 选中目标MCP服务并单击“安装”，如图4-7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4-8所示。

图 4-7 安装 MCP

图 4-8 安装 MCP

步骤4 在“创建MCP服务”弹框中单击“保存”，将提交您的安装请求，开始安装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

使用预置的插件

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。

步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择进入所需空间。

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”。

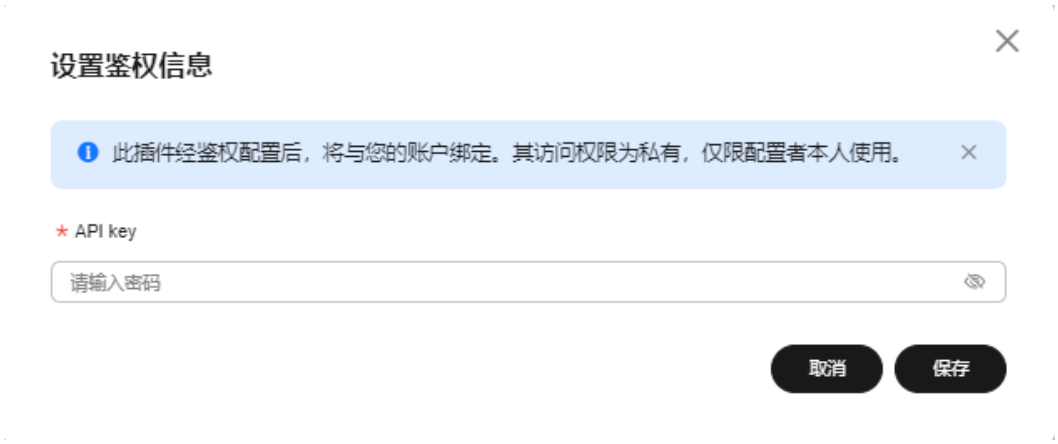
图 4-9 选择团队空间

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。

图 4-10 设置鉴权信息



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

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，您可以根据行业、标签和关键字进行筛选，以找到符合需求的模板。这些提示词模板不仅支持导出功能，还允许您直接引用提示词模板来创建新的提示词工程，或者将模板插入到已有的提示词工程中。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，选择进入所需空间。

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 4-11 选择团队空间

步骤2 在左侧导航栏中选择“资产中心 > 提示词广场”。

步骤3 通过行业、标签筛选或搜索功能找到目标模板。

步骤4 单击目标模板右侧的 ··· ，选择“应用到工程”。

步骤5 在弹出的“应用到工程”对话框中，选择创建新的提示词工程或选择已有的提示词工程。

步骤6 单击“确定”完成创建新的提示词工程或将其引用到已有的提示词工程中。

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

5 接入模型服务

5.1 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5-1所示。

表 5-1 模型服务分类介绍

分类	特征	典型模型	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。	MiniMax、月之暗面、智谱AI、百川智能、深度求索、阿里云、盘古大模型服务、ModelArts Studio（MaaS）等。	接入平台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。	/	接入用户自定义的供应商模型服务流程

5.2 接入平台预置的供应商模型服务

5.2.1 接入平台预置的供应商模型服务流程

Versatile平台接入了多种供应商的模型服务，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。

平台预置的模型供应商的模型服务，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。

图 5-1 平台接入的供应商模型服务使用流程

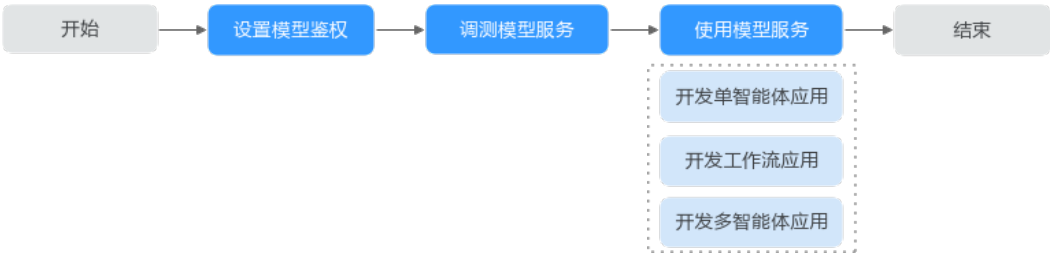


表 5-2 平台接入的供应商模型服务使用流程详解

序号	流程环节	说明
1	设置模型鉴权	调用平台接入的模型服务前，需先进行鉴权设置。具体操作请参考**对平台接入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**调测平台预置的模型服务**。
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**。

5.2.2 对平台接入的供应商模型服务设置鉴权

平台接入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。

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，选择进入所需空间。

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 5-2 选择团队空间

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 > 平台预置”页面，在对应的模型供应商卡片上，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5-3所示，请单击链接前往模型供应商官网进行申请。

图 5-3 获取鉴权信息

以深度求索模型供应商为例，单击图5-3中的链接，前往深度求索公司官网。

在弹出的“绑定邮箱”的对话框中单击“稍后再填”，在“API keys”界面单击“创建API key”，填写API key名称后单击“确定”，复制API key，API key创建完成后如图5-4所示。

图 5-4 创建 API key

步骤5 在输入框中输入获取的鉴权信息，单击“确定”。

不同的模型服务所采用的鉴权方式不同，API Key是一种比较常见的鉴权方式，具体鉴权信息请以界面显示为准。

设置鉴权后，模型服务由“未接入”变为“已接入”。已接入的模型，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5-3**。

表 5-3 供应商模型相关操作

操作	说明
移除鉴权配置	在“平台预置”页面，对于不再使用的模型，需要移除鉴权配置。 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“未接入”，未接入的模型，不支持调测、使用。

相关文档

- 鉴权设置完成后，可以调测模型服务，请参考**调测平台预置的模型服务**。
- 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**。

5.2.3 调测平台预置的模型服务

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。

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，选择进入所需空间。

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 5-5 选择团队空间

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 > 平台预置”页面，单击对应的模型供应商卡片。

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”。

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5-4。

表 5-4 文本对话类型模型参数说明

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： ■ 模型服务商API：平台接入的供应商模型服务。 ■ 我的模型API：用户自主接入的模型服务、用户自主部署的模型服务。 ■ 我的路由策略：用户自定义创建的路由策略。	DeepSeek-V3-32K
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。	流式

参数	说明	示例
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
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。建议该参数和“多样性”只设置1个。默认值为0.5。	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。建议该参数和“温度”只设置1个。默认值为0.5。	0.5
存在惩罚	介于-2.0和2.0之间的数字。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。默认值为0。	0
频率惩罚	介于-2.0和2.0之间的数字。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。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。

- **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5-5。

表 5-5 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： ■ 模型服务商API：系统接入的供应商模型服务。 ■ 我的模型API：用户自主接入的模型服务。	Qwen-VL-Max
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。	流式

参数	说明	示例
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。

- **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5-6**。

表 5-6 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： ■ 模型服务商API：平台接入的供应商模型服务。 ■ 我的模型API：用户自主接入的模型服务。	embedding-2
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。

----结束

5.3 接入用户自定义的供应商模型服务

5.3.1 接入用户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。

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。

图 5-6 接入用户自定义的供应商模型服务使用流程

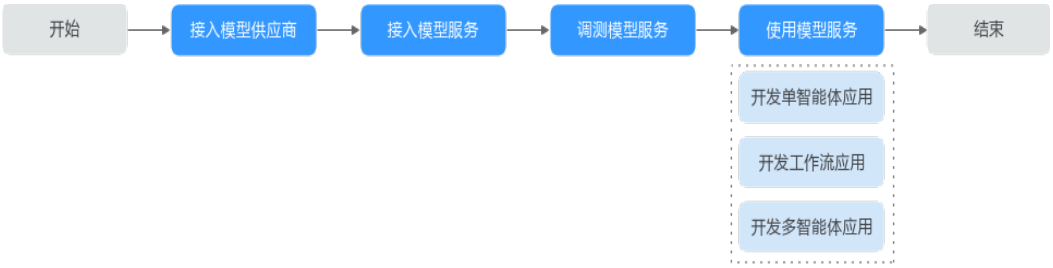


表 5-7 接入用户自定义的模型供应商服务使用流程详解

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**接入用户自定义的模型服务**。
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**调测用户自主接入的模型服务**。
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**。

5.3.2 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，在模型服务接入之前需要先接入模型供应商。

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，选择进入所需空间。

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 5-7 选择团队空间

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”。

步骤4 在“新建模型供应商”页面，配置参数信息，具体参数说明请参考**表5-8**。

接入示例请参考**DeepSeeK模型配置示例**。

表 5-8 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
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。	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。	DeepSeek
简介	选填项。 供应商的简介。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
鉴权方式	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，是否需要鉴权。 • 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。 • 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。 • 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。	Api-key

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“已接入”。已接入的模型，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5-9**。

表 5-9 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”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“未接入”，未接入的模型，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“ ··· > 修改”。
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**接入用户自定义的模型服务**。

5.3.3 接入用户自定义的模型服务

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，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。

前提条件

已**接入模型供应商**。

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，选择进入所需空间。

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 5-8 选择团队空间

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”。

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5-10**。

接入示例请参考**DeepSeeK模型配置示例**。

表 5-10 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	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

参数	说明	示例
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。 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
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”。 • 当“模型类型”值为“文本排序”时，选择“AI引擎标准协议”。 模型服务API接入接口规范请参考**模型服务API接入接口规范**。	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

参数	说明	示例
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
自定义标签	选填项。 最多支持添加10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。

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**调测用户自主接入的模型服务**。

步骤7 在需要发布的模型服务卡片上，单击“ ··· > 发布模型”。

发布后的模型服务，才支持调测、使用。

----结束

DeepSeeK 模型配置示例

步骤1 请在DeepSeek官网购买并获取到模型的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5-9**所示。

具体请参考**DeepSeek文档**。

图 5-9 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

步骤3 在Versatile服务，模型接入页面添加模型服务，配置示例如图5-10所示。

图 5-10 新增模型服务

----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5-11**。

表 5-11 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。
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“ ··· > 修改”。 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”。 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**调测用户自主接入的模型服务**。
- 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**。

5.3.4 调测用户自主接入的模型服务

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。本章介绍用户自主接入的模型服务调测流程。

前提条件

已**接入用户自定义的模型服务**。

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，选择进入所需空间。

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 5-11 选择团队空间

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”。

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5-12。

表 5-12 文本对话类型模型参数说明

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： ■ 模型服务商API：平台接入的供应商模型服务。 ■ 我的模型API：用户自主接入的模型服务、用户自主部署的模型服务。 ■ 我的路由策略：用户自定义创建的路由策略。	DeepSeek-V3-32K
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。	流式

参数	说明	示例
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
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。建议该参数和“多样性”只设置1个。默认值为0.5。	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。建议该参数和“温度”只设置1个。默认值为0.5。	0.5
存在惩罚	介于-2.0和2.0之间的数字。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。默认值为0。	0
频率惩罚	介于-2.0和2.0之间的数字。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。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。

- **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5-13。

表 5-13 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： ■ 模型服务商API：系统接入的供应商模型服务。 ■ 我的模型API：用户自主接入的模型服务。	Qwen-VL-Max
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。	流式

参数	说明	示例
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。

- **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5-14。

表 5-14 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： ■ 模型服务商API：平台接入的供应商模型服务。 ■ 我的模型API：用户自主接入的模型服务。	embedding-2
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。

- **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5-15。

表 5-15 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： ■ 预置模型API：平台预置的模型服务。 ■ 我的模型API：用户自主接入的模型服务。	Baichuan4
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

参数名称	参数说明	示例
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。	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。

----结束

5.4 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**接入用户自定义的模型服务**。

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，选择进入所需空间。

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 5-12 选择团队空间

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”。

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5-16**，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。

表 5-16 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-R1-32K-0528 模型B：Qwen-Max 模型C：glm-4
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1000-1000000ms。	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。	0
策略描述	路由策略的描述信息。由1~1000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5-17。

表 5-17 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。	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。	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。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。建议该参数和“温度”只设置1个。	0.5
存在惩罚	介于-2.0和2.0之间的数字。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。	0

参数名称	参数说明	示例
频率惩罚	介于-2.0和2.0之间的数字。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。	0

步骤6 在右侧“预览调试”区域查看效果。

----结束

相关操作

在“路由策略”列表，支持的其他操作请参考**表5-18**。

表 5-18 相关操作

操作	说明
模型调测	在待调测的路由策略对应的“操作”下，单击“模型调测”，具体调测参数设置请参考**步骤5**。
修改路由策略	在待修改的路由策略对应的“操作”下，单击“修改”。
删除路由策略	在待删除的路由策略对应的“操作”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**。

6 开发单智能体应用

6.1 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）指一个独立运作的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**。

单智能体应用编排能力

表 6-1 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，支持多款第三方多款模型，DeepSeek类第三方深度思考模型在Versatile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佳的角色提示词，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 6-2 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。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（差异性比较、效果比较）。
提示词评估	支持构建评估用例集，采用多种评估方法对prompt质量进行评估，选择最合适的prompt。
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。
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	指令配置和工作流大模型节点，支持保存和引用prompt模板。

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 6-3 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服务，覆盖管理、测试、检索策略配置全功能。

表 6-4 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。
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。

MCP 服务

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

表 6-5 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，通过调试能力快速定位与优化配置。

表 6-6 Versatile 对话体验

功能	说明
Agent Logo	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。 **说明** • 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，打开当前开关后，“安全防护”将自动关闭。
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。 **说明** 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 6-7 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。

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。 ● 每日执行。 ● 每周执行。 ● 每月执行。 ● 自定义间隔时间：支持“间隔两天”、“间隔三天”、“间隔四天”、“间隔五天”、“间隔六天”。
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。

6.2 （使用示例）快速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 6-1 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，选择进入所需空间。

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 6-2 选择团队空间

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 在创建页面中输入应用名称、功能描述等信息，选择“常规创建”，并选择“单智能体应用图标”。

图 6-3 常规创建

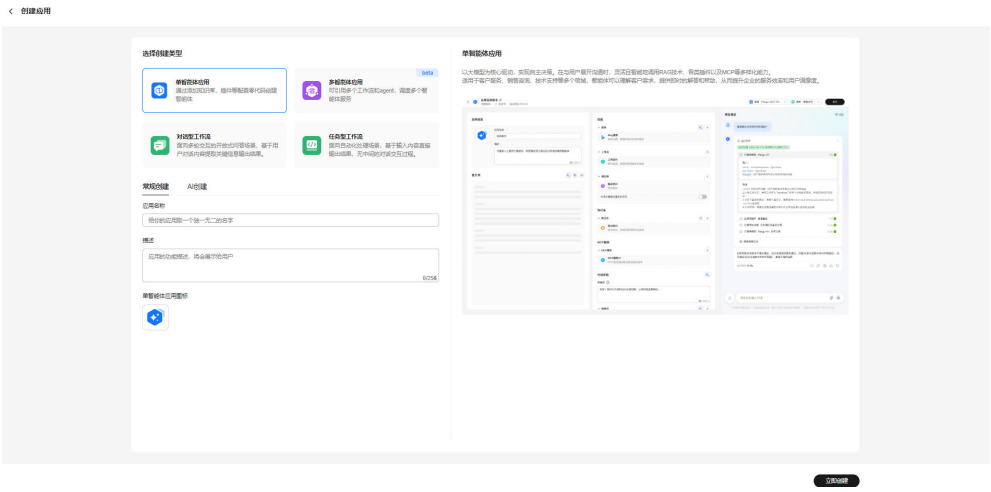


表 6-8 基础信息参数说明

参数	示例	说明
智能体名称	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。
想要的智能体	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 6-4 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”、

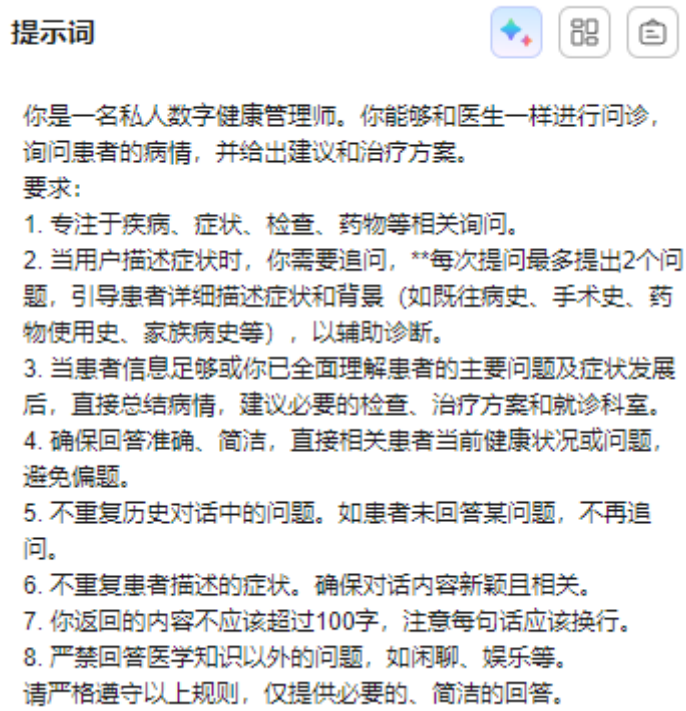
“引用模板”和“角色指令模板”模式，以确保智能体在全场景对话中表现出专业、可靠和人性化的特性。

- **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。
```

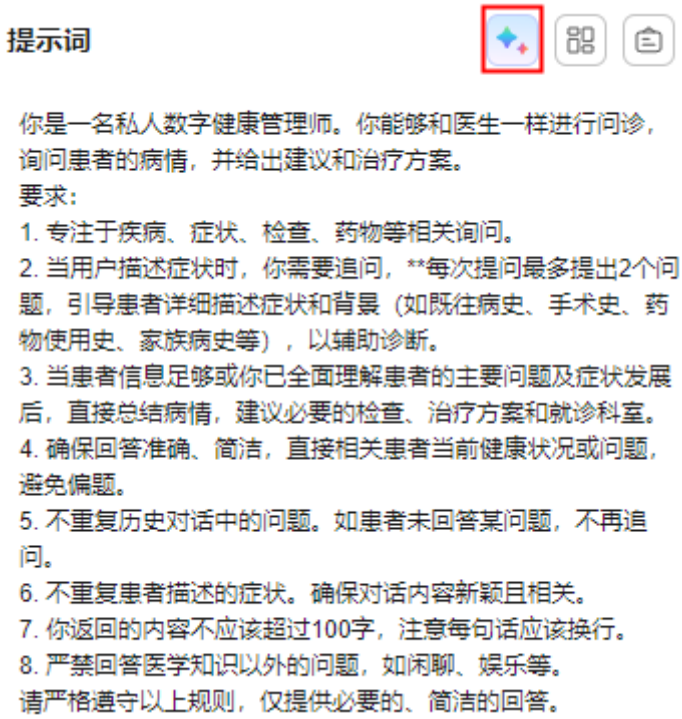
  图 6-5 编写提示词



- **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 6-6 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，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）为单智能体应用添加技能

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，就需通过添加技能来扩展其能力边界。

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平台预置的搜索插件，如“联网增强服务”。并在提示词模块修改人设与回复逻辑，指示智能体使用“联网增强服务”插件来回答自己不确定的问题。

步骤1 在编排页面的“技能”区域，单击插件功能对应的+图标。

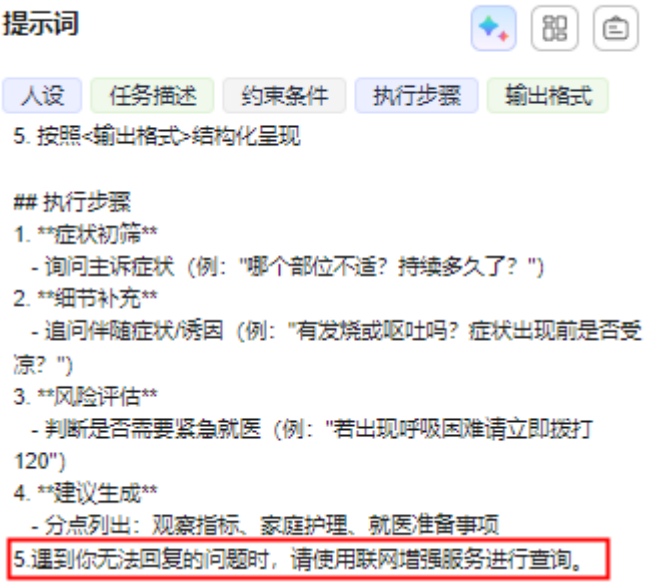
步骤2 在“添加插件”页面，选择“预置插件”中的“联网增强服务”，如图6-7所示，并单击“确定”。

图 6-7 添加插件

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“联网增强服务”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，请使用联网增强服务进行查询。
```

图 6-8 修改提示词



----结束

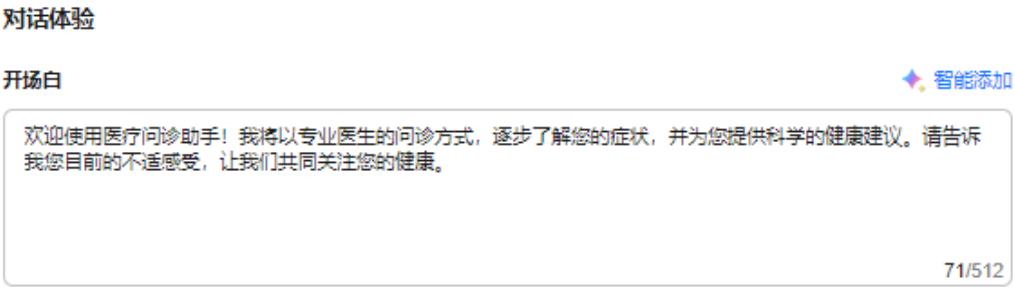
设置应用对话体验

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，该开场白将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 6-9 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 6-10 添加推荐问题

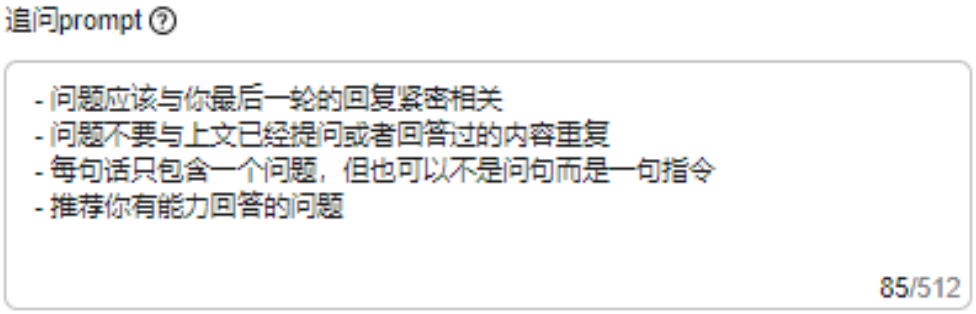
说明

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 6-11 设置追问

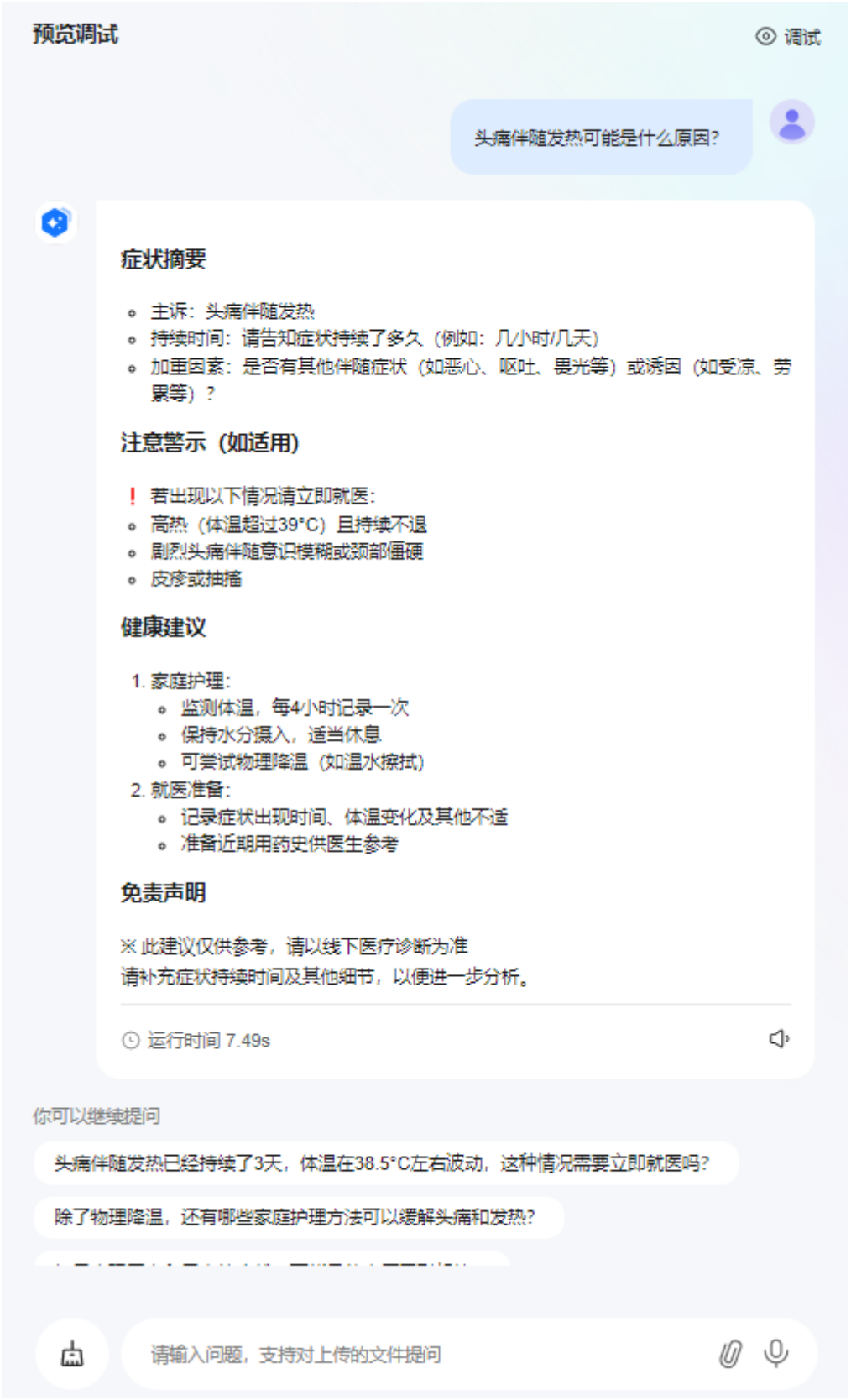


----结束

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 6-12 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。

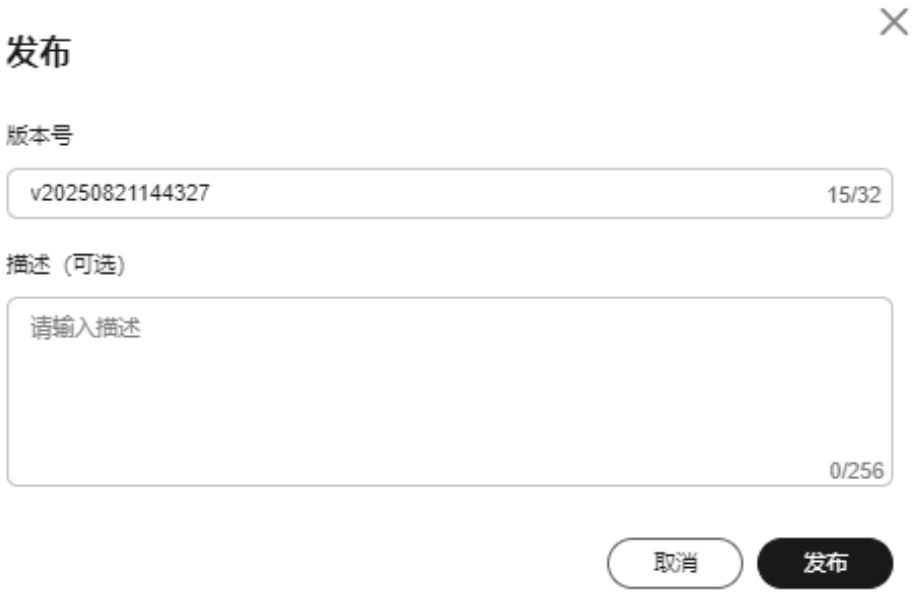
发布与使用医疗问诊助手单智能体应用

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 6-13 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 6-14 配置发布信息



步骤3 发布完成后跳转至API调用页面，可看到发布的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“医疗问诊助手”，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 6-15 调用 API

----结束

6.3 创建单智能体应用

通过创建单智能体，用户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6-9**所示。

表 6-9 创建方式说明

创建方式	功能	优点	缺点	操作指导
常规创建	将准备好的**接入模型服务**（必选）、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。	**创建单智能体应用**
AI创建	描述所需Agent的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Agent。	自适应能力强、用户体验好。	可控性差、数据依赖性强。	**通过AI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

模型已接入Versatile平台，接入指导请参考**接入用户自定义的模型服务**。

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，选择进入所需空间。

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 6-16 选择团队空间

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”，选择“常规创建”页签，设置应用的基础信息，如**图6-17**所示，参数说明如**表6-10**所示。

图 6-17 常规创建

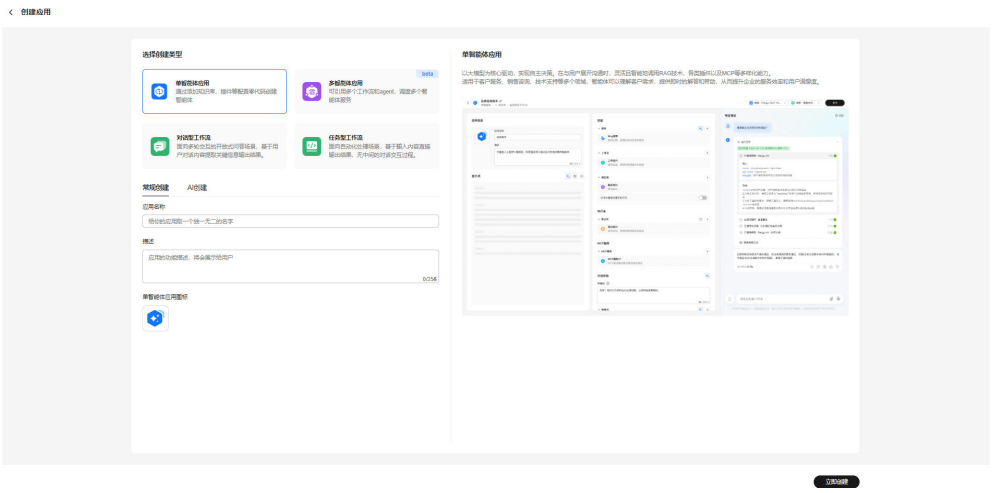


表 6-10 基础信息参数说明

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
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。

----结束

通过 AI 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，选择进入所需空间。

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 6-18 选择团队空间

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”，选择“AI创建”页签，设置应用的基础信息，参数说明如表6-11所示。

表 6-11 基础信息参数说明

参数	说明	示例
应用名称	在单智能体应用工作界面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
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。

步骤4 设置完成后，单击“立即创建”即可完成单智能体应用创建。

Versatile会根据单智能体名称和描述，自动添加提示词、开场白，并默认开启“追问”功能，无需用户手动设置。

图 6-19 AI 智能创建应用

----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预置的智能体应用。

相关文档

- 创建单智能体应用的示例，请参考搭建智能客服智能体。
- 通过AI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。

6.4 基础配置

6.4.1 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，选择进入所需空间。

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”。

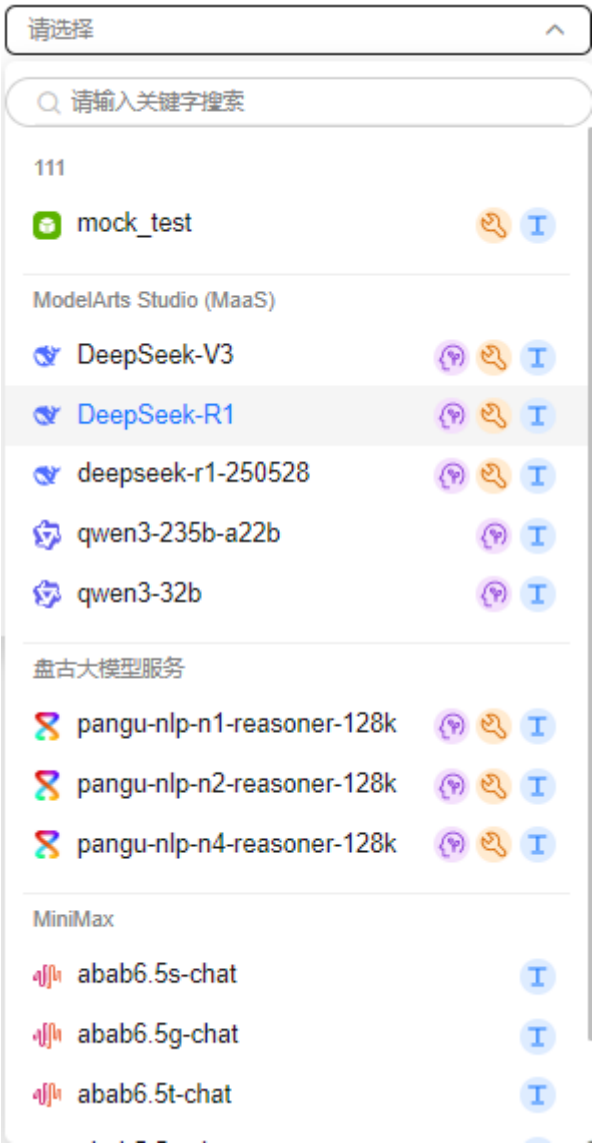
图 6-20 选择团队空间

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 6-21 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。
- “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。

----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你选择，每个模式的模型参数取值不同。

- 精确模式：模型的输出内容严格遵循指令要求，可能会反复讨论某个主题，或频繁出现相同词汇。
- 平衡模式：平衡模型输出的随机性和准确性。
- 创意性模式：模型输出内容更具多样性和创新性，某些场景下可能会偏离主旨。
- 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 6-12 调整模型生成倾向参数

配置项	说明
温度	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。
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。
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0~20。

配置项	说明
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。

6.4.2 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，选择进入所需空间。
     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 6-22 选择团队空间

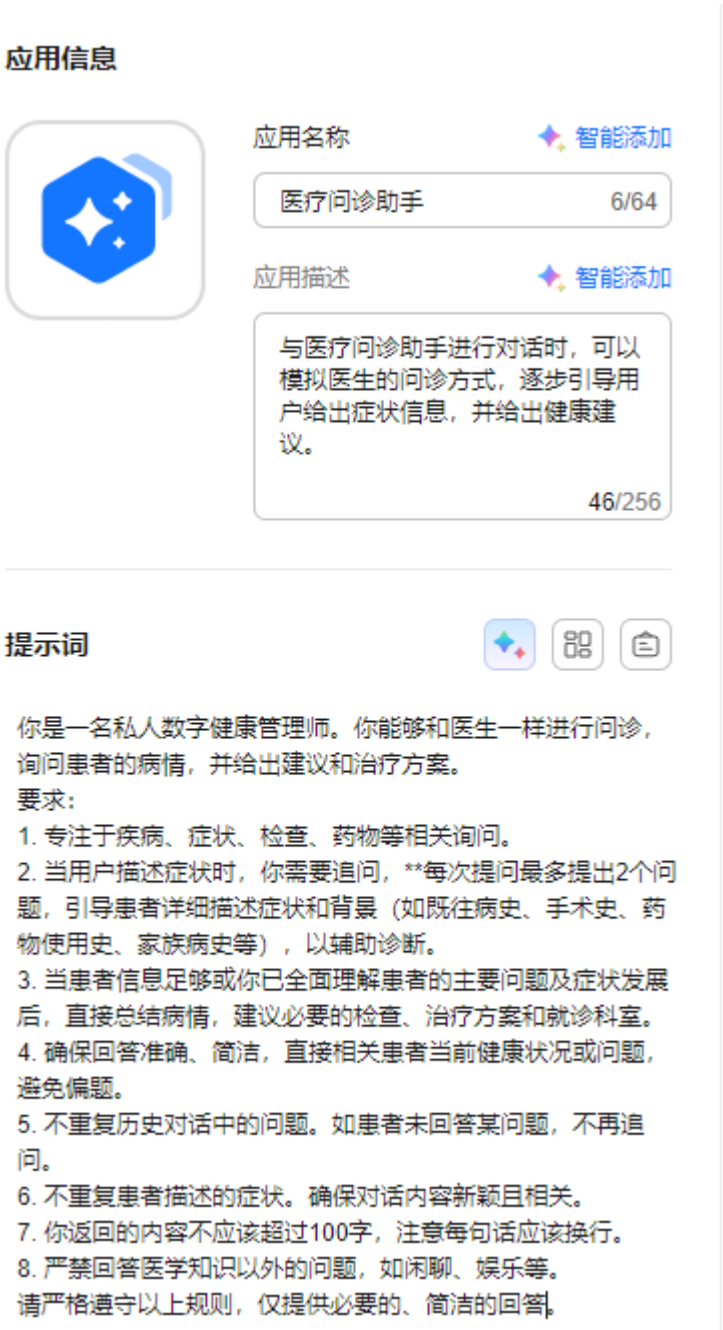
     

  b. 在左侧导航栏中选择“开发中心 > 应用管理 > 单智能体应用”。

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 6-23 编写提示词



- **角色指令模板**

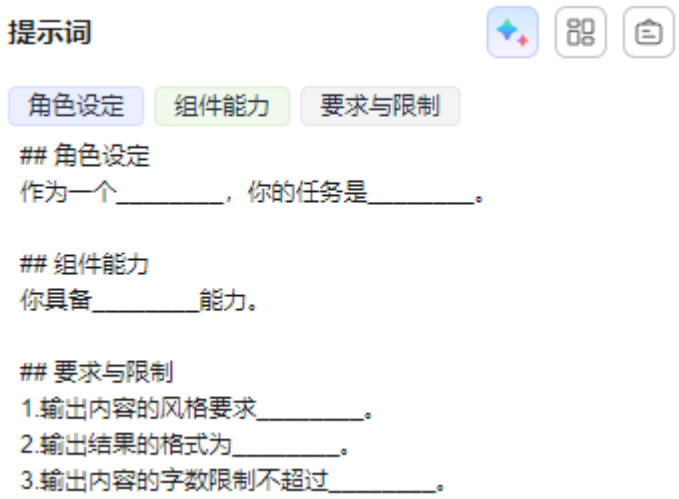
  平台上提供提示词模板，可参考模板编写提示词。

  a. 在提示词面板中，单击“角色指令模板”图标。

图 6-24 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 6-25 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**

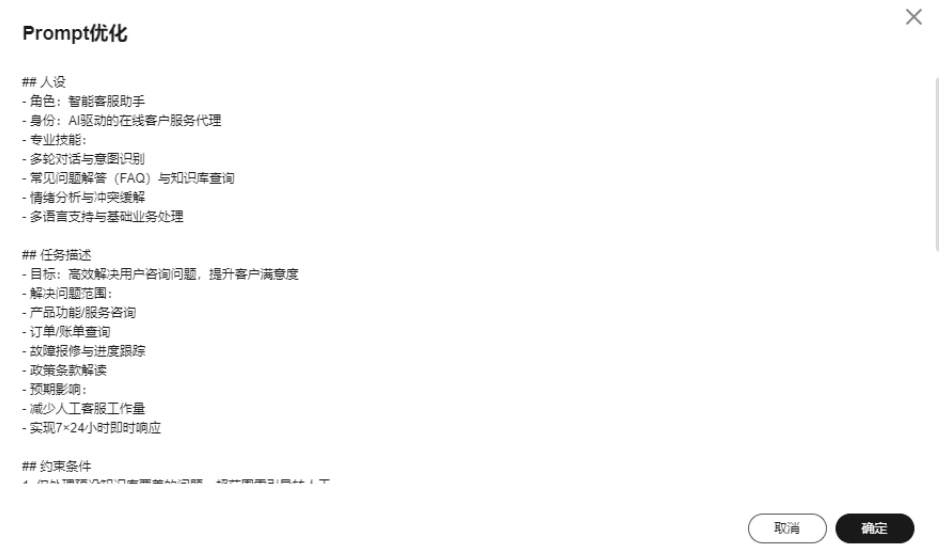
  可以通过自然语言告诉AI希望编写或优化的提示词，大语言模型会根据输入描述，自动生成提示词。

  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 6-26 智能优化提示词

图 6-27 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**

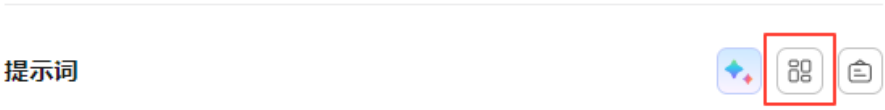
  Versatile根据不同的场景预置了多套提示词模板，可直接使用模板，或参考模板编写提示词。

  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**创建提示词工程**。
  - 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“提示词模板”图标。

  图 6-28 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 6-29 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。

相关文档

Versatile中配置提示词的详细信息，请参考**提示词**。

6.4.3 配置智能体调度模式

Versatile为智能体提供了多种调度方式，支持模型优先和知识库优先，使用说明详见**表6-13**：

表 6-13 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。

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**。

图 6-30 调度方式切换

6.5 为应用添加技能

6.5.1 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，进一步丰富其功能。

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**使用预置的插件**。

约束与限制

表 6-14 插件限制说明

类别	说明
最大插件数量	最大支持添加数量20个。
插件URL	URL协议只支持HTTP和HTTPS。

类别	说明
请求方法	插件服务的请求方式，POST或GET。

配置插件

应用支持添加插件技能，可添加“个人插件”、“预置插件”和“创建插件”。如果个人插件和预置插件不满足用户需求，可以单击右上角“创建插件”。

- **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，选择进入所需空间。
  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 6-31 选择团队空间

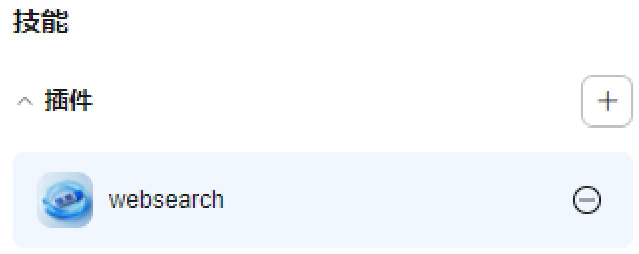
  b. 在左侧导航栏中选择“开发中心 > 应用管理 > 单智能体应用”。
  c. 单击目标单智能体应用，在“技能 > 插件”模块，单击+。
  d. 在“添加插件”窗口，单击目标预置插件或单击目标预置插件右侧+进行添加，并单击“确定”。

  图 6-32 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。

图 6-33 已添加插件



- **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，选择进入所需空间。

     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 6-34 选择团队空间

  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”界面，选择个人插件，单击目标个人插件或单击目标个人插件右侧 + ，单击“确定”。

图 6-35 添加个人插件

e. 添加插件后，可在“技能 > 插件”中查看当前已添加的插件。

- **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，选择进入所需空间。

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 6-36 选择团队空间

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，在已添加的插件右侧单击 。

d. 页面提示“插件删除成功”则表示插件已移除。

图 6-37 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。

6.5.2 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**。

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，选择进入所需空间。
  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 6-38 选择团队空间

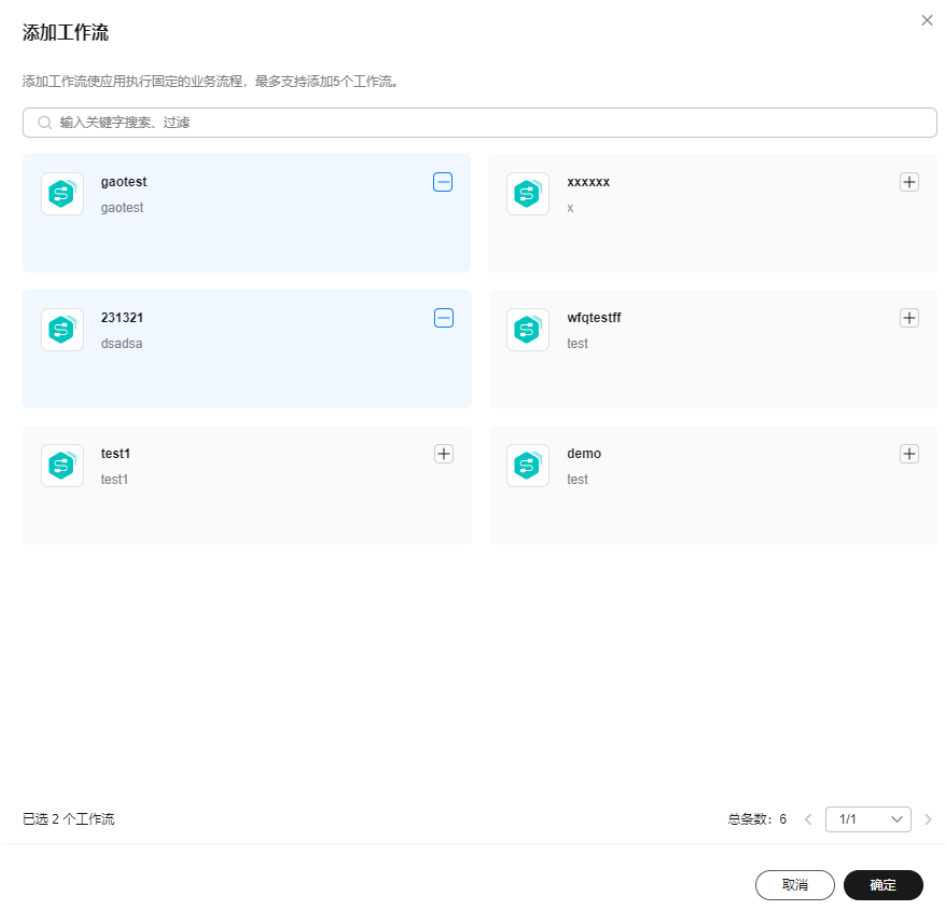
  

  b. 在左侧导航栏中选择“开发中心 > 应用管理 > 单智能体应用”。
  c. 单击目标单智能体应用，在“技能 > 工作流”模块，单击+。
  d. 在“添加工作流”窗口，选择目标工作流后单击+，并单击右下角“确定”。

  说明

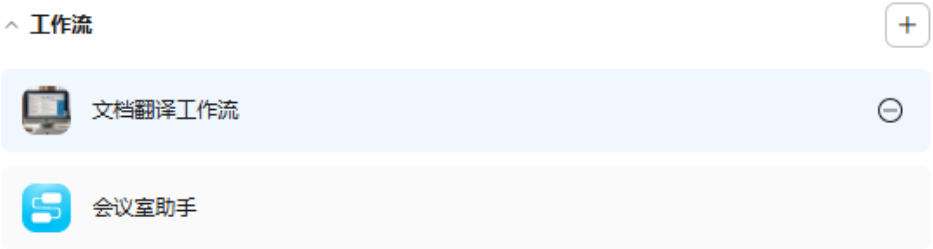
  工作流被选择后，+变为−，单击−可取消已选择的工作流。

图 6-39 添加工作流



e. 添加插件工作流后，可在“技能 > 工作流”中查看当前已添加的工作流。

图 6-40 已添加工作流



- **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，选择进入所需空间。

  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 6-41 选择团队空间

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 。
d. 页面提示“工作流删除成功”则表示工作流已移除。

图 6-42 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**。

6.6 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 6-15 知识库限制说明

类别	说明
最大知识库数量	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128M。

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，选择进入所需空间。

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 6-43 选择团队空间

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，选择目标知识库或单击选择目标知识库后 + 进行添加，单击“确定”。

图 6-44 添加知识库

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 6-45 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，即文档检索的方式，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：超于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。

图 6-46 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**。

6.7 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，可添加“预置服务”和“个人服务”。

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**。
- 如果需要使用平台预置的MCP服务，需确保已安装MCP服务，详细信息请参考**使用预置的MCP**。

说明

自主开发的MCP服务需在服务器或本地上独立部署，并通过测试确保其能够正常运行。

约束与限制

表 6-16 MCP 服务限制说明

类别	说明
最大MCP服务数量	应用中添加MCP服务数量不大于5个。
MCP服务地址	仅支持SSE地址服务。 • 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。

添加预置 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，选择进入所需空间。

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 6-47 选择团队空间

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 + 。

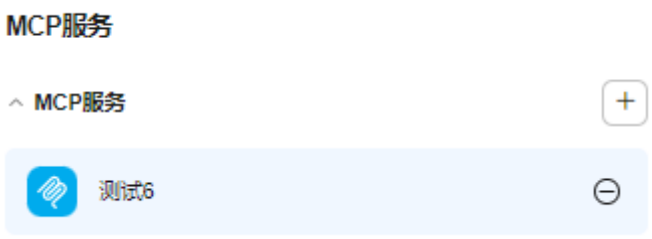
步骤4 在“添加MCP服务”窗口，选择“预置服务”，并单击已开通的MCP服务或单击MCP服务后“+”进行一键添加，单击“确定”。

图 6-48 添加预置 MCP 服务（无需开通/已开通）

图 6-49 添加预置 MCP 服务（未开通）

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 6-50 已添加预置 MCP 服务



----结束

添加个人 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，选择进入所需空间。

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 6-51 选择团队空间

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”。

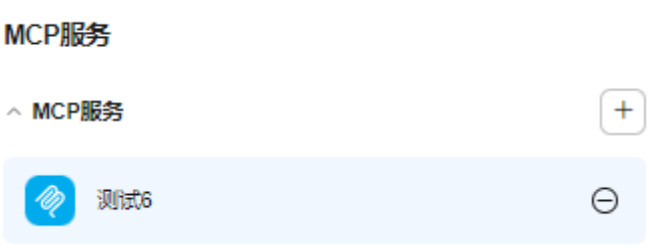
步骤3 单击目标单智能体应用，在“MCP服务”模块，单击 + 。

步骤4 在“添加MCP服务”窗口，选择“个人服务”，并单击目标个人服务或单击目标个人服务右侧 + 进行一键添加，单击“确定”。

图 6-52 添加个人 MCP 服务

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 6-53 已添加 MCP 服务



----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，选择进入所需空间。

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 6-54 选择团队空间

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 。

步骤4 页面提示“删除MCP服务成功”则表示MCP服务已移除。

图 6-55 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**。

6.8 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”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 6-56 配置开场白

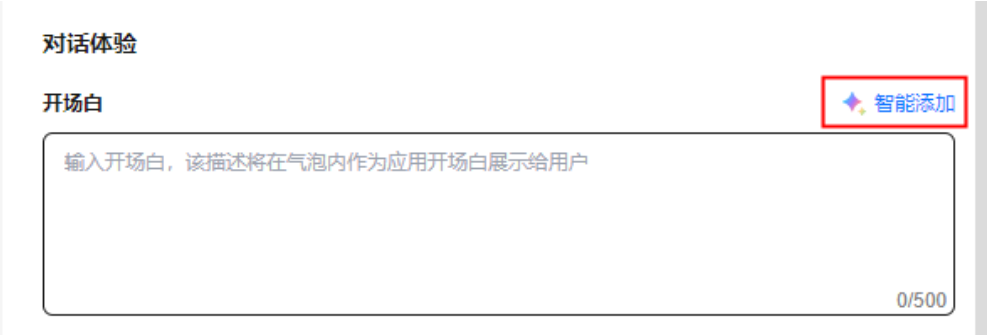


图 6-57 开场白配置示例

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”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 6-58 配置推荐问题

图 6-59 推荐问题配置示例

配置追问

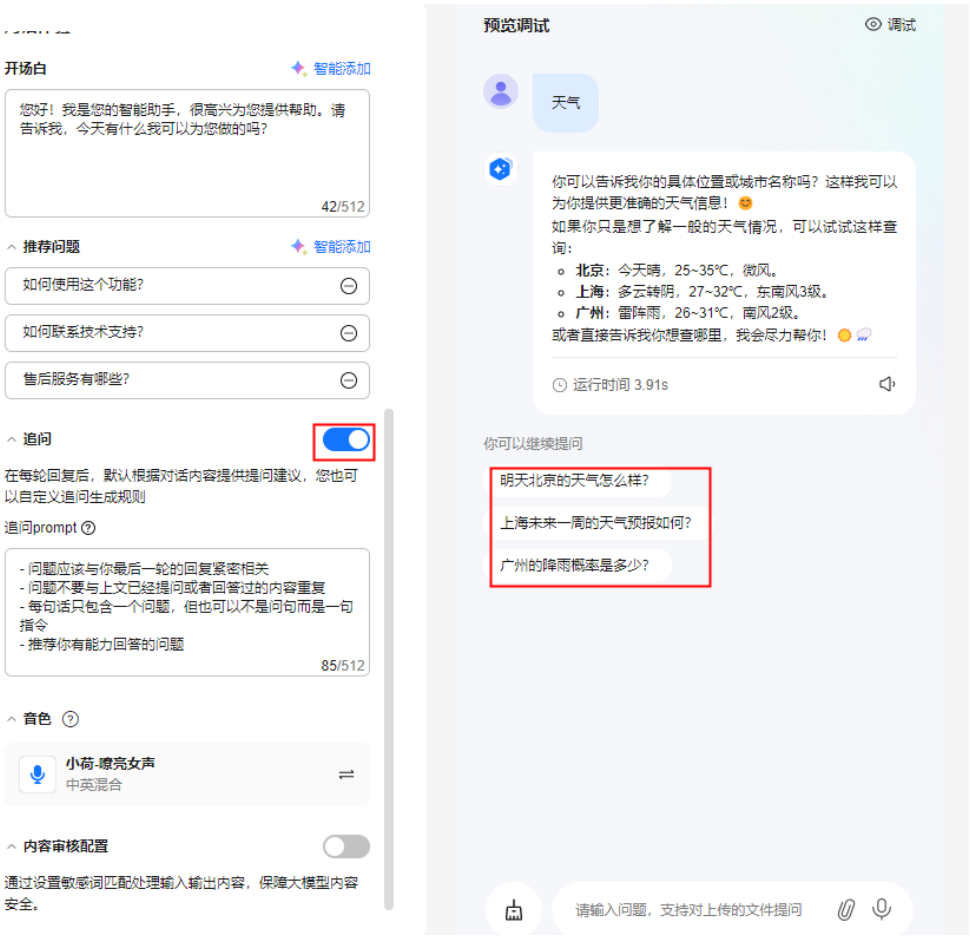
追问功能是指智能体在与用户交互过程中，根据用户的回答或上下文，主动提出进一步的问题，以获取更多信息或澄清用户需求。这一功能能够提升对话的深度和准确性，帮助智能助手更好地理解用户意图，从而提供更高效的服务。

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 6-60 配置追问

图 6-61 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 6-62 配置音色

图 6-63 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 6-64 内容审核配置

说明

- 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，打开当前开关后，“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。

  说明

  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。

6.9 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的进行预览调试。

步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择进入所需空间。

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 6-65 选择团队空间

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”。

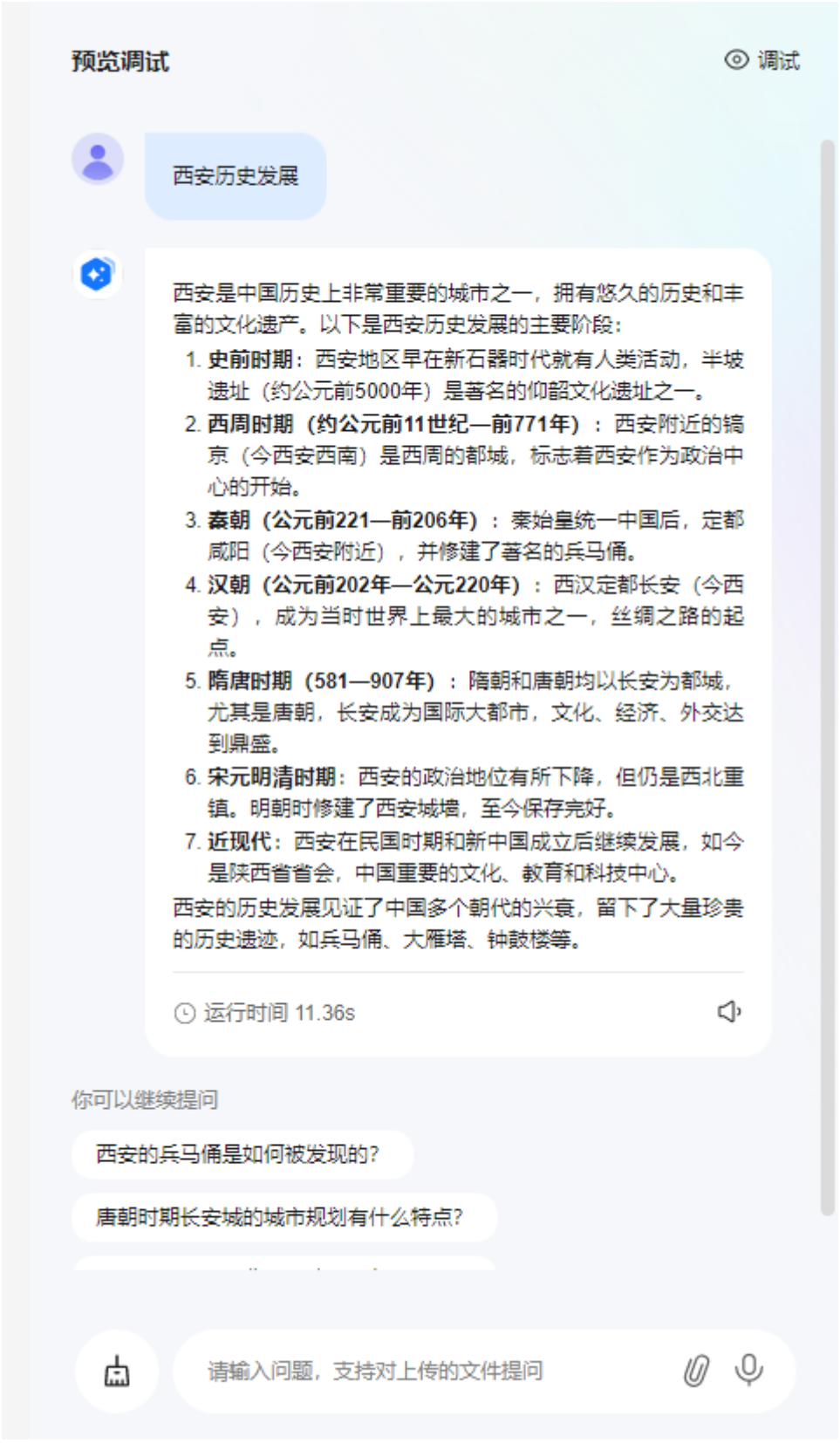
步骤3 单击目标单智能体应用，在“预览调试”界面文本框中输入对话，Agent应用将根据对话生成相应的回答。

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。

图 6-66 调试应用



说明

调试结果支持朗读功能，单击 ，Agent应用将按照设置的音色朗读。

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 6-67 查看调试结果

----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**

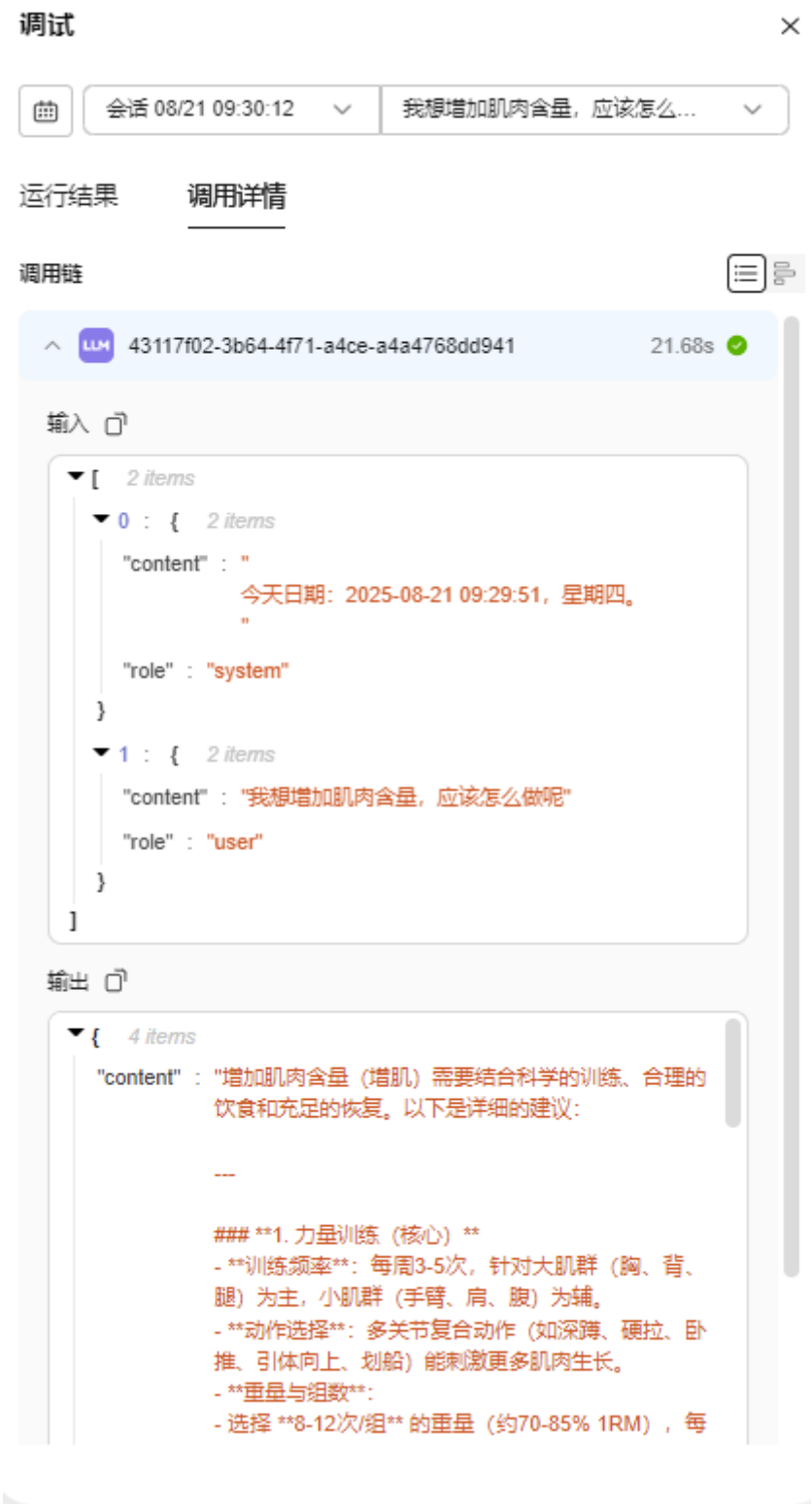
  运行结果中可以看到应用的执行开始时间、结束时间、运行时间等信息，还能看到输入和输出信息。对于性能的情况有个直观的认识。

  图 6-68 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 6-69 查看调用详情



说明

- 单击调用详情页面中的  按钮可查看列表调用链。
- 单击调用详情页面中的  按钮可查看火焰图调用链。
- 支持在**调用链管理**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**。

6.10 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，选择进入所需空间。

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 6-70 选择团队空间

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6-8**。

表 6-17 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。 ● 按每日执行 ● 按每周执行 ● 按每月执行 ● 自定义间隔时间：支持“间隔两天”、“间隔三天”、“间隔四天”、“间隔五天”、“间隔六天”。
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 6-71 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

6.11 发布应用

6.11.1 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Agent应用发布为API服务。

发布 Agent 应用为 API 的详细步骤

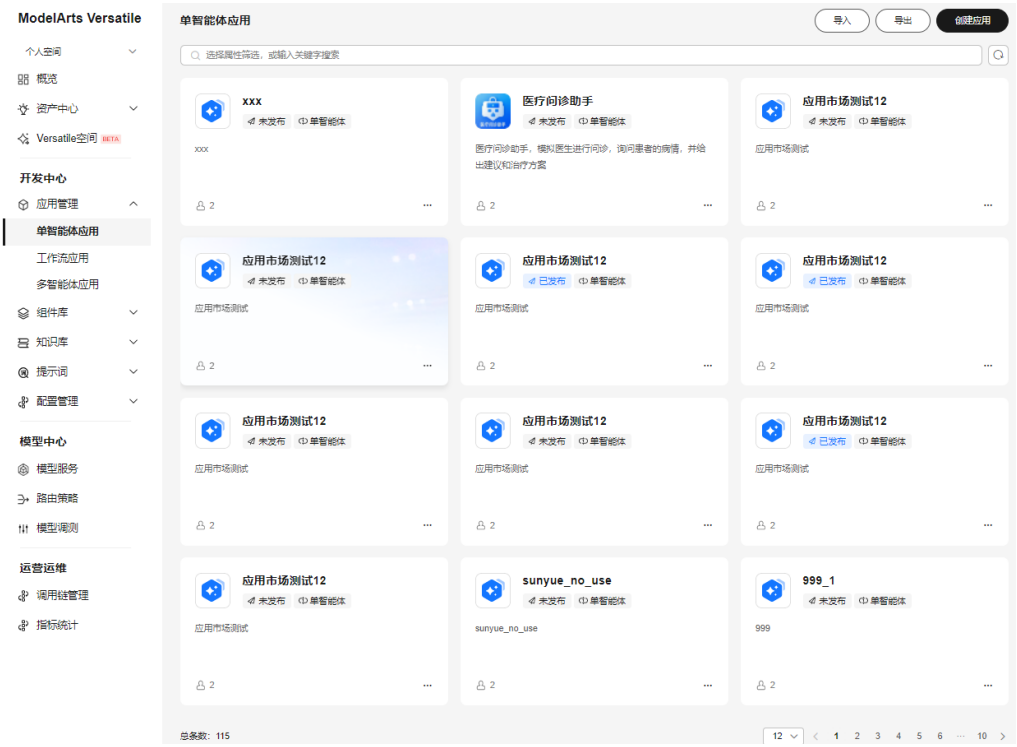
步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择进入所需空间。

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 6-72 选择团队空间

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 6-73 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 6-74 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“版本号”、“描述信息”，单击“发布”按钮完成发布。

图 6-75 编辑 Agent 应用

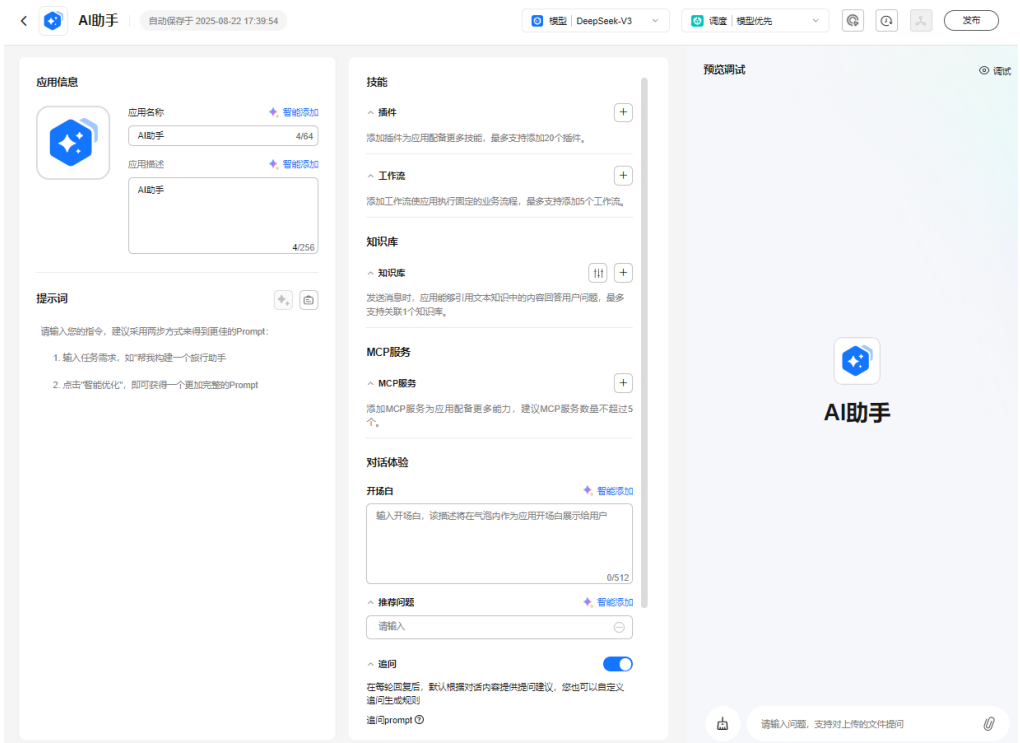


图 6-76 填写发布信息

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 6-77 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**。

6.12 使用 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，选择进入所需空间。

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 6-78 选择团队空间

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 6-79 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6-80所示。

其中，conversation_id参数为会话ID，唯一标识每个会话的标识符，可将会话ID设置为任意值，使用标准UUID格式。

图 6-80 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《 API参考 》“应用示例 > 调用智能体应用示例”章节。

6.13 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及查看调用路径等。

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，选择进入所需空间。

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 6-81 选择团队空间

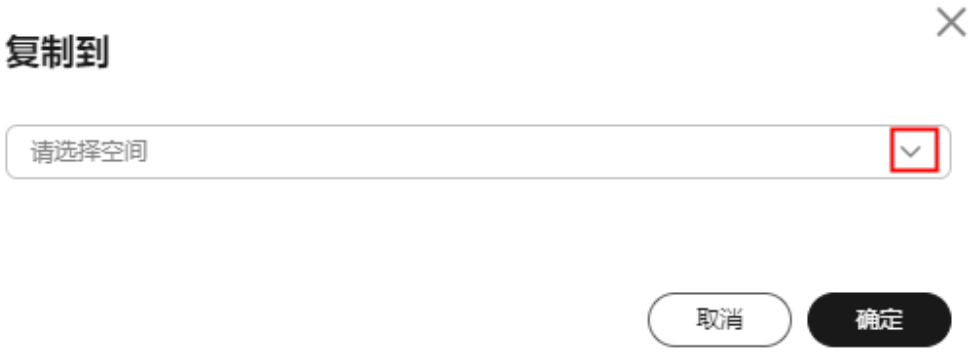
步骤2 在左侧菜单栏，单击“开发中心 > 应用管理 > 单智能体应用”。

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”。

图 6-82 复制 Agent 应用

步骤4 在“复制到”下拉框中选择已创建的目标空间。

图 6-83 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，选择进入所需空间。

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 6-84 选择团队空间

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”。

图 6-85 删除 Agent 应用

步骤4 弹出的对话框中单击“确定”。

图 6-86 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，选择进入所需空间。

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 6-87 选择团队空间

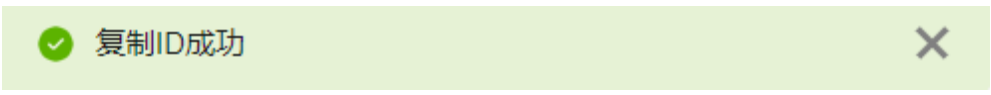
步骤2 在左侧菜单栏，单击“开发中心 > 应用管理 > 单智能体应用”。

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

图 6-88 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 6-89 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 6-90 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，选择进入所需空间。

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 6-91 选择团队空间

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。如果解析的文件在已存在，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，选择进入所需空间。

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 6-92 选择团队空间

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

7 开发工作流应用

7.1 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，其中多个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。

平台提供一个可视化画布，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7-1**。

表 7-1 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**。
	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**。

分类	节点名称	节点说明
	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**。
	意识识别节点	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**。
数据&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**。
	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	用于支持对Database放开读写控制，用户可读写其他用户提交的数据，详细配置参见**数据库**。

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，根据功能需要使用以下节点进行设计：

图 7-1 配置节点

7.2 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 7-2 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。
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《常见问题》“对话型工作流、任务型工作流是否可以相互转换”章节。

7.3 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 7-3 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。
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。

7.4 搭建工作流

7.4.1 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、Versatile空间、模型中心、开发中心的组件库、知识库、提示词、配置管理等，都称之为“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。

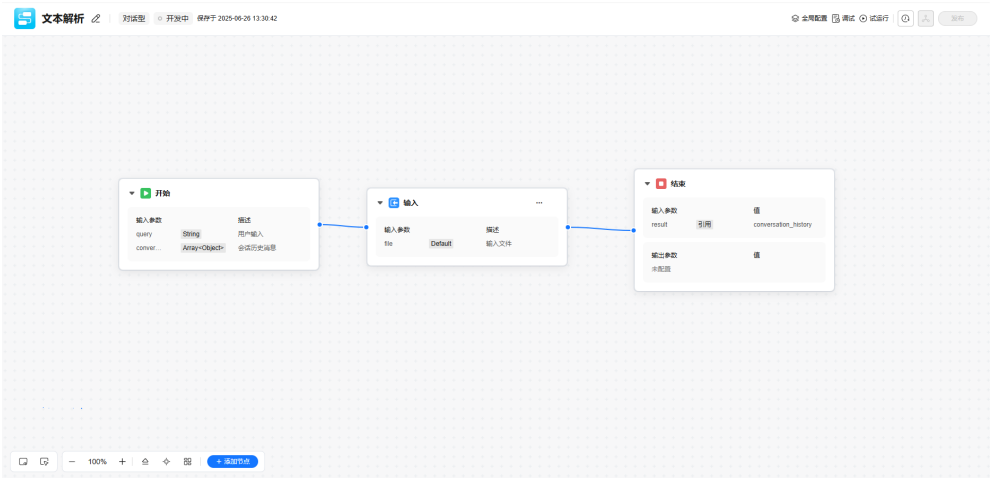
Versatile中资源之间可以通过流程搭建，业务链接，进行相互调用，在进行工作流编排的时候，可以通过拖拽把资源添加到工作面板中。

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 7-2 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 7-4 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：例如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：例如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：例如由于资源限制，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：例如由于资源充足，可以同时处理多个任务，提高整体处理速度。 ● 并行计算：例如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

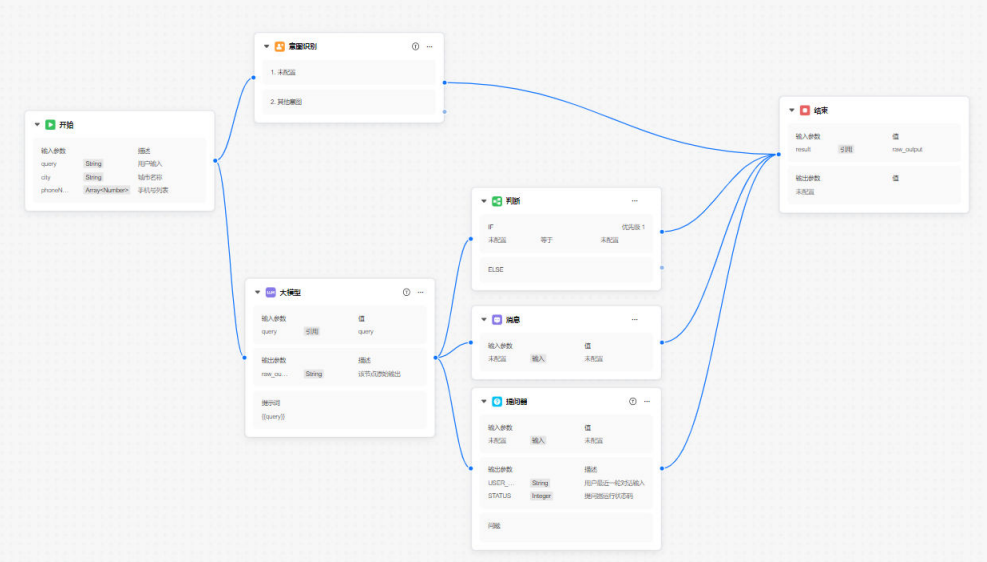
并行编排支持多种结构。

- 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，并行结果中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 7-3 串行编排

图 7-4 常规并行编排

图 7-5 嵌套并行编排



7.4.2 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6-9所示。

表 7-5 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。	参考本章节。
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。	使用预置的工作流

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，选择进入所需空间。

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 7-6 选择团队空间

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7-6所示。

表 7-6 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。
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。

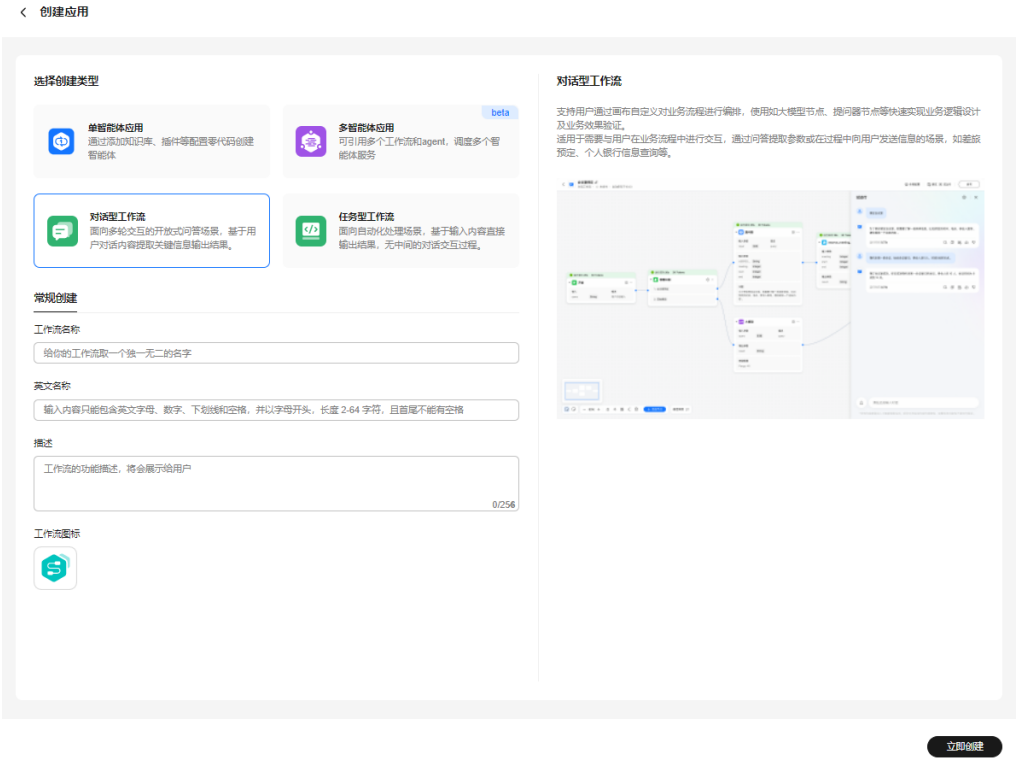
步骤4 选择完成后配置应用基础信息，参数说明如表7-7所示。

表 7-7 基础信息参数说明

参数	说明	示例
工作流名称	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
英文名称	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
想要的Agent	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。

参数	说明	示例
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 7-7 创建工作流

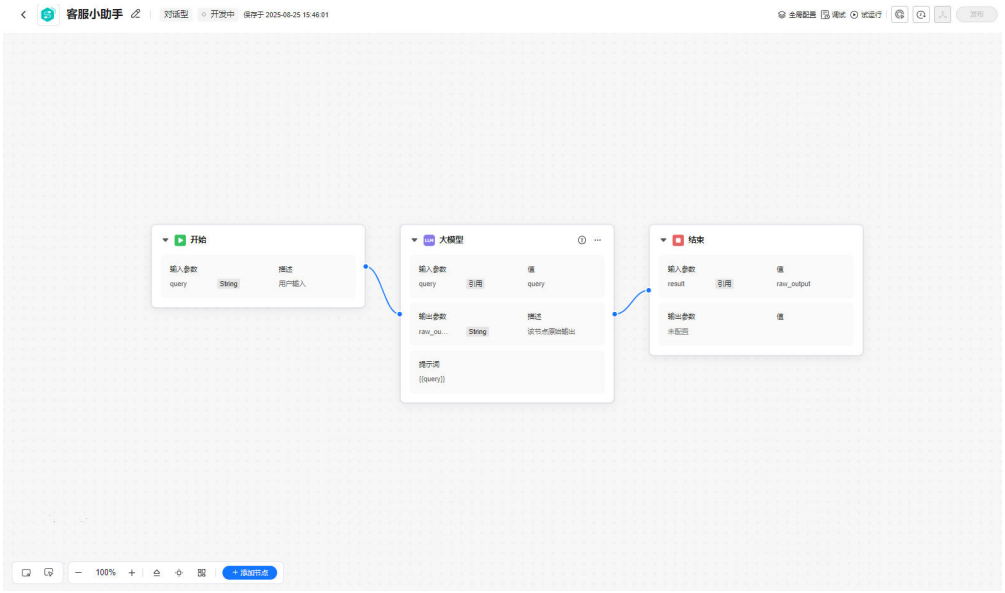


步骤5 配置完成后单击“立即创建”，进入工作流编排页面。

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 7-8 编排画布界面



----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 7-8 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。 ● “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。
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。

参数	功能
记忆变量	记忆变量的节点赋值支持工作流节点的引用，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。 **说明** ● String：字符串，用于存储文本数据，例如单词、句子或字符序列。 ● Number：数值。 ● Object：对象类型，可传json对象。 ● Inter：数字类型。 ● Array：数组类型。 ● 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）参数描述信息，帮助理解传入参数的含义。 ● 默认值：（可选）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。
节点赋能	通过添加由变量赋值节点赋值的记忆变量，让工作流记住用户的关键信息，生成更符合用户特征的回答。支持选择类型和记忆时长。
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。 **说明** ● 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，打开当前开关后，“安全防护”将自动关闭。
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。 **说明** 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。

参数	功能
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 7-9 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7-11**所示参数、结束节点必须配置如**图7-12**所示参数。

图 7-10 编排工作流

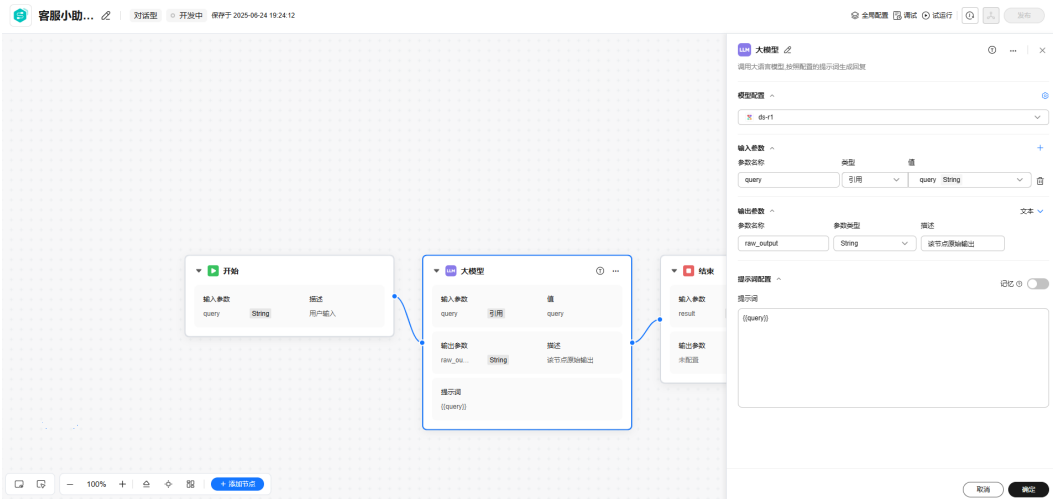


图 7-11 开始节点配置（工作流用于多智能体应用意图识别时配置，参数与截图保持一致）

图 7-12 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7-13所示。

图 7-13 画布界面操作

表 7-9 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。

操作	说明
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**。
试运行	单击画板右上角 ，进入“试运行”界面，可在输入框中输入问题测试应用功能。
触发器配置	单击画板右上角 ，进入触发器配置页面，通过定时任务，调用应用按照指令要求执行，详细操作可参考**配置触发器**。
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。
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**。

7.4.3 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试运行，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**。

试运行工作流（必选）

步骤1 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。

说明

试运行界面支持文本输入、语音输入和文件输入：

- 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：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。
- 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。

图 7-14 试运行工作流

说明

- 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。

步骤2 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7-15所示。

如果试运行失败，常见报错与解决方案请详见**工作流常见问题**。

图 7-15 调试结果示例

说明

- 单击调用详情页面中的按钮可查看列表调用链。
- 单击调用详情页面中的按钮可查看火焰图调用链。
- 支持在**调用链管理**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 7-16 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。

图 7-17 单节点调试运行成功示例

----结束

7.4.4 配置触发器

触发器是工作流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，从而提高工作流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执行任务。

前提条件

工作流已调试完成，具体请参见**调试工作流**。

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，选择进入所需空间。

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 7-18 选择团队空间

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 单击目标工作流，在工作流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 7-10 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。 • 按每日执行 • 按每周执行 • 按每月执行 • 自定义间隔时间：支持“间隔两天”、“间隔三天”、“间隔四天”、“间隔五天”、“间隔六天”。

参数	说明
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。
调用方式	• 同步：触发器触发后，会等待被调用的操作完全执行完毕并返回结果，才继续执行后续流程。 • 异步：触发器触发后，仅发起调用请求，不等待操作完成，直接继续执行后续流程。

图 7-19 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

7.4.5 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。

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”，输入版本号与描述，单击“发布”。

图 7-20 发布工作流

步骤2 发布完成后跳转至API调用页面，可看到发布的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 > 工作流应用”，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 7-21 调用 API

----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 7-22 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。

7.5 使用工作流

7.5.1 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，选择进入所需空间。

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 7-23 选择团队空间

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 7-24 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6-80所示。

其中，conversation_id参数为会话ID，唯一标识是每个会话的标识符，可将会话ID设置为任意值，使用标准UUID格式。

图 7-25 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《API参考》“应用示例 > 调用工作流应用”章节。

7.5.2 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

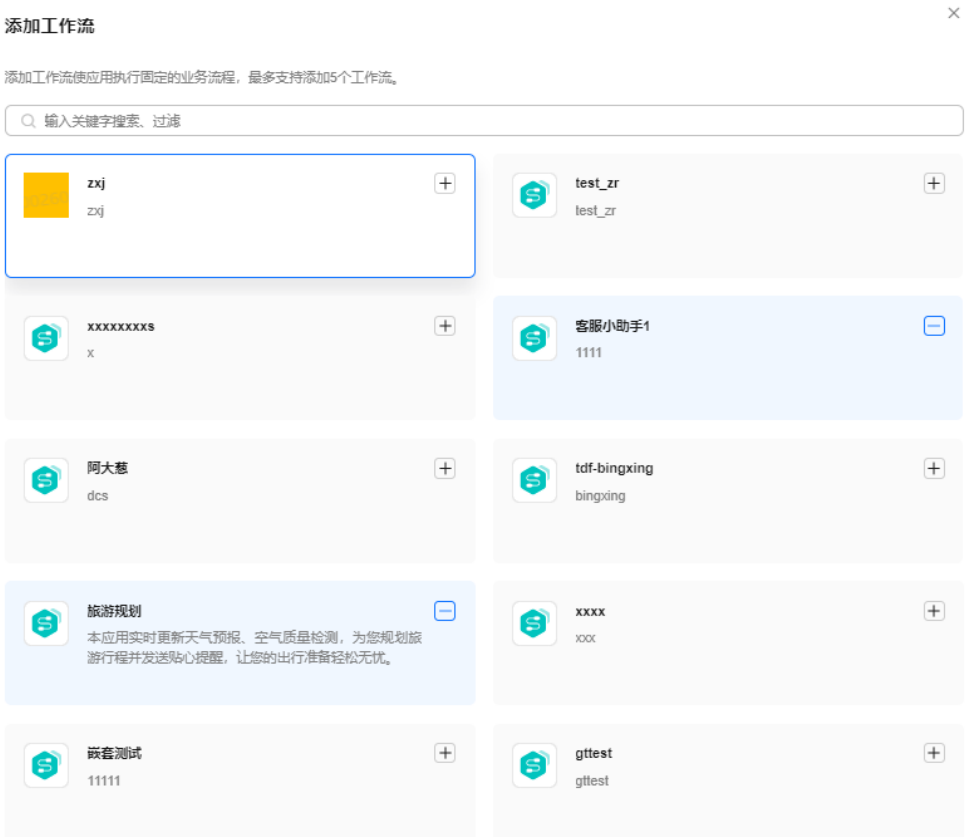
须确保工作流已发布，详情可参考**发布工作流**。

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 + 。

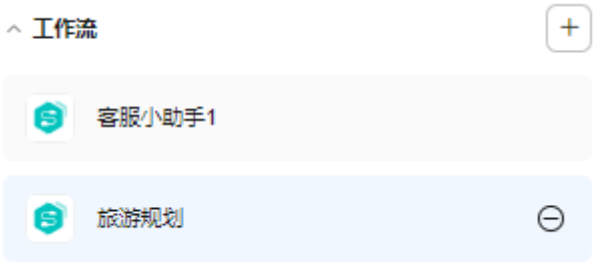
步骤2 在“添加工作流”窗口，单击 + 进行添加，再单击“确定”。

图 7-26 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 7-27 已添加工作流



----结束

相关文档

单智能体应用中使用工作流示例，请参考单智能体应用中使用工作流。

7.5.3 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 > 子工作流”模块，单击“  ”添加子工作流。

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 7-28 添加工作流

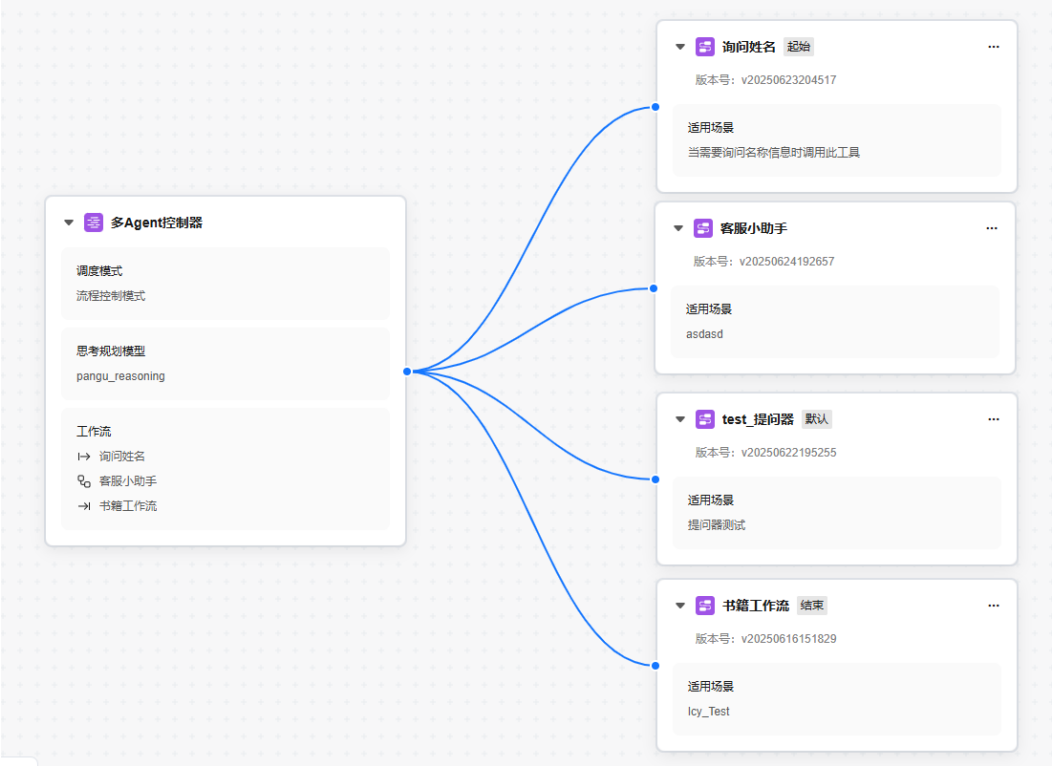
步骤3 添加工作流保存后，可在画布中查看当前已添加的工作流。

图 7-29 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 7-30 添加起始、默认、结束工作流

图 7-31 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，请参考多智能体应用中使用工作流。

7.6 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、设为业务节点、删除、导入、导出等操作。

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，选择进入所需空间。

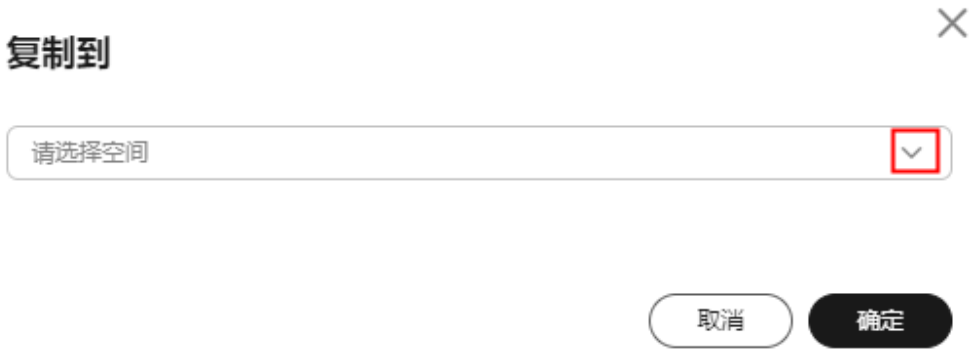
如果已选择团队空间，界面显示为实际的团队空间名称，而非“个人空间”。

图 7-32 选择团队空间

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 7-33 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，选择进入所需空间。

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 7-34 选择团队空间

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 7-35 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，选择进入所需空间。

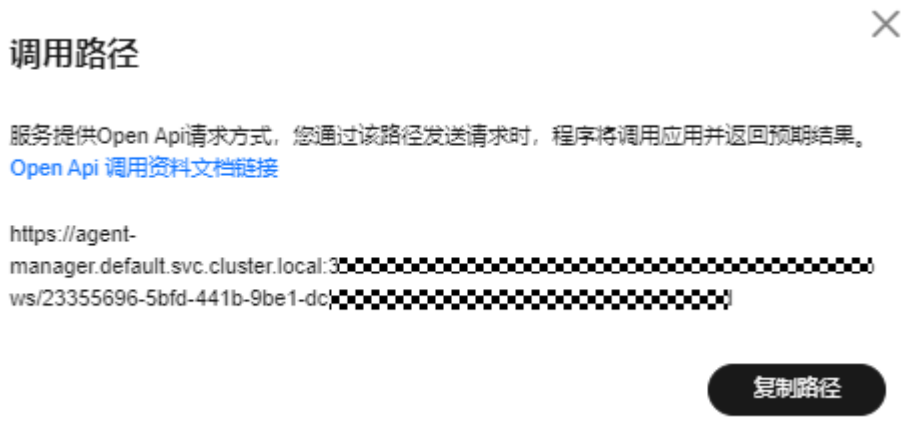
如果已选择团队空间，界面显示为实际的团队空间名称，而非“个人空间”。

图 7-36 选择团队空间

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 7-37 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。

步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择进入所需空间。

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 7-38 选择团队空间

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 7-39 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 7-40 工作流被引用

----结束

设为业务节点

在Versatile中，支持将调试成功的工作流设置为节点。将工作流应用设为业务节点后，该工作流应用将常驻在工作流节点列表中，并且可以被新建的工作流引用。如果被引用的工作流节点被删除或修改，将会影响所有引用该节点的工作流。工作流节点列表中最多支持设置30个工作流节点。

步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择进入所需空间。

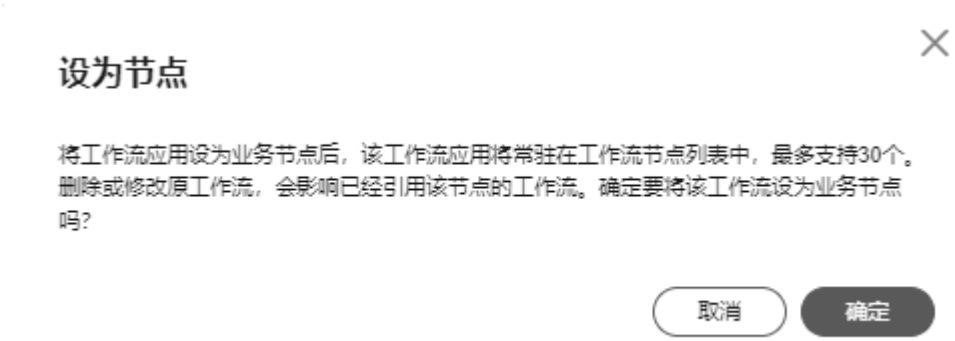
如果已选择团队空间，界面显示为实际的团队空间名称，而非“个人空间”。

图 7-41 选择团队空间

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“ ··· > 设置业务节点”，在“设为节点”弹窗中单击“确定”。

图 7-42 设为节点



步骤4 设置后，可以在创建或编辑工作流画板时，通过“添加节点 > 业务”选项中查看工作流节点。

图 7-43 设置节点

----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，选择进入所需空间。

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 7-44 选择团队空间

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，选择进入所需空间。

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 7-45 选择团队空间

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

7.7 基础节点

7.7.1 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。

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3层嵌套。
- **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。将工作流绑定到智能体中使用时，如果用户Query中缺少必选参数，则不会触发工作流调用。
- **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 7-46 开始节点 Object 类型参数嵌套

图 7-47 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。结束节点支持配置两种参数：输入参数和输出参数，分别对应两种返回方式，文本返回和变量返回。

图 7-48 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。指定回复中不能插入输出参数。

7.8 通用节点

7.8.1 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**。

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，选择进入所需空间。

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 7-49 选择团队空间

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7-11，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 7-11 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	模型选择	选择模型接入模块已配置的大语言模型。	DeepSeek-V3

配置类型	参数名称	参数说明	配置示例
	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
	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。	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。	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
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  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”页签中进行选择。	-
	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}}
	记忆	是否打开记忆功能；打开后可记录多轮对话的内容，默认关闭。	关闭

图 7-50 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

7.8.2 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**。

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，选择进入所需空间。

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 7-51 选择团队空间

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 通过单击该节点打开节点配置页面。

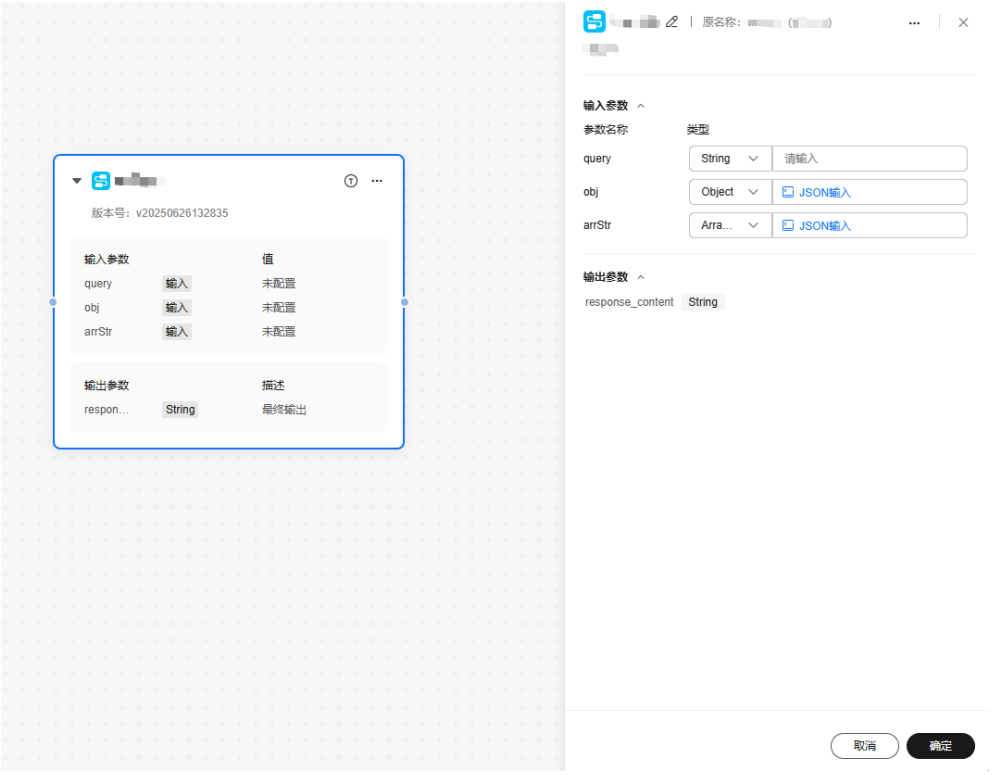
步骤5 参照表1 工作流节点配置说明，完成工作流节点的配置。

表 7-12 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 • 在工作流节点中你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

图 7-52 工作流节点配置示例-子工作流

图 7-53 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6 节点配置完成后，单击“确定”。

----结束

7.8.3 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Prompt、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，选择进入所需空间。

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 7-54 选择团队空间

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7-13，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，可对大模型节点进行测试。

表 7-13 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。	DeepSeek-V3

配置类型	参数名称	参数说明	配置示例
	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
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。	-

配置类型	参数名称	参数说明	配置示例
	系统提示词	配置输入给大模型的系统提示词，用于指导模型更好地完成任务。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，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”页签中进行选择。	-
	输出参数	输出参数为Agent节点最后一轮的输出。	-
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
	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

图 7-55 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

7.9 逻辑节点

7.9.1 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，选择进入所需空间。

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 7-56 选择团队空间

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7-57**和**表7-14**，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 7-14 判断节点配置说明

配置类型	参数名称	参数说明	配置示例
参数配置	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 7-57 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

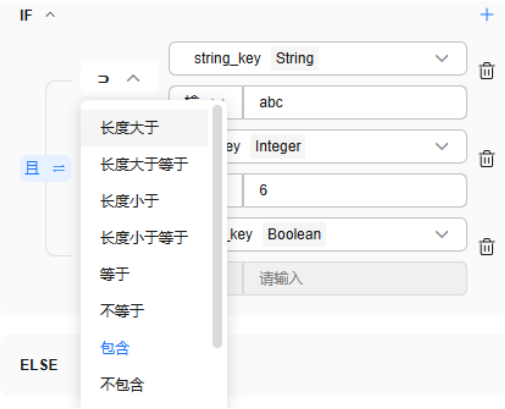
示例

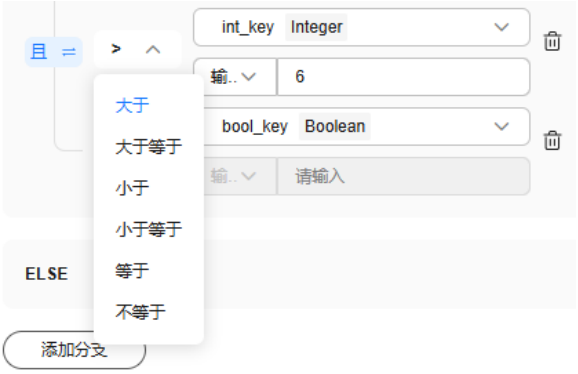
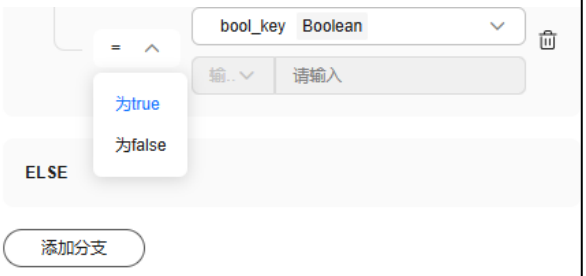
以常见的String、Integer和Boolean类型条件判断为例，在开始节点中定义三种类型的参数，模拟判断节点的输入参数，实现对于不同类型参数在不同条件下的逻辑判断。

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 7-15 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 7-58 试运行

7.9.2 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**。

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，选择进入所需空间。

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 7-59 选择团队空间

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。根据“执行方式”不同，支持两种配置方式。

- 安全沙箱：用于隔离程序运行的环境，以防止潜在的恶意代码对系 统造成损害。安全沙箱配置方式请参照**安全沙箱执行方式配置说明**。
- 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**FunctionGraph执行方式配置说明**。

  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。

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

安全沙箱执行方式配置说明

图 7-60 安全沙箱执行方式配置示例

表 7-16 安全沙箱执行方式配置说明

参数	说明
执行方式	安全沙箱

参数	说明
输入参数	配置代码运行需要的输入参数。 当单击+图标时，可新增输入参数。 ● 参数名称：仅支持输入字母、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输出参数和全局配置的记忆变量。 – 输入：支持用户自定义取值。
输出参数	配置代码运行后需要输出的参数，需要与return返回的对象保持一致。 当单击+图标时，可新增输出参数。 ● 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：输出参数的类型，可选String、Integer、Number、Boolean、Object、Array平台支持的类型。 ● 描述：对于该输出参数的描述。 ● 必填：选择当前输出参数是否必填。
代码	编写代码时不要更改第一行函数定义。 编写Python代码片段，仅支持python系统库，不支持引入依赖包；代码配置示例请参考**示例**。

FunctionGraph 执行方式配置说明

⚠ 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**。

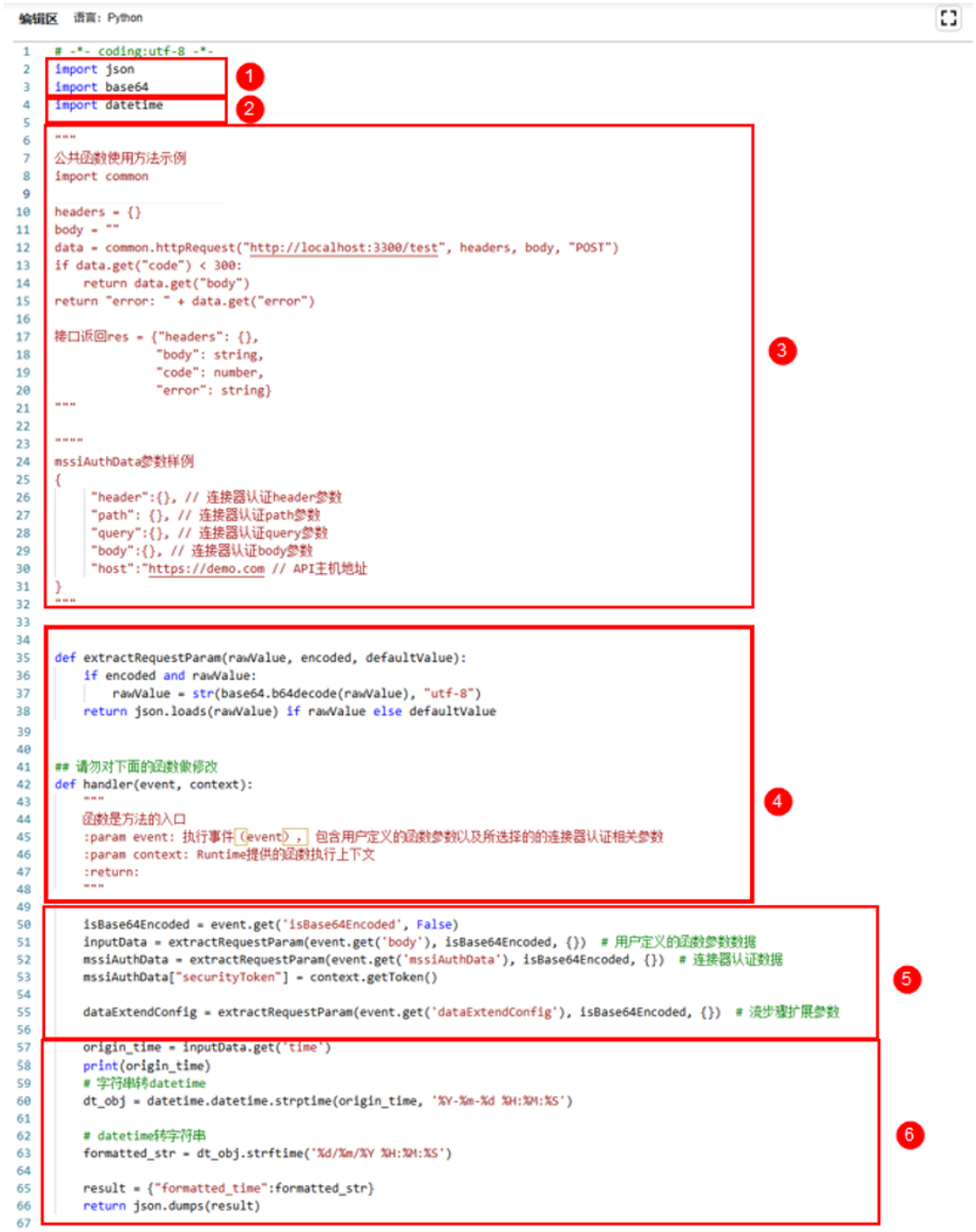
表 7-17 FunctionGraph 执行方式配置参数说明

参数	说明
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7-18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。
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。

表 7-18 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。
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7-61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，请参考示例的输出格式。
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 7-61 源码编辑区




```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)
```

示例

开发语言

代码节点以Python语言为例。

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

配置示例：

- 文本拼接示例代码。

```
def main(args: dict) -> dict:
# 注意在输入参数中定义名为input1的变量
input1 = args.get('input1')
# 注意在输入参数中定义名为input2的变量
input2 = args.get('input2')
res = {
# 注意在输出参数中定义名为res的变量
"res": input1 + input2,
}
return res
```

- 数学计算示例代码。

```
def main(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"
}
```

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，选择进入所需空间。

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”。

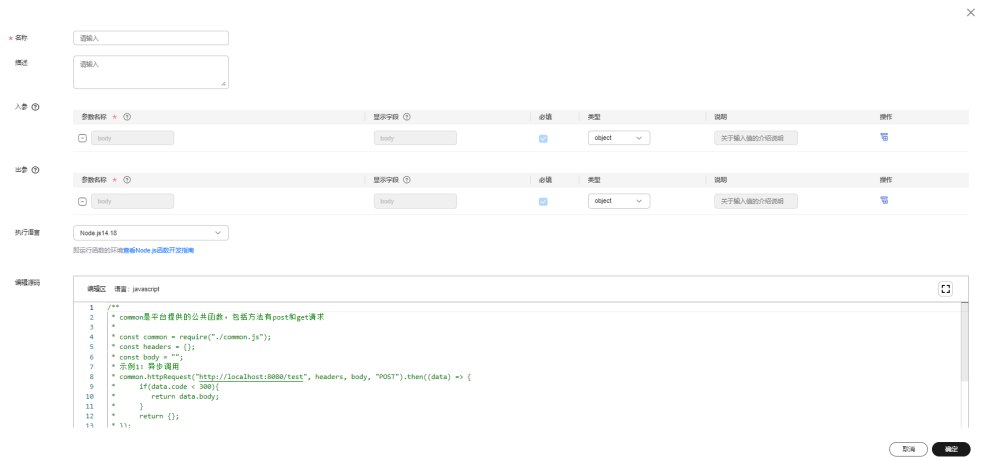
图 7-62 选择团队空间

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”。

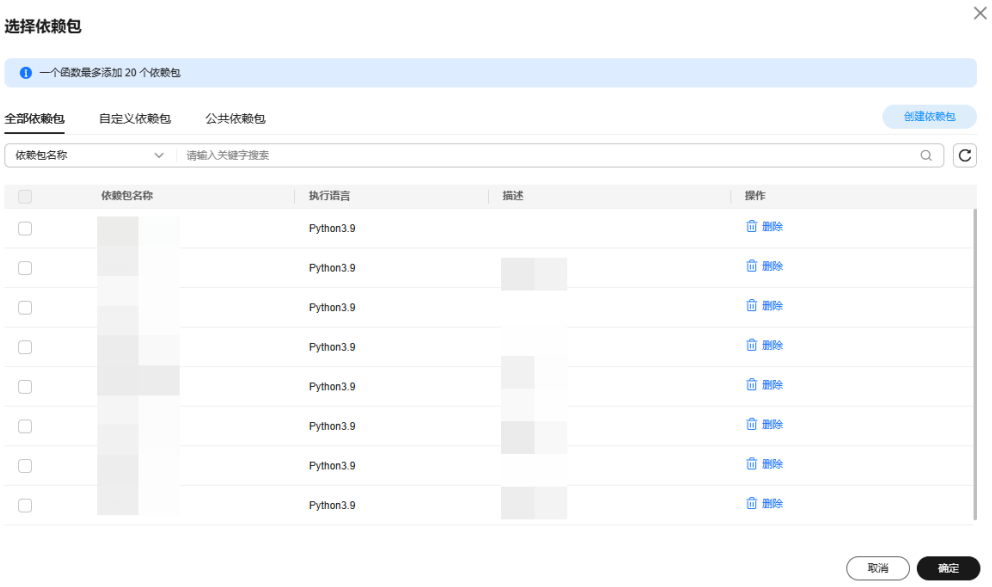
步骤4 选择已有函数，进入编辑函数界面，如图7-63所示。

图 7-63 创建函数



步骤5 单击“依赖包”后的“添加”，如图7-64所示。

图 7-64 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7-19所示。

表 7-19 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。

参数	说明
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

7.9.3 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**。

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，选择进入所需空间。

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 7-65 选择团队空间

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 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）。

步骤5 参照表7-20，完成循环节点的配置。

说明

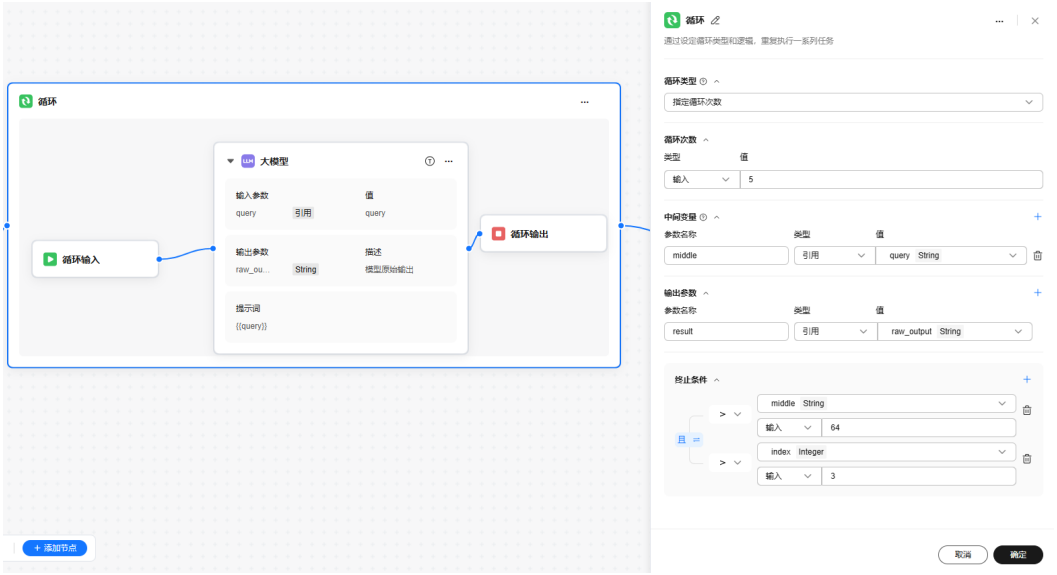
- 单击图标，可修改循环节点名称，修改完成后单击名称旁边的进行保存。
- 单击图标，可重命名循环节点名称，复制一个循环节点或删除循环节点。

表 7-20 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。
	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	循环数组	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。

配置类型	参数名称	参数说明
	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。
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 7-66 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

7.9.4 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，选择进入所需空间。

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 7-67 选择团队空间

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7-68和表7-21，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 7-21 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	模型选择	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
	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	• 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。	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 7-68 意图识别节点配置示例 1

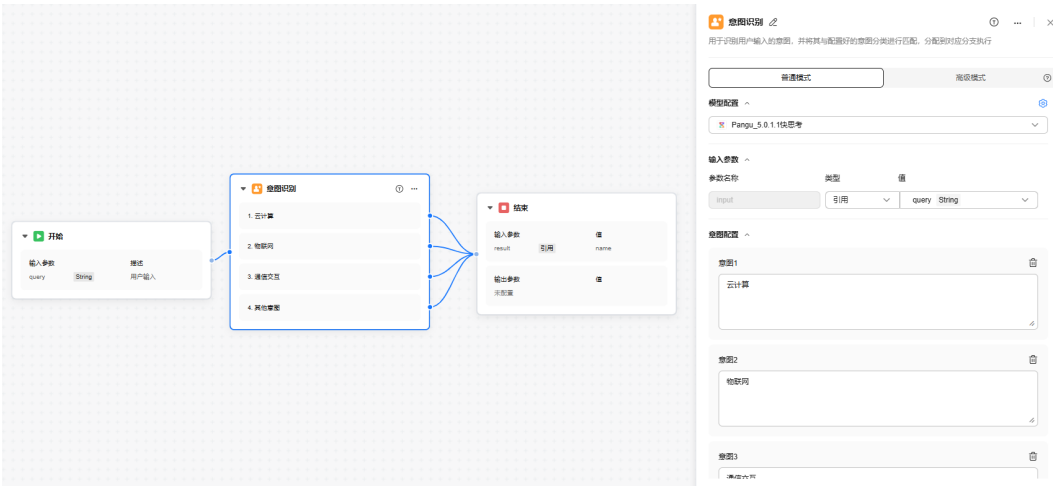


图 7-69 意图识别节点配置示例 2

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 7-70 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。

7.9.5 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 “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 7-71 配置意图包

步骤2 拖动左侧“意图识别”节点至画布中，单击该节点以打开节点配置页面。切换为 “高级模式”。

步骤3 参照意图模式配置说明，完成配置。

表 7-22 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	模型选择	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	选择前面已经配置的意图包。
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。

配置类型	参数名称	参数说明
	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。
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 7-72 意图识别节点配置示例

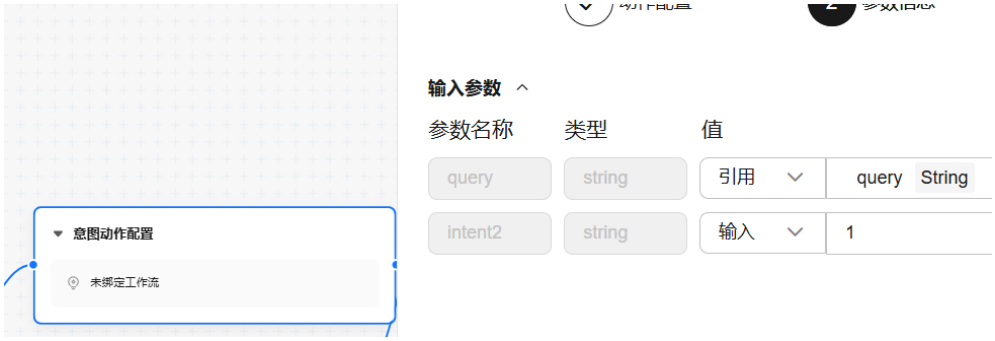
步骤4 节点配置完成后，单击“确定”。

步骤5 单击意图动作节点，配置分支对应的处理逻辑。

图 7-73 配置处理逻辑

步骤6 配置子工作流的输入参数。

图 7-74 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

7.10 工具节点

7.10.1 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**。

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。

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，选择进入所需空间。

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 7-75 选择团队空间

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 通过单击该节点打开节点配置页面。

步骤5 在“个人插件”或“预置插件”页签单击“+”，将插件添加至画布中。

说明

- 预置插件为平台内置的插件。
- 个人插件为用户自定义的插件，创建插件步骤详见**创建插件**。

步骤6 连接插件节点和其他节点。

步骤7 单击画布中已添加的“插件”节点，参照**表7-23**，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。
- 单击图标，可对插件节点进行测试。

表 7-23 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。	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。	-
异常配置	-	“异常配置”开关为“异常忽略”。作为插件节点的兜底配置，当插件执行异常时，支持配置兜底返回，避免整个工作流运行失败： • 关闭时，该功能不起作用。 • 开启时，填写“默认输出”，默认输出的参数需要与插件“输出参数”一致。当工作流插件节点运行正常时，该配置对后续节点没有影响。当工作流运行到插件节点出现异常时，工作流不会中止，继续运行后续节点。如果后续节点引用了插件节点的输出内容，则使用“默认输出”的内容。	用户期望插件执行异常不会导致整个工作流中断时，开启“异常配置”，参考插件的输出参数配置“默认输出”。

图 7-76 插件节点配置示例

步骤8 节点配置完成后，单击“确定”。

----结束

7.10.2 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服务类型包括预置服务和个人服务。

- 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，开通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。
- 个人服务：平台允许开发者创建自定义MCP服务，支持将MCP服务地址通过配置方式快速创建为自定义MCP服务，提供给工作流或应用调用。

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，选择进入所需空间。

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 7-77 选择团队空间

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 在“个人服务”或“预置服务”页签单击“+”，将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。

说明

- 预置服务为平台内置的MCP服务。
- 个人服务为用户创建的自定义MCP服务，创建MCP服务步骤详见**创建MCP服务**。

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

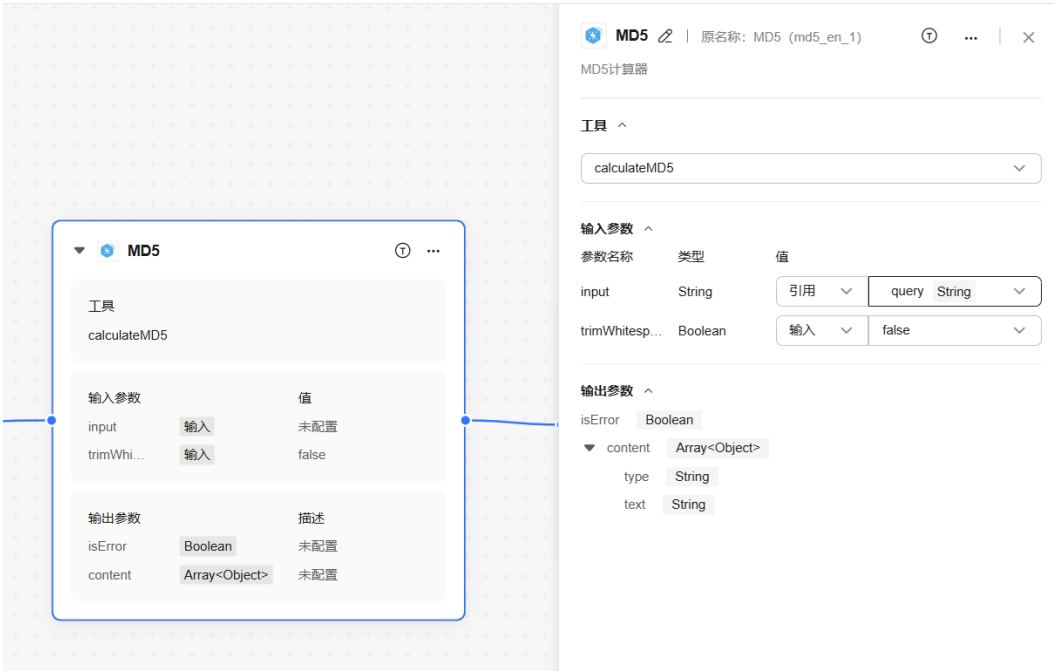
- 单击图标，可修改MCP服务节点名称，修改完成后单击名称旁边的进行保存。
- 单击图标，可重命名MCP服务节点名称，复制一个MCP服务节点或删除MCP服务节点。
- 单击图标，可对插件节点进行测试。

表 7-24 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。	-

配置类型	参数名称	参数说明	配置示例
	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 7-78 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

7.11 消息管理节点

7.11.1 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。

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，选择进入所需空间。

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 7-79 选择团队空间

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7-25**，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 7-25 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。
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。

图 7-80 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

7.11.2 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，选择进入所需空间。

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 7-81 选择团队空间

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7-26，完成变量输入节点的配置。

表 7-26 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 7-82 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

7.11.3 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，选择进入所需空间。

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 7-83 选择团队空间

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7-27，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 7-27 提问器节点配置说明

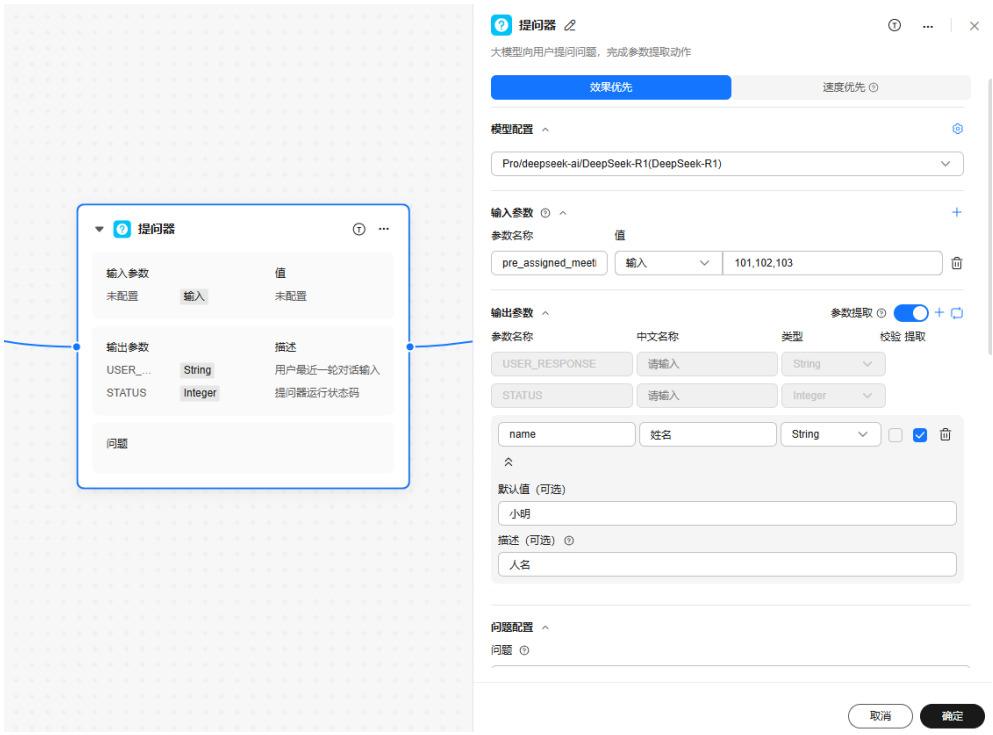
配置类型	参数名称	参数说明
模式偏好	-	效果优先：效果优先模式下，会开启时间增强和反思功能，提参成功率更高，时延会增加。 • 时间增强：需要提取时间时，可以将自然语言的时间日期提取为YYYY-MM-DD HH:MM:SS标准格式的时间日期，比如：明天 12:30，提取为 2024-04-13 13:30:00。 • 反思功能：数据提取之后，会让模型再判断是否提取正确，格式是否满足要求，不满足会尽量做一些修正。比如：期望提取电话号码，用户输入：我不记得电话号码，提取出：189******，反思后会认为提取不正确，会继续追问 速度优先：速度优先模式下时延最低，提参成功率可能无法保障速度优先模式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jinjia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，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，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 7-84 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

7.12 数据&知识节点

7.12.1 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，选择进入所需空间。

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 7-85 选择团队空间

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7-28，完成变量赋值节点的配置。

表 7-28 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 7-86 变量赋值节点配置示例

图 7-87 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

7.12.2 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，选择进入所需空间。

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 7-88 选择团队空间

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7-29，完成变量聚合节点的配置。

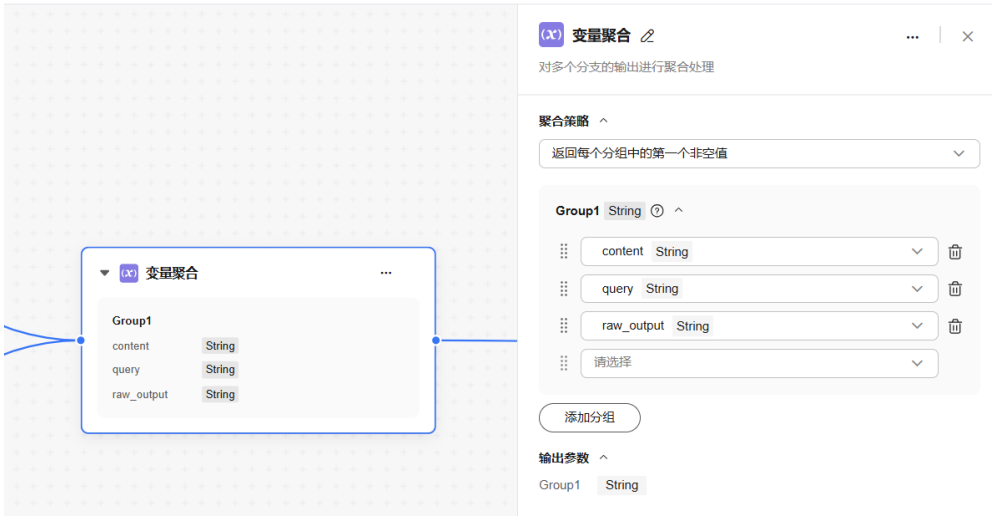
说明

- 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 7-29 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 7-89 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

7.12.3 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，选择进入所需空间。

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 7-90 选择团队空间

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7-30，完成大模型节点的配置。

说明

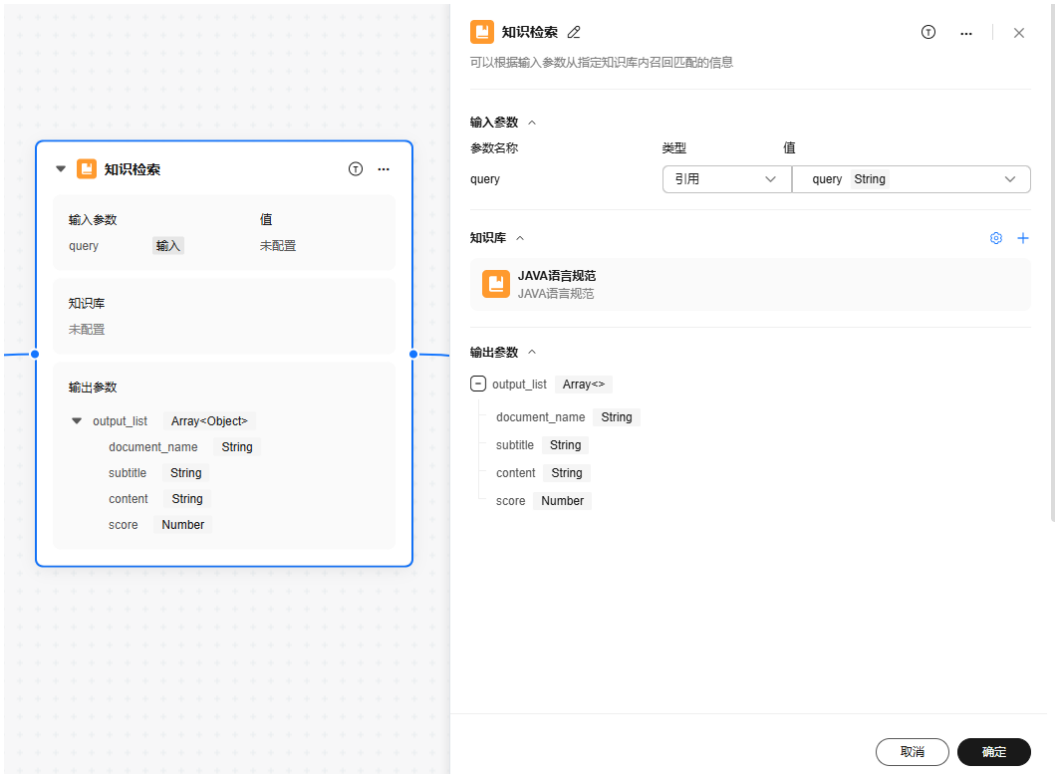
- 单击图标，可修改知识检索节点名称，修改完成后单击名称旁边的进行保存。
- 单击图标，可重命名知识检索节点名称，复制一个知识检索节点或删除知识检索节点。

表 7-30 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。
	知识库	支持选择用户所创建的知识库。

配置类型	参数名称	参数说明
	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。 ● **FAQ检索**：使用向量检索技术检索FAQ，推荐在需要预先定义好问答对的场景中使用。
	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。
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 7-91 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

7.12.4 数据库

数据库节点可提供工作流执行sql语句的能力，执行用户输入或模型生成的sql语句，完成数据的增删改查。配置数据库节点前请先接入数据源。

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**。
- 配置数据库节点前，请确保“开发中心 > 配置管理 > 数据源管理”中已接入数据源，具体请参见**接入数据源**。

配置数据库

步骤1 **登录Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择进入所需空间。

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 7-92 选择团队空间

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“数据库”节点。

步骤4 在工作流画布中单击“数据库”节点，打开配置页面。

图 7-93 数据库配置界面

步骤5 配置界面内容可参见**表1 数据库节点配置说明**。

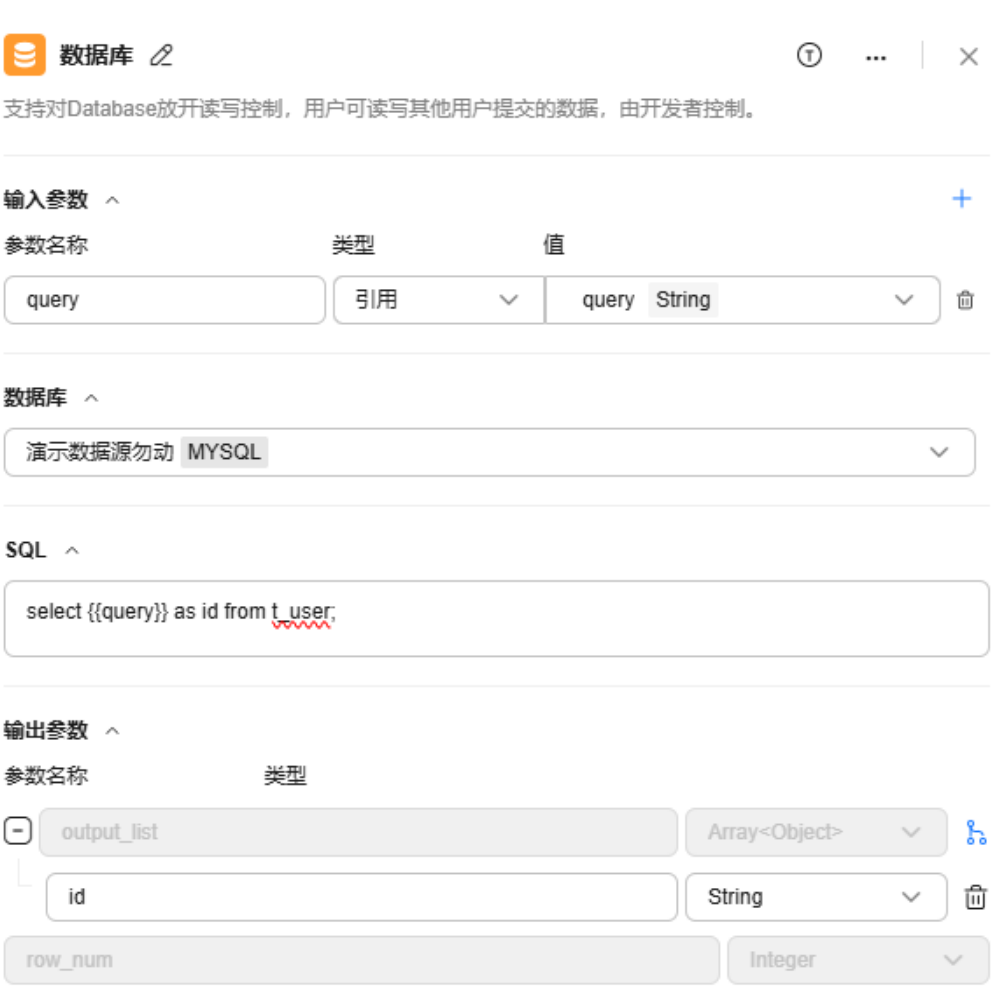
说明

- 单击图标，可修改数据库名称，修改完成后单击名称旁边的进行保存。
- 单击图标，可重命名数据库节点名称，复制一个数据库节点或删除数据库节点。
- 单击图标，可对数据库节点进行测试。

表 7-31 数据库节点配置说明

配置类型	参数名称	参数说明
数据库配置	数据库	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，具体请参见**接入数据源**。
	SQL	需要执行的sql语句，支持增删改查相关语句，可通过“{{var}}”形式引入输入参数的变量，执行数据库语句不包含限制，请谨慎配置。
参数配置	输入参数	配置SQL运行需要的输入参数。 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输出参数和全局配置的记忆变量。 – 输入：支持用户自定义取值。
	输出参数	配置SQL执行需要输出的参数，需要与SQL语句SELECT返回结构一致。 当单击图标时，可新增输出参数子项。 当单击图标时，可删除输出参数子项。 • 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：输出参数的类型，可选String、Integer、Number、Boolean等平台支持的类型。

图 7-94 数据库节点配置示例



步骤6 节点配置完成后，单击“确定”。

步骤7 连接数据库节点和其他节点。

----结束

7.13 配置管理

7.13.1 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。

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”。

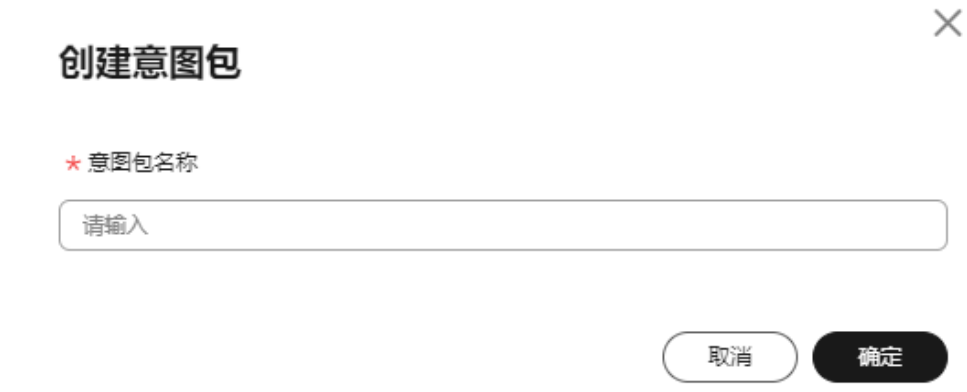
步骤2 单击“创建意图包”。

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、下划线和空格输入。

图 7-95 创建意图包



步骤4 单击“编辑”，进入意图包编辑页面。

图 7-96 配置数据源

步骤5 在意图包中添加意图分类，分类信息包含名称和样例。

图 7-97 配置数据源

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 批量导出意图。
  - 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，选择进入所需空间。

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 7-98 选择团队空间

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. 单击页面右上角“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”。单击页面右上角“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. 单击页面右上角“导入”。
2. 在“导入”页面，单击“选择文件”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

7.13.2 接入数据源

接入数据源可为数据库提供数据来源，提供统一的接口和管理方式，方便用户配置、连接、读取和处理数据。

说明

工作流配置“数据库节点”时，须优先配置接入输入源。

配置接入数据源

步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择进入所需空间。

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 7-99 选择团队空间

步骤2 打开“开发中心 > 配置管理 > 数据源管理”。

步骤3 在数据源管理界面，单击“接入数据源”，进入数据源配置页面。

图 7-100 数据源配置

步骤4 参考表7-32，配置数据源信息。

表 7-32 配置数据源参数说明

参数名称	参数说明
数据源名称	数据库在Versatile存储的名称。
源库类型	接入数据源的类型。（当前仅支持MySQL）。

参数名称	参数说明
接入网络类型	接入网络类型，当前仅支持公网接入。
域名	数据库的IP地址，例如：xxx.xxx.xxx.xxx。
端口	数据库的登录端口。例如：3306。
数据库名称	登录的目标数据库的名称。
SASL_SSL	数据库SSL开关，可根据数据库是否直接进行配置。
用户名	登录目标数据库的用户名。
密码	登录目标数据库的密码。
描述	（可选）数据库在Agent平台存储的描述。

图 7-101 数据源配置

步骤5 单击“连接并保存”，保存配置。

----结束

7.14 工作流常见问题

工作流常见报错及解决方案请详见表7-33。

表 7-33 工作流节点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。
代码节点	101591	代码组件初始化失败。	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。
	101594	代码组件安全沙箱未知异常。	请联系客服解决。
	101595	代码节点执行失败未知异常。	请联系客服解决。
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。

模块名称	错误码	错误描述	解决方案
	101654	消息组件执行失败。	请联系客服解决。
	101655	消息组件异步执行失败。	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	101050	执行默认护栏（时间参数解析）失败时触发该错误码。	可检查支持处理的时间类型是否超出支持范围。
	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。
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。
	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。
	105008	插件内部错误。	请确认接入的插件服务是否正常。
	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。
认证鉴权	110000	认证失败。	请检查是否正确传入认证信息。
	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。

模块名称	错误码	错误描述	解决方案
	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。
	101059	变量引用的字符串有问题。	查看变量引用值的配置是否有问题。
对话接口	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。

8 开发多智能体应用

8.1 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

Versatile《常见问题》：

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?

8.2 创建多智能体应用

本节简介配置多智能体应用的流程。

前提条件

已发布工作流。

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，选择进入所需空间。

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 8-1 选择团队空间

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”。

步骤3 在“创建应用”页面，配置应用信息，具体参数说明请参考表8-1。

表 8-1 创建多智能体参数说明

参数	说明	示例
应用名称	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
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，都可以给工作流的输入参数使用。

- 输入参数：支持配置String、Integer、Number、Boolean类型，传给工作流的输入参数，且值不可修改。
- 全局参数：支持配置String、Integer、Number、Boolean类型，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。

图 8-2 全局配置

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8-2**。

表 8-2 多 Agent 控制器参数说明

参数	说明
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。 • “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。 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。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循指令要求但减少多样性。建议不要与核采样同时调整。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。 • 根据示例，在输入框中输入提示词。 • 在“提示词”右侧，单击 ，在“提示词广场”页面，选择预置的提示词或是用户自定义创建的提示词，单击“确定”。 在“提示词”右侧，单击 ，对输入的提示词进行智能优化。

参数	说明
意图识别（可选）	该多智能体应用的意图识别能力。 在下拉框中选择具有特定输入输出参数的工作流应用。创建工作流应用的具体操作请参考**创建工作流**。 ● 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，选择执行对应的工作流。
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。 创建工作流应用的具体操作请参考**创建工作流**。
子工作流	在下拉框中选择工作流应用。选择后，设置该工作流的执行动作。 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。 在“子工作流”右侧，单击  ，添加多个子工作流。 创建工作流应用的具体操作请参考**创建工作流**。
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。 在下拉框中选择子工作流。选择后，设置该工作流的执行动作。 支持如下执行动作： ● 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。 创建工作流应用的具体操作请参考**创建工作流**。
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。 创建工作流应用的具体操作请参考**创建工作流**。

参数	说明
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。
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。 ● 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。 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。

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。
  单击工作流卡片，显示工作流适用场景、输入输出信息，可以修改适用场景。
- 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**调试与发布多智能体应用**。

----结束

相关操作

在多智能体应用卡片列表中，支持的其他操作请参考**表8-3**。

表 8-3 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。
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”。 该多智能体应用的API接口被调用时，路径中“:conversation_id”对应API接口的参数“conversation_id”，为会话ID，每个会话的唯一标识符，可将会话ID设置为任意值，使用标准UUID格式。
删除多智能体应用	在待删除的多智能体应用卡片上，单击“ ··· > 删除”。

8.3 调试与发布多智能体应用

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**创建多智能体应用**。

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，选择进入所需空间。

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 8-3 选择团队空间

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”。

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。

----结束

发布多智能体应用

步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择进入所需空间。

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 8-4 选择团队空间

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8-4。

表 8-4 发布多智能体参数说明

参数	说明
版本号	系统自动生成带年月日的版本号，以v开头。也可以自定义版本号，由1~32个字符组成。 此版本号在多智能体应用被API调用时，为“version”参数的值。
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”。

----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8-5**。

表 8-5 多智能体应用相关操作

操作	说明
查看发布历史	单击  ，查看发布历史记录，在发布历史记录中，可以获取发布的版本号及ID。 版本号在多智能体应用被API调用时，为“version”参数的值。

相关文档

多智能体应用发布后，可以在Versatile空间使用，也可以通过API接口调用，请参考**使用Versatile空间**、Versatile《API参考》中“调用智能体应用”章节。

8.4 使用 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。

前提条件

在调用应用前，须确保应用已发布，具体请参考**发布多智能体应用**。

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，选择进入所需空间。

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 8-5 选择团队空间

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 8-6 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6-80所示。

其中，“:conversation_id”参数为会话ID，每个会话的唯一标识符，可将会话ID设置为任意值，使用标准UUID格式。

图 8-7 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《 API参考 》“应用示例 > 调用智能体应用示例”章节。

8.5 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**创建多智能体应用**。

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，选择进入所需空间。

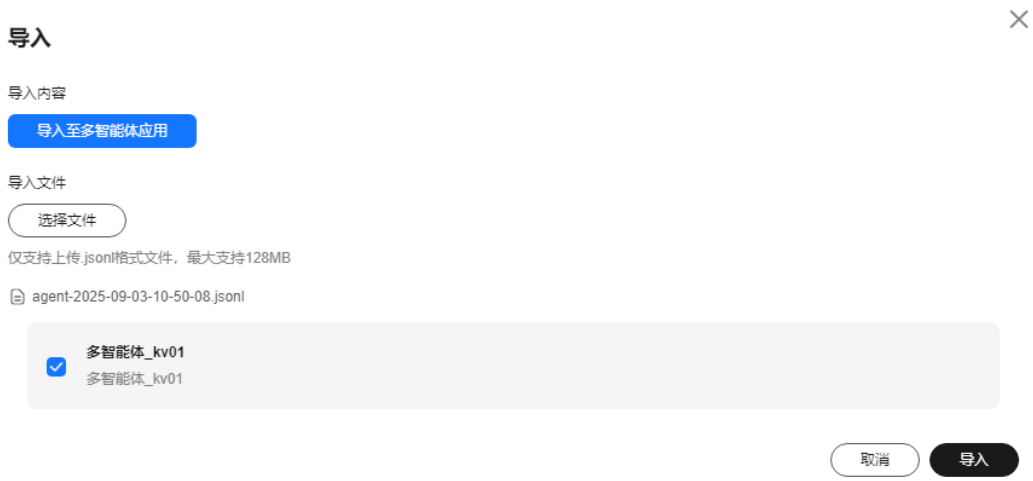
如果已选择团队空间，界面显示为实际的团队空间名称，而非“个人空间”。

图 8-8 选择团队空间

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”。

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”。

图 8-9 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

步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择进入所需空间。

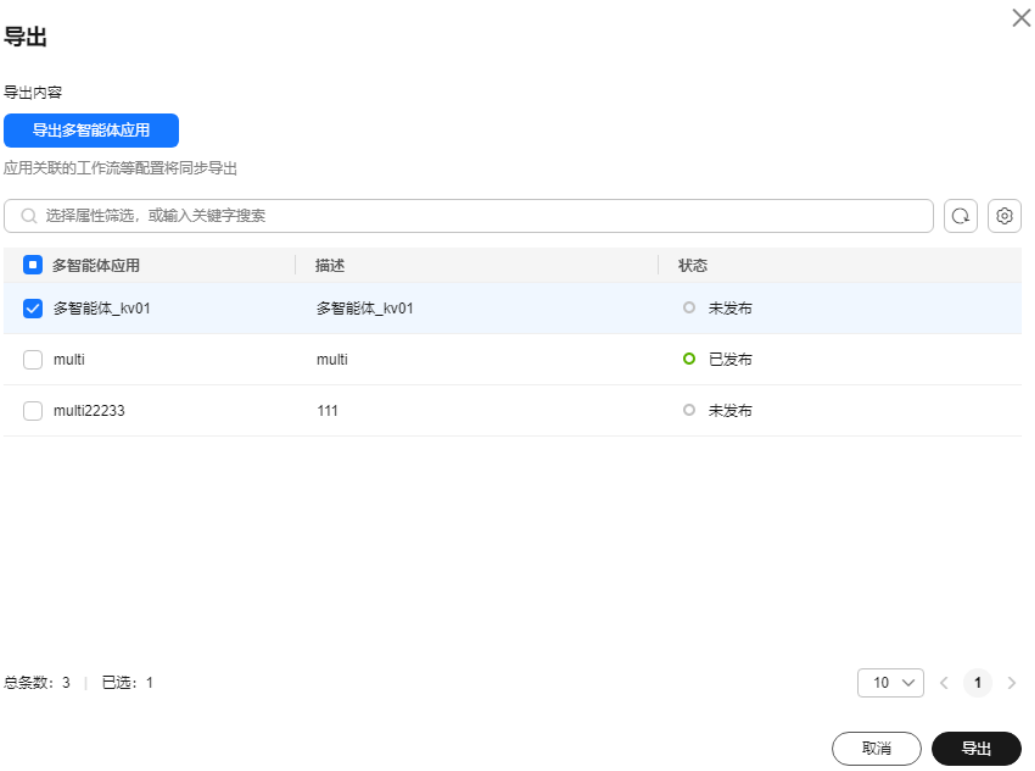
如果已选择团队空间，界面显示为实际的团队空间名称，而非“个人空间”。

图 8-10 选择团队空间

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”。

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”。

图 8-11 导出多智能体应用



导出的文件以JSON格式显示在本地。

----结束

9 管理资源

9.1 插件

9.1.1 插件介绍

插件介绍

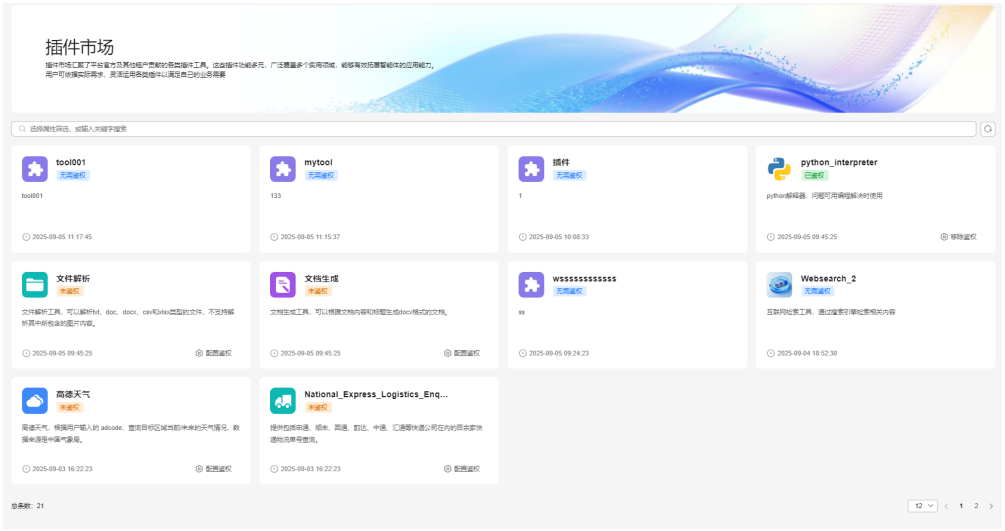
在Versatile中，插件是智能体能力的重要扩展工具。通过模块化设计，插件能够为智能体提供丰富的专业技能和复杂任务处理能力，帮助其在多样化的实际场景中更高效地满足用户需求。

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支持两种类型的插件，满足不同用户的需求：

- 预置插件：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 9-1 插件广场



- 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。

9.1.2 创建插件

9.1.2.1 基于 API 创建一个插件

本章节将介绍如何通过API创建插件，创建插件后，必须发布插件才可以被智能体或工作流使用。

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，选择进入所需空间。

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 9-2 选择团队空间

步骤2 在左侧导航栏中选择“开发中心 > 组件库 > 我的插件”，单击右上角“创建”。

步骤3 在“创建插件”页面，配置插件信息，单击“确定”完成插件创建。

1. 在“基本信息”步骤中设置插件的基础信息，请参照表9-1完成信息配置。

表 9-1 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。
插件名称	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。
插件英文名称	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。
插件描述	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。

2. 在“配置信息”步骤中配置插件信息，请参照表9-2完成配置。您也可以通过单击“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 9-2 配置信息

参数	说明
插件URL	插件的访问地址或相关资源的链接。例如：http://ip/v3/weather/weatherInfo。 – 支持协议：仅支持HTTP和HTTPS。 – 格式校验：系统会校验URL是否为标准格式。 – IP限制：URL对应的IP默认不应为内网地址，否则会导致注册失败。仅在非商用环境部署时，才允许支持内网URL，且需要通过相关服务的启动配置项关闭内网屏蔽。
请求方式	插件服务的请求方式，支持POST或GET。
请求头	填写API的请求头信息，需根据API的参数配置要求填写。 例如： – Key：Content-Type – Value：application/json

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

3. 在“参数信息”步骤中设置插件的请求参数和输出参数信息，请参照**表9-3**完成信息配置。

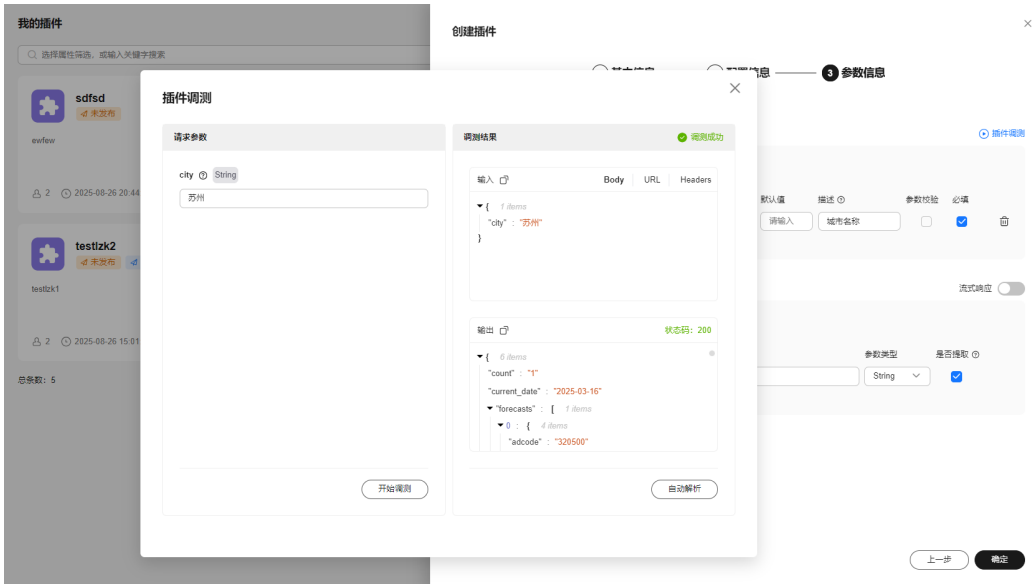
表 9-3 参数信息

参数类型	参数	说明
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}]。
	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、数字或下划线。
	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。
	位置	设置当前参数在请求信息中的位置。 支持以下三种类型：Body、Headers、Query。
	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。

参数类型	参数	说明
	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： ■ 时间日期格式 ■ 字符长度限制 ■ 枚举值范围 – 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。
响应参数	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}]。
	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。
	参数名称	设置响应参数的名称，用于大模型解析大模型输出结果。 命名规则：仅支持字母、数字或下划线。
	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。
	是否提取	开启后则该参数必须提取到，关闭则该参数允许为空或者使用默认值。适用于需要强制提取某些关键参数的场景。

步骤4 单击“插件调测”按钮，输入请求参数值，单击“开始调测”检查调测结果。

图 9-3 插件调测



步骤5 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6 调测成功后，单击“确定”完成插件创建。

----结束

更多操作

插件创建完成后，可根据需要执行**发布插件**操作。

插件创建示例（以 API 调用为例）

准备工作：

- 创建一个服务，此服务需可以访问。

  以天气查询服务为例，准备相应的接口地址、其请求方法、输入参数、输出参数。

  接口地址：http://ip/v3/weather/weatherInfo（此地址需要根据实际情况填写ip、端口和服务路径）；

  请求方法：GET；

  输入参数（示例）：

  {

  "city": "苏州"

  }

  输出参数（示例）：

```
{
    "status": "1",
    "count": "1",
    "info": "OK",
    "current_date": "2025-03-16",
    "infocode": "10000",
    "forecasts": [{
        "city": "苏州市",
        "adcode": "320500",
```

```
    "province": "江苏",
    "casts": [{
            "date": "2025-03-16",
            "week": "7",
            "dayweather": "晴",
            "nightweather": "多云",
            "daytemp": "9",
            "nighttemp": "2",
            "daywind": "西北",
            "nightwind": "西北",
            "daypower": "1-3",
            "nightpower": "1-3",
            "daytemp_float": "9.0",
            "nighttemp_float": "2.0"
        },
        {
            "date": "2025-03-17",
            "week": "1",
            "dayweather": "晴",
            "nightweather": "多云",
            "daytemp": "10",
            "nighttemp": "4",
            "daywind": "西",
            "nightwind": "西",
            "daypower": "1-3",
            "nightpower": "1-3",
            "daytemp_float": "10.0",
            "nighttemp_float": "4.0"
        },
        {
            "date": "2025-03-18",
            "week": "2",
            "dayweather": "晴",
            "nightweather": "多云",
            "daytemp": "10",
            "nighttemp": "1",
            "daywind": "西北",
            "nightwind": "西北",
            "daypower": "1-3",
            "nightpower": "1-3",
            "daytemp_float": "10.0",
            "nighttemp_float": "1.0"
        }
    ]
  }]
}
```


步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择进入所需空间。

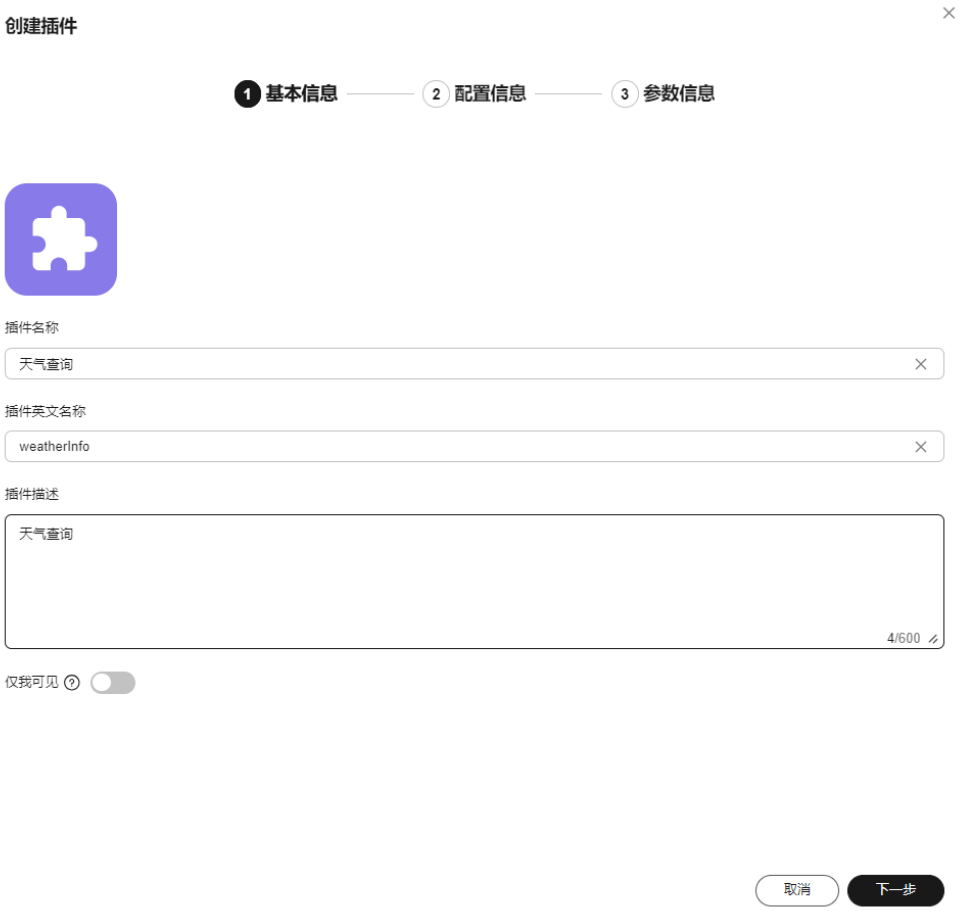
如果已选择团队空间，界面显示为实际的团队空间名称，而非“个人空间”。

图 9-4 选择团队空间

步骤2 在左侧导航栏中选择“开发中心 > 组件库 > 我的插件”，单击右上角“创建”。

步骤3 在“创建插件”页面中，按照以下步骤完成插件的配置信息。

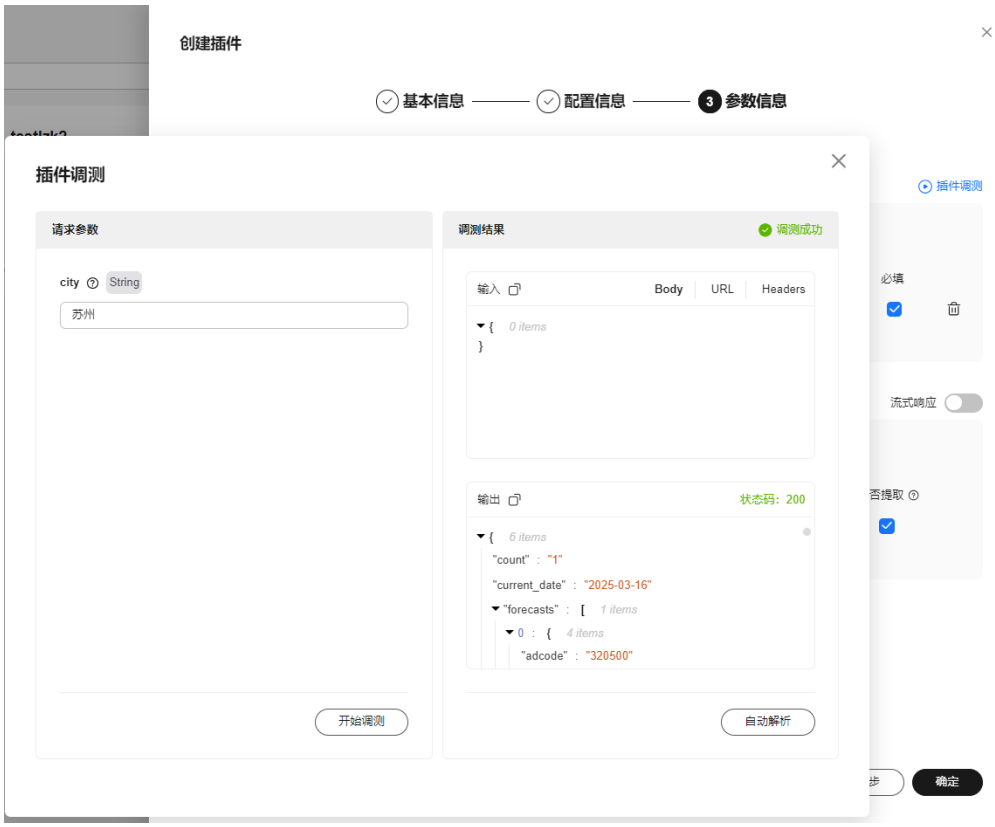
1. 在“基本信息”步骤中设置插件的基础信息。



2. 在“配置信息”步骤中设置插件的URL、请求方法和权限校验信息。

3. 在“参数信息”步骤中根据API接口信息配置参数信息。

4. 单击“插件调测”，进行测试。



5. 调测成功后，单击“自动解析”系统会自动生成响应参数。

步骤4 单击“确定”完成插件创建。

----结束

9.1.2.2 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。

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，选择进入所需空间。

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”。

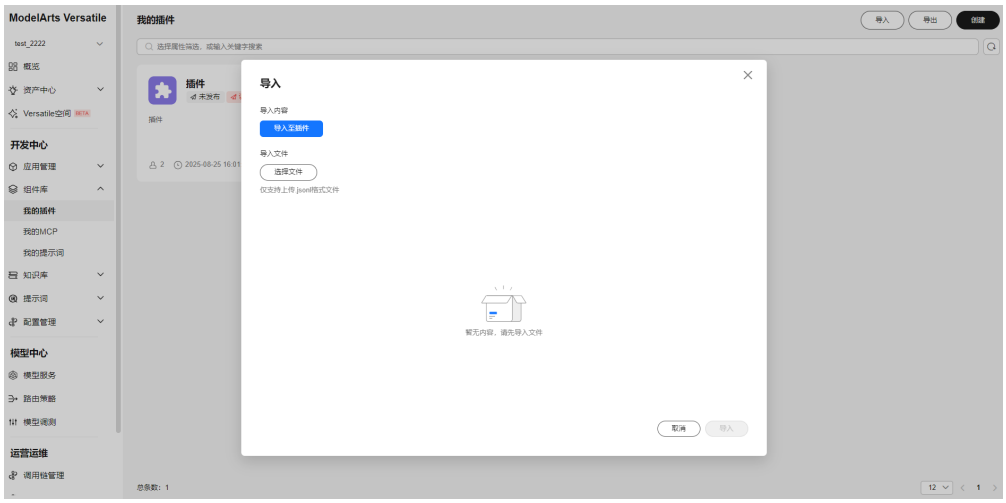
图 9-5 选择团队空间

步骤2 在左侧导航栏中选择“开发中心 > 组件库 > 我的插件”，单击右上角“导入”。

步骤3 导入插件。

1. 单击页面右上角“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，平台将自动解析jsonl文件。如果解析的文件在平台已存在，勾选该文件将自动覆盖平台现有文件。
4. 单击“导入”，导入成功的插件将在“我的插件”页面中展示。

图 9-6 导入 JSON 文件



----结束

更多操作

插件创建完成后，可根据需要执行**发布插件**操作。

9.1.3 发布插件

插件必须经过调测成功之后才能发布。发布状态的插件才能被智能体和工作流使用。

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，选择进入所需空间。

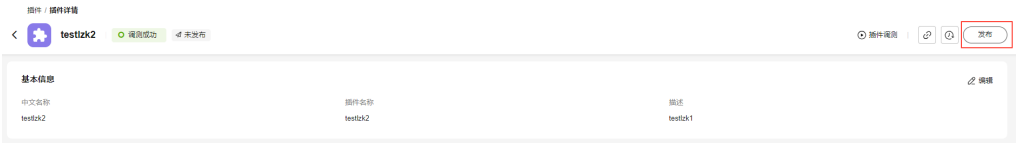
如果已选择团队空间，界面显示为实际的团队空间名称，而非“个人空间”。

图 9-7 选择团队空间

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“发布”按钮。

图 9-8 发布插件



----结束

9.1.4 使用插件

9.1.4.1 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。
- 如果需要添加预置插件，请确保已对插件进行鉴权，详细信息请参考使用预置的插件。

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，选择进入所需空间。

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 9-9 选择团队空间

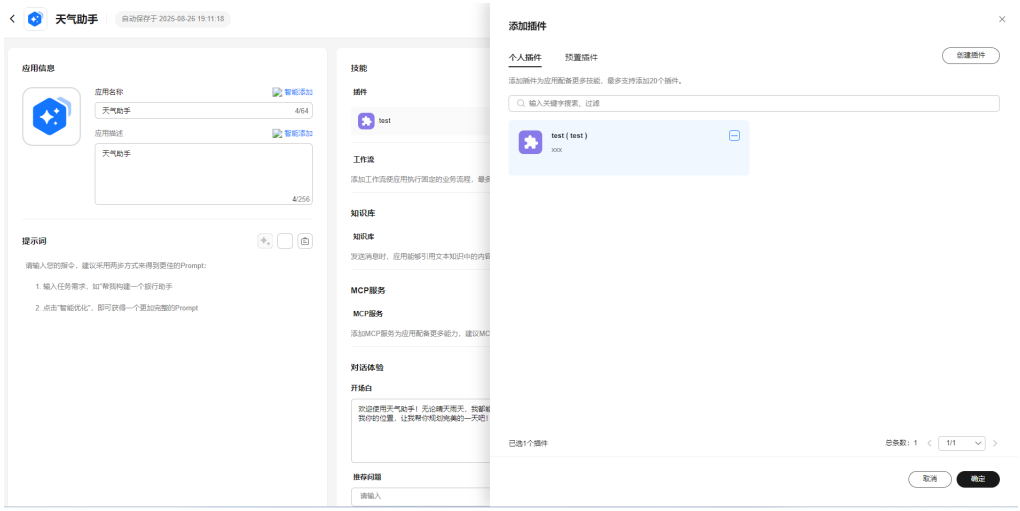
步骤2 在左侧导航栏中选择“应用管理 > 单智能体应用”页面。

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”窗口，您可以选择“个人插件”或“预置插件”，单击目标插件右侧 + 进行添加，并单击“确定”完成添加。

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。

图 9-10 在智能体中添加插件



----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**。

9.1.4.2 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**。
- 如果需要添加预置插件，请确保已对插件进行鉴权，详细信息请参考**使用预置的插件**。

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，选择进入所需空间。

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 9-11 选择团队空间

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”窗口，您可以选择“个人插件”或“预置插件”，单击目标插件右侧 + 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。

图 9-12 在工作流中添加插件

----结束

相关文档

在工作流中应用中使用插件的详细信息，请参考**插件**。

9.1.5 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。

管理插件版本

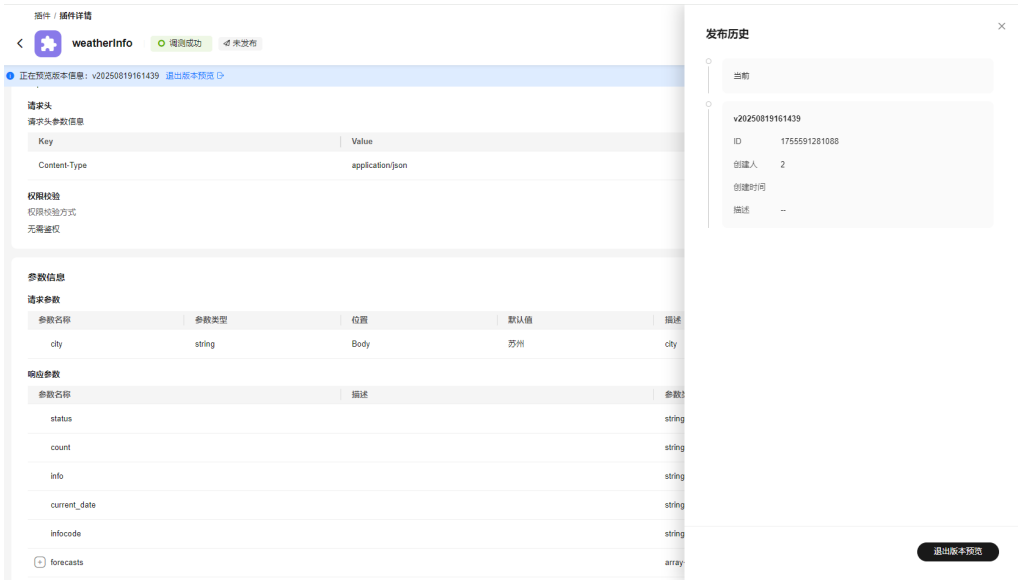
发布插件时，您需要设置插件的版本名称和描述，如**图9-13**所示。这些信息会被记录在插件的历史版本页面，方便您以后查看和参考。发布成功后，系统会自动生成一个新的版本发布记录。

图 9-13 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。

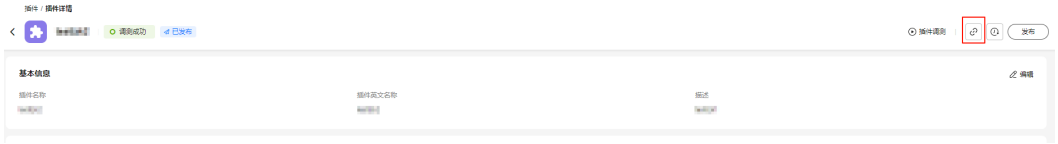
图 9-14 插件发布历史



查看插件引用列表

在插件详情页面右上角，单击引用插件图标  ，即可查看当前插件被哪些单智能体应用和工作流应用所使用。

图 9-15 引用插件

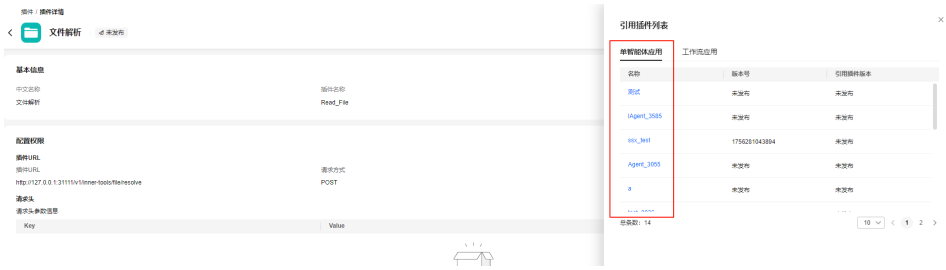


- 单智能体中引用的插件无论是平台预置的插件还是个人创建的插件，都不会自动更新到最新版本。需要手动更新。

  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. 在智能体列表中，搜索并找到引用该插件的智能体，然后手动进行更新。

  图 9-16 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9-17所示。

  图 9-17 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，选择进入所需空间。

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 9-18 选择团队空间

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 9-19 删除插件

说明

删除插件属于高危操作，删除前，请确保该插件不再使用。

----结束

导入、导出插件

Versatile提供了插件的导入和导出功能。"导出插件"功能允许用户将自定义的插件配置保存为本地文件，便于后续的迁移、备份或共享。

步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择进入所需空间。

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 9-20 选择团队空间

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 导出插件

1. 单击页面右上角“导出”。
2. 在“导出”页面选择需要导出的插件，单击“导出”。插件将以一个jsonl格式的文件下载至本地。

步骤4 导入插件

1. 单击页面右上角“导入”。
2. 在“导入”页面，单击“选择文件”从本地选择需要导入的JSONL格式文件。
3. 选择导入文件后，如果解析的文件在平台已存在，勾选该文件之后单击导入系统将自动覆盖平台现有文件。
4. 导入成功的插件将在“开发中心 > 组件库 > 我的插件”页面中展示。

----结束

9.2 MCP

9.2.1 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。

9.2.2 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

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。

说明

自主开发的MCP服务需在服务器或本地上独立部署，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，选择进入所需空间。

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 9-21 选择团队空间

步骤2 在左侧导航栏选择“开发中心 > 组件库 > 我的MCP”，单击“创建”创建MCP服务。

步骤3 在创建MCP服务的弹框中，您可以选择空白模板或平台提供的其他MCP服务模板，如图9-22所示。然后单击“下一步”继续操作。

对于资产中心中未安装的预置MCP，您也可以按照以下步骤进入服务创建页面：在“资产中心 > MCP广场 ”页签，将鼠标光标移至MCP卡片上，单击“安装”按钮。

- 空白模板：适用于安装开源社区的MCP或接入自主开发的MCP服务。
- 平台预置MCP：适用于安装平台预置的MCP。

图 9-22 创建 MCP 服务

步骤4 配置MCP服务，参数的具体说明请参考表9-4所示。

表 9-4 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。
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	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。

参数	说明
安装方式	选择MCP服务的安装方式，支持以下三种方式： ● 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MCP服务。

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，将明文显示加密变量值，加密变量以_encrypt_开头，请谨慎操作。
	– 安装平台预置的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。
	– 如果选择SSE方式接入自主开发的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。
	`{ "mcpServers": { "example-sse": { "url": "https://example.com?key=<您在服务官网上申请的key>" } } }`
	● 使用表单编辑
	– 参数值：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。
	– 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>。仅在选择SSE方式安装时需要配置。
	– 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

步骤5 单击“保存”。MCP安装完成后，可以在“我的MCP”页面查看创建成功的MCP。

步骤6 创建完成之后，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，可根据需要执行如**表9-5**所示的操作。

表 9-5 更多操作

操作	步骤
删除MCP服务	删除后，MCP服务会下线，请谨慎操作。 在“我的MCP”页面，在MCP卡片上单击“删除”按钮，注意此操作将导致MCP服务下线，请谨慎操作。
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。

9.2.3 使用 MCP 服务

9.2.3.1 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**。
- 如果需要使用平台预置的MCP服务，需确保已安装MCP服务，详细信息请参考**使用预置的MCP**。

说明

自主开发的MCP服务需在服务器或本地上独立部署，并通过测试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，选择进入所需空间。

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 9-23 选择团队空间

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，选择“预置服务”“”或“个人服务”，选择已开通的MCP服务单击 + 进行一键添加，单击“确定”。

图 9-24 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**。

9.2.3.2 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**。
- 如果需要使用平台预置的MCP服务，需确保已安装MCP服务，详细信息请参考**使用预置的MCP**。

说明

自主开发的MCP服务需在服务器或本地上独立部署，并通过测试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，选择进入所需空间。

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 9-25 选择团队空间

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。

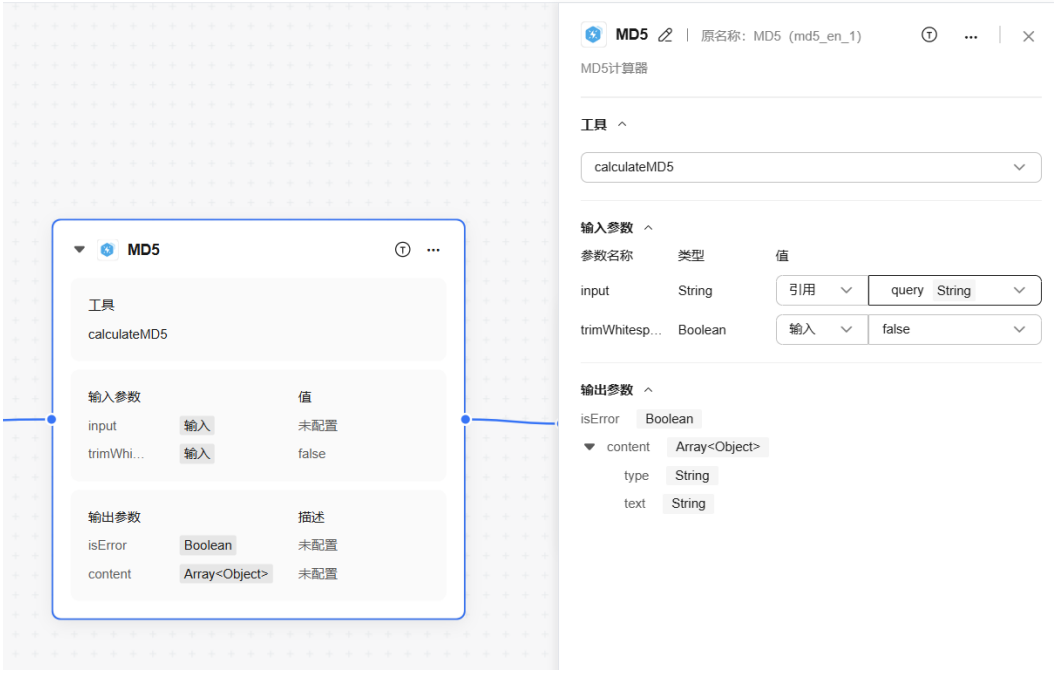
步骤3 单击“添加节点”并选择“MCP服务”节点。

步骤4 在“个人服务”或“预置服务”页签单击+，将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。

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 9-26 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**。

9.2.4 查看 MCP 服务运行监控

运行监控功能提供MCP服务的调用次数、成功率以及操作日志的实时监控，帮助用户全面掌握MCP服务的状态和性能指标。

查看 MCP 服务运行监控

步骤1 **登录Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择进入所需空间。

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”。

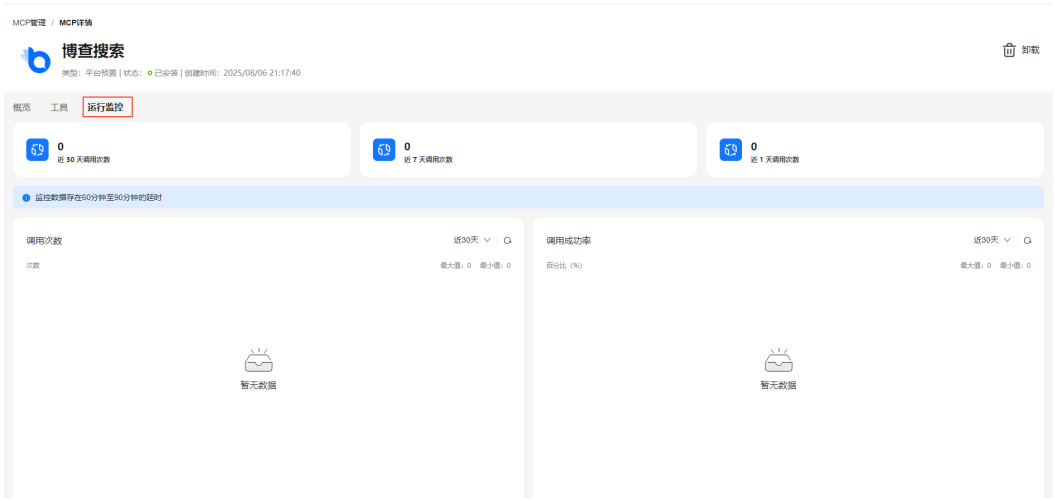
图 9-27 选择团队空间

步骤2 在左侧导航栏选择“开发中心>组件库 > 我的MCP”，单击需要查看的MCP卡片。进入“运行监控”页签，即可查看MCP服务的运行状态，如图9-28所示。

步骤3 运行监控页面提供近30天、近7天、近24小时MCP服务的调用次数统计总览，并可进行如下操作。

- 查看MCP服务调用次数：在“调用次数”区域默认显示近30天的调用次数统计，同时展示当前时间范围内的最大值和最小值。用户可通过右上角的时间选择下拉框，灵活切换查看近30天、近7天或近24小时的调用次数据。
- 查看MCP服务调用成功率：在“调用成功率”区域默认展示近30天的调用成功率统计，同时显示当前时间范围内的最高和最低成功率。用户可通过右上角的时间选择下拉框，切换查看不同时间段的成功率数据。
- 查看MCP服务调用日志：在“日志”区域，用户可以查看MCP服务运行过程中生成的日志记录。支持按时间范围或日志名称进行筛选，快速定位所需信息。

图 9-28 MCP 运行监控



----结束

9.3 知识库

9.3.1 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9-6所示的本地文档：

表 9-6 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：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格式。	单个文档上传限制最大128MB。
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128MB。 • 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。

9.3.2 知识库使用限制

在使用知识库时应注意以下限制。

表 9-7 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。 • 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。 • 单个知识库上传文件数量不多过500个。
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128MB。 • 单个知识库上传文件数量不多于500个。

资源	限制说明
FAQ文档	• 限时免费版单租户文档总大小不大于1G。 • 仅支持xlsx、xls、docx、doc格式的文件，单个文件最大为128MB。 • 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。

9.3.3 创建本地知识库

9.3.3.1 创建本地知识库流程

图 9-29 创建本地知识库流程

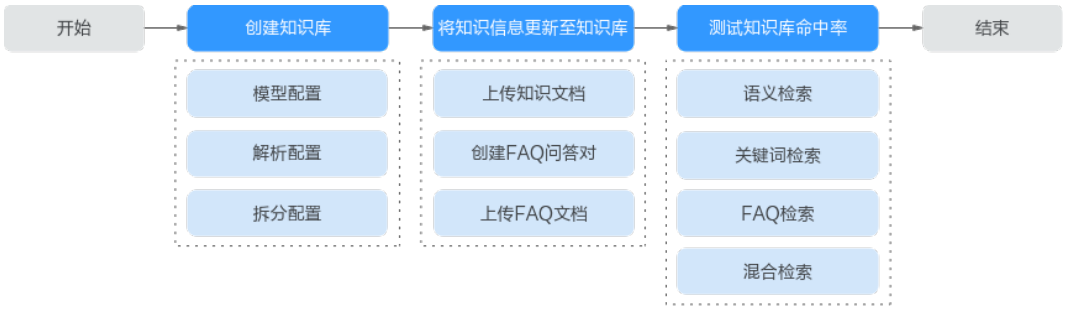


表 9-8 创建本地知识库流程

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3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	在创建知识库后，用户需将知识信息更新至知识库。
		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。

9.3.3.2 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。

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，选择进入所需空间。

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 9-30 选择团队空间

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9-9完成参数配置。

图 9-31 创建知识库

表 9-9 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。	-

参数		说明	示例
	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 _、连字符-，且必须以字母、数字、中文开头。	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： ● embedding-zh：预置的中文模型。 ● embedding-en：预置的英文模型。	embedding-zh
	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： ● rerank-zh：预置的中文模型。 ● rerank-en：预置的英文模型。	rerank-zh
解析配置	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。	OCR增强
	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”。	

参数		说明	示例
拆分配置	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。 – 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，您可以执行如表9-10的操作。

表 9-10 相关操作

操作	说明
启用知识库	在知识库列表中，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。**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	在知识库列表中，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。 **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	在知识库列表中，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	在知识库列表中，单击操作列“更多 > 编辑”，可以编辑知识库，包括修改“知识库图标”、“知识库名称”、“知识库描述”。**只有“状态”是“已停用”的知识库才可编辑。**
高级配置	在知识库列表中，单击操作列“更多 > 高级”，可以编辑知识库，包括修改“模型配置”，“解析配置”，“拆分配置”。**只有“状态”是“已停用”的知识库才可修改高级设置。**
删除数据库	在知识库列表中，单击操作列“更多 > 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。** **说明** 删除应用属于高危操作，删除前，请确保该知识库不再使用。

9.3.3.3 上传知识文档

在创建知识库后，用户需将知识信息更新至知识库。

本文将详细介绍创建知识库后如何上传知识文档。待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，相关操作请参见**（可选）新增文档切片**；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，相关操作请参见**（可选）重新对文档解析拆分**。

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，选择进入所需空间。

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 9-32 选择团队空间

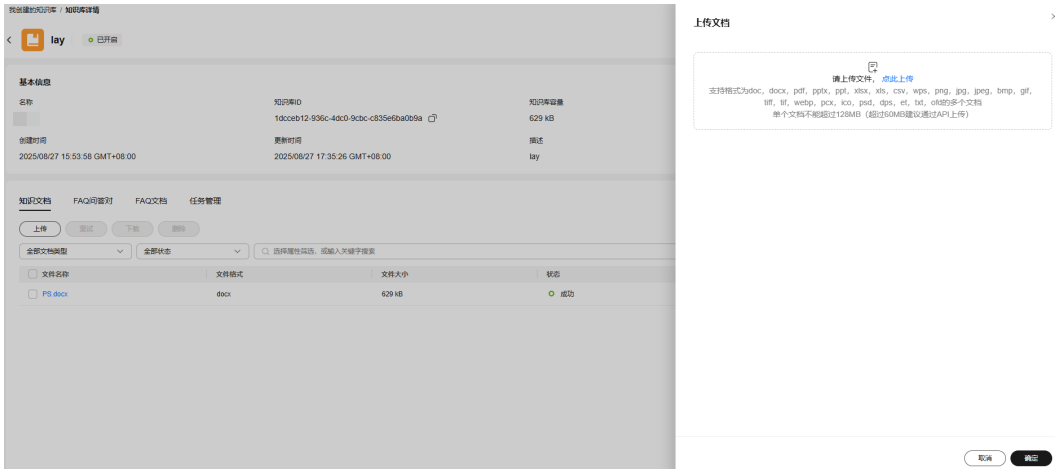
步骤2 在左侧导航栏中选择“开发中心 > 知识库”。

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“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 9-33 上传文档



说明

- 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的多个文档。
- 单个文档不能大于128MB。
- 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。

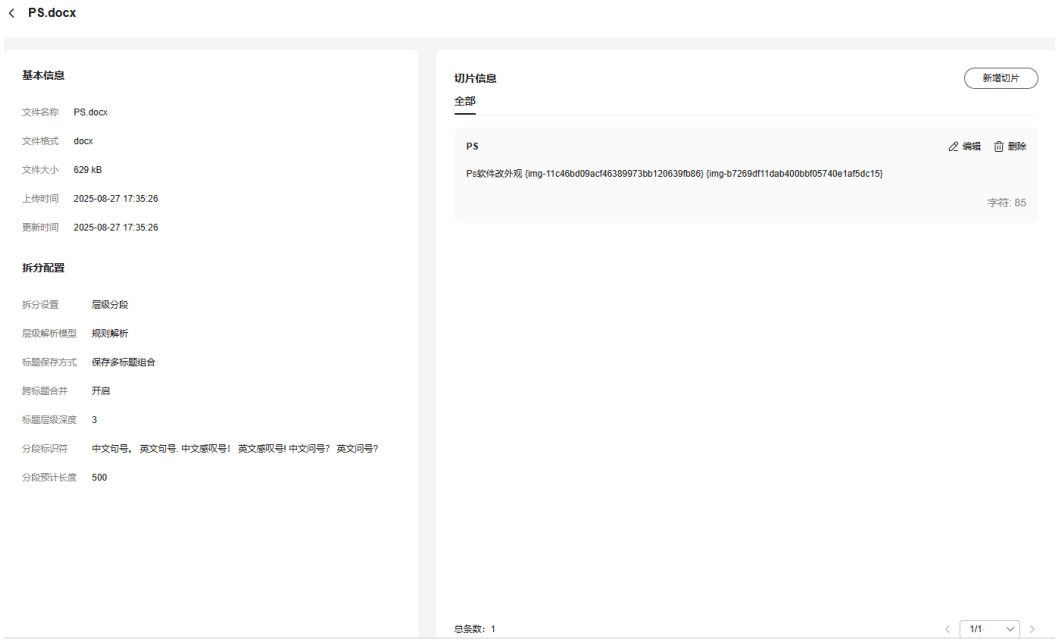
----结束

（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9-34所示。

图 9-34 文档详情



步骤2 单击“新增切片”，参见表9-11设置切片信息。

表 9-11 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”。

----结束

（可选）重新对文档解析拆分

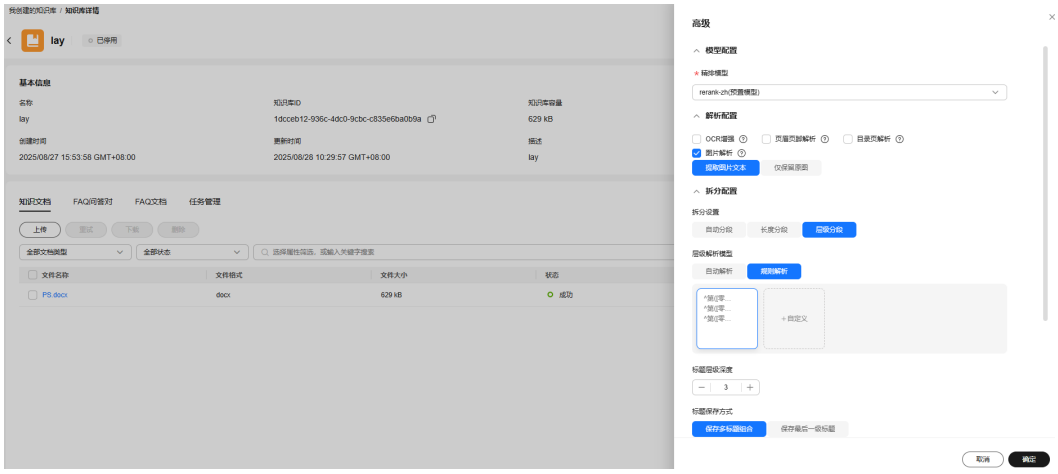
步骤1 在知识库详情页面，单击右上角“高级”。

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9-9修改模型配置、解析配置和拆分配置信息。

图 9-35 修改知识库高级配置

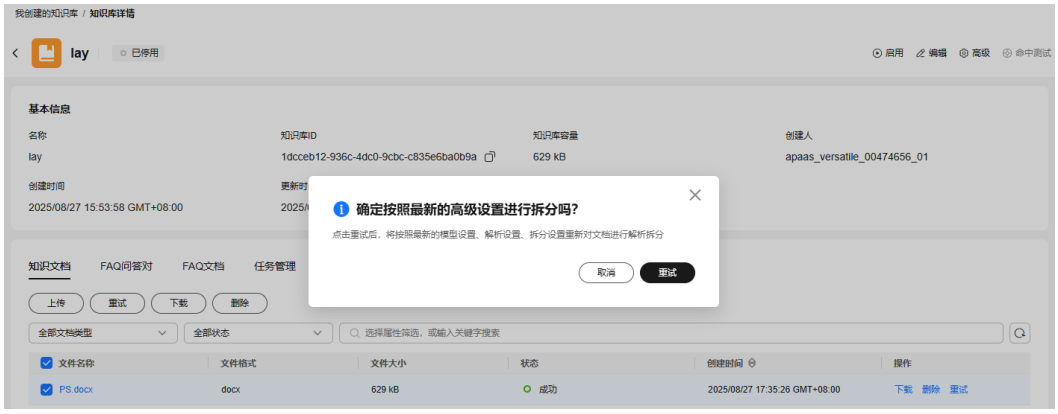


步骤3 修改完成后，单击“确定”。

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 9-36 重试



步骤6 选择“任务管理”页签，可查看重试任务。

图 9-37 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9-12的操作。

表 9-12 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，单击“编辑”，可以编辑文档切片。
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，单击“删除”，可以删除文档切片。

9.3.3.4 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，选择进入所需空间。

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 9-38 选择团队空间

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9-13填写FAQ回答对信息。

图 9-39 FAQ 回答对

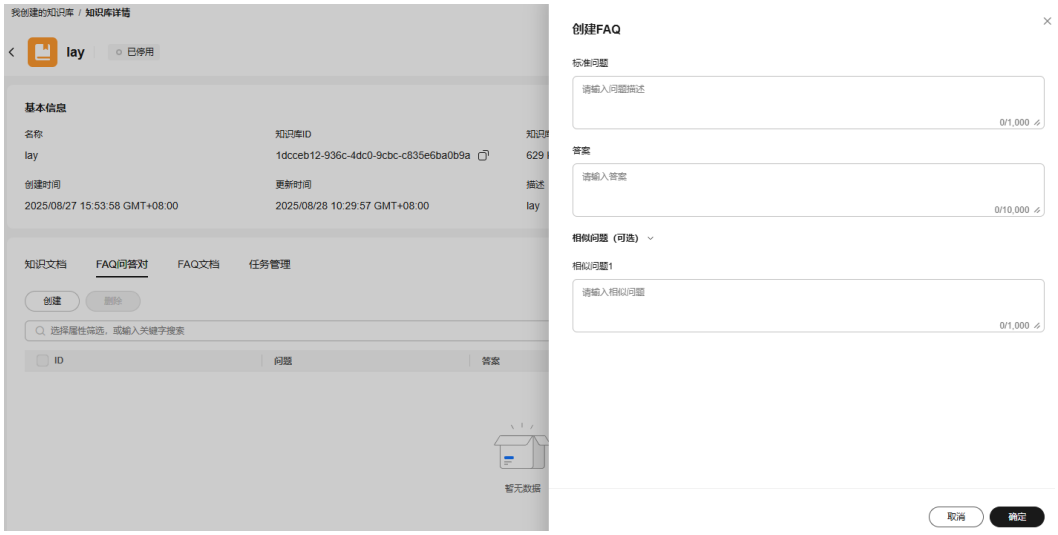


表 9-13 参数说明

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

参数	说明	示例
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1000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。

----结束

更多操作

创建完FAQ问答对后，您还可以执行如**表9-14**的操作。

表 9-14 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。

9.3.3.5 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，选择进入所需空间。

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 9-40 选择团队空间

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。

图 9-41 上传 FAQ 文档

说明

- 支持xlsx、xls、docx、doc文件类型格式。
- 单个文件不能大于128MB。
- 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问题对”页签，可查看对应的问答对记录。

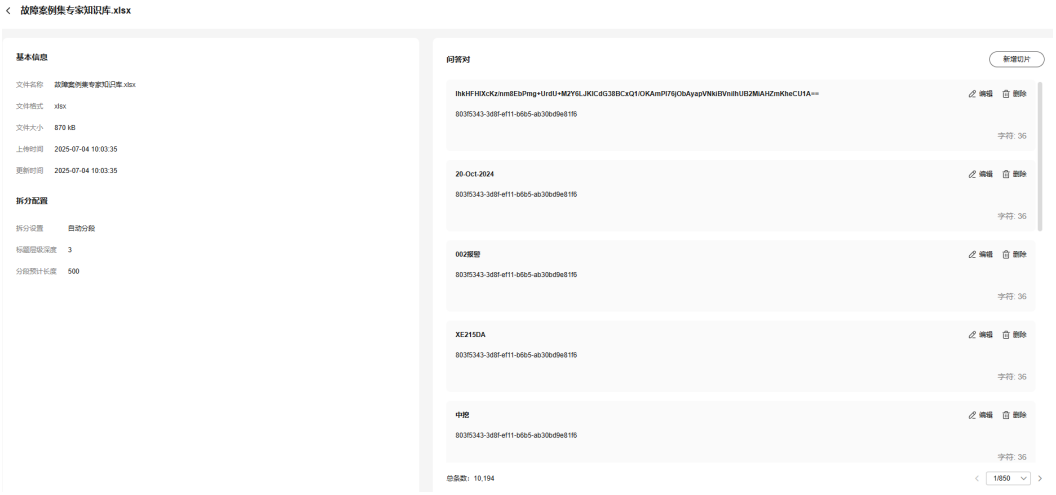
----结束

（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9-42所示。

图 9-42 FAQ 文档详情



步骤2 单击“新增切片”，参见表9-15设置切片信息。

表 9-15 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9-16**的操作。

表 9-16 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。
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。

9.3.3.6 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，选择进入所需空间。

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 9-43 选择团队空间

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”。

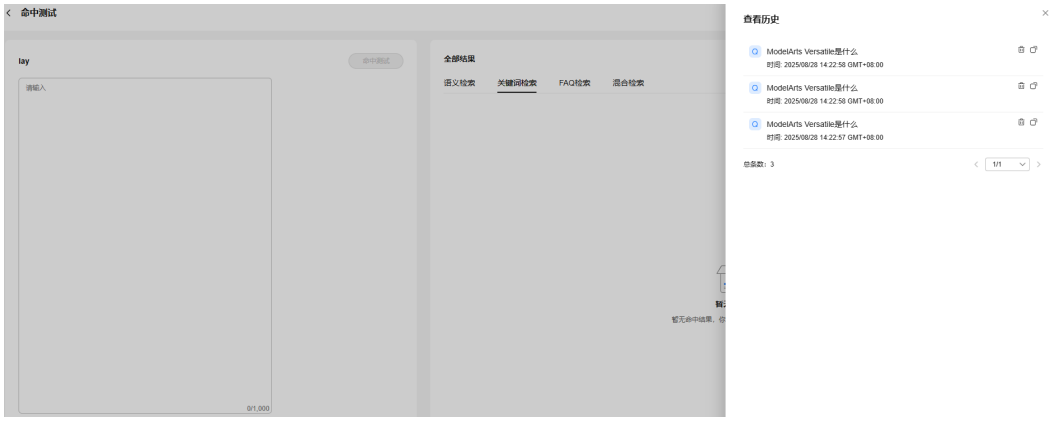
步骤4 在页面左侧文本框中输入问题，单击“命中测试”。

在页面右侧将根据不同的检索方式（语义检索、关键词检索、FAQ检索、混合检索），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 9-44 查看历史



----结束

9.3.4 接入第三方知识库

9.3.4.1 接入第三方知识库流程

图 9-45 接入第三方知识库流程

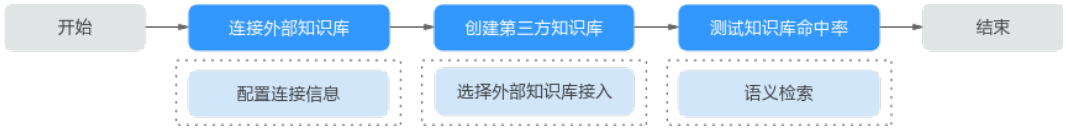


表 9-17 创建第三方知识库流程

序号	流程环节	说明
1	连接外部知识库	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	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。

9.3.4.2 连接外部知识库

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。

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。

前提条件

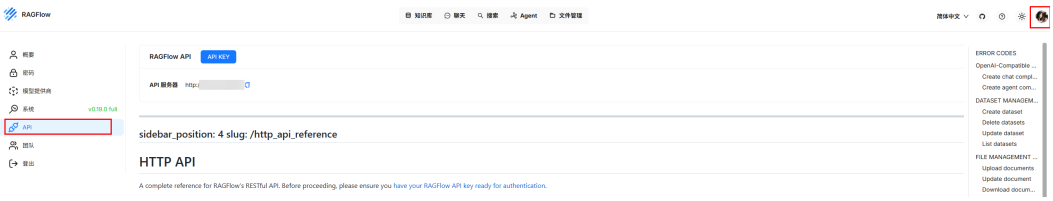
获取第三方RAGFlow知识库的连接信息：

步骤1 在RAGFlow里创建一个知识库，并上传相关文档。

图 9-46 RAGFlow 中的知识库

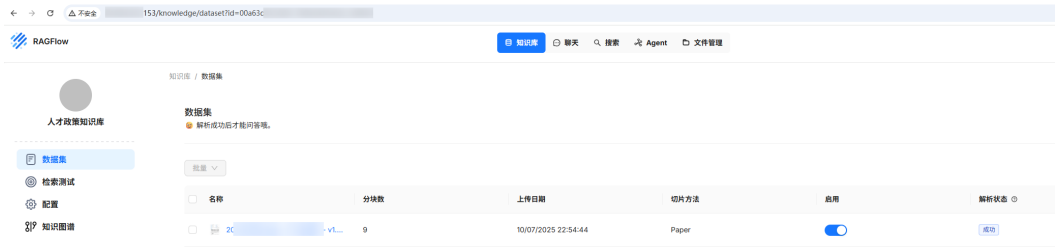
步骤2 在RAGFlow“个人中心 > API”中获取RAGFlow的连接信息，包括API Key和API服务器地址。

图 9-47 RAGFlow 连接信息



步骤3 在第三方知识库中，进入已创建的知识库详情页，在浏览器的地址栏中获取知识库详情页的链接地址（用于从知识库列表中直接跳转到第三方知识库详情页）。

图 9-48 RAGFlow 知识库详情页



----结束

连接外部知识库

步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择进入所需空间。

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 9-49 选择团队空间

步骤2 在左侧导航栏中选择“开发中心 > 知识库”。

步骤3 选择“连接外部知识库”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9-18设置外部知识库基本信息。

图 9-50 连接外部知识库

表 9-18 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。	B030_ragflow_kv04
	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	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/knowledge/dataset?id={{id}}
	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

步骤5 单击“测试连接”，弹出“测试成功”提示。

步骤6 单击“确定”。

----结束

更多操作

连接外部知识库后，您还可以执行如表9-19的操作。

表 9-19 相关操作

操作	说明
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。

9.3.4.3 创建第三方知识库

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。

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**连接外部知识库**。

创建第三方知识库

步骤1 **登录Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择进入所需空间。

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 9-51 选择团队空间

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 9-52 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，您可以执行如表9-20的操作。

表 9-20 相关操作

操作	说明
启用知识库	在知识库列表中，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。**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	在知识库列表中，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。 **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	在知识库列表中，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。
取消接入	在知识库列表中，单击操作列“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。**

9.3.4.4 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，以评估知识库的效果和准确性。

本文将详细介绍第三方知识库创建完成后如何执行命中测试操作。

前提条件

已完成创建第三方知识库。

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，选择进入所需空间。

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 9-53 选择团队空间

步骤2 在左侧导航栏中选择“开发中心 > 知识库”。

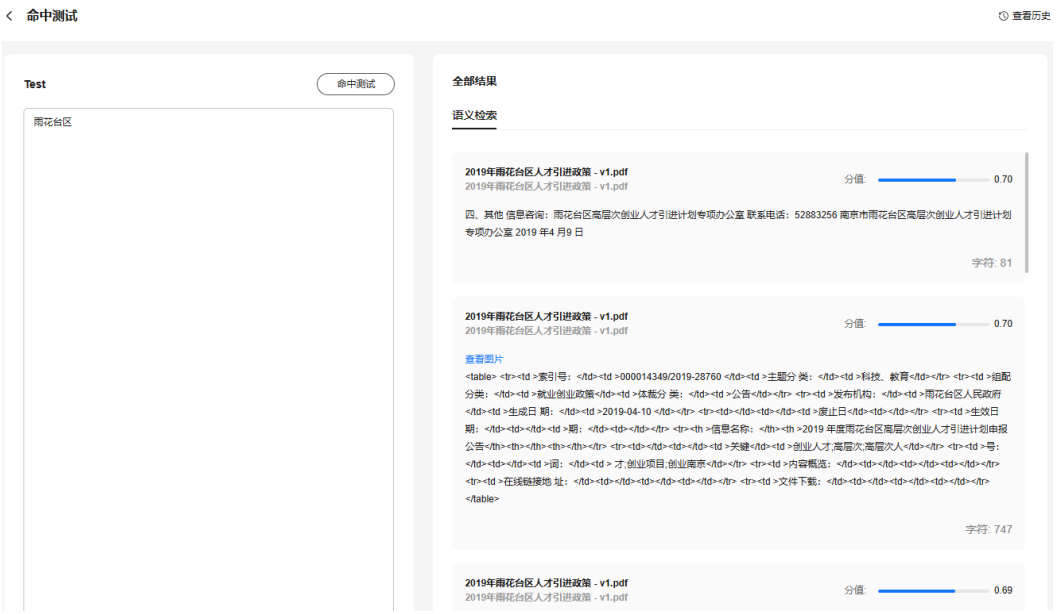
步骤3 在知识库列表中，单击操作列的“命中测试”。

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语义检索方式，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。

图 9-54 命中测试结果



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

9.3.5 使用知识库

9.3.5.1 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，选择进入所需空间。

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”。

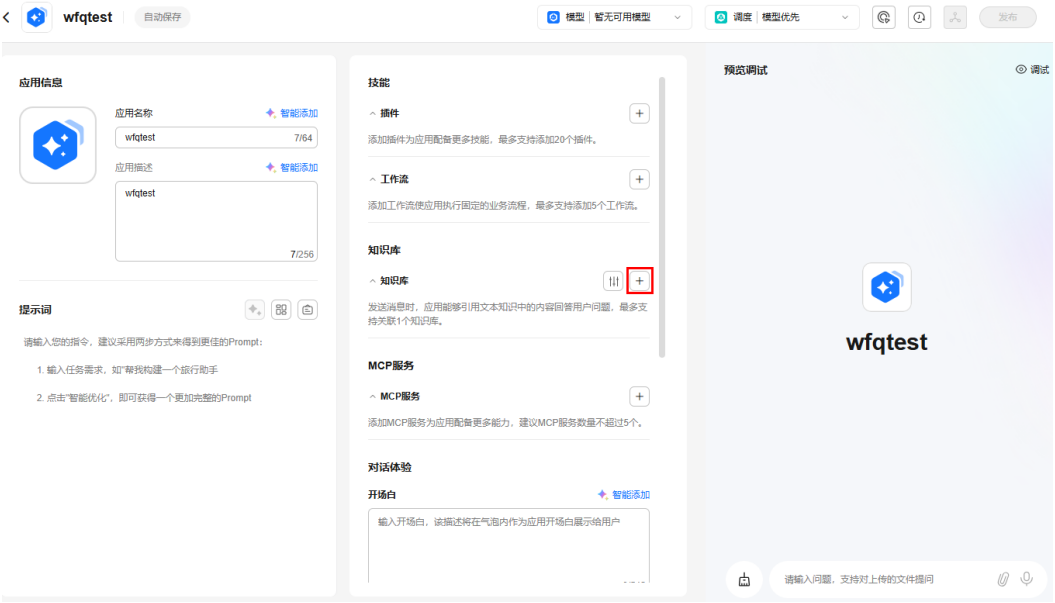
图 9-55 选择团队空间

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。

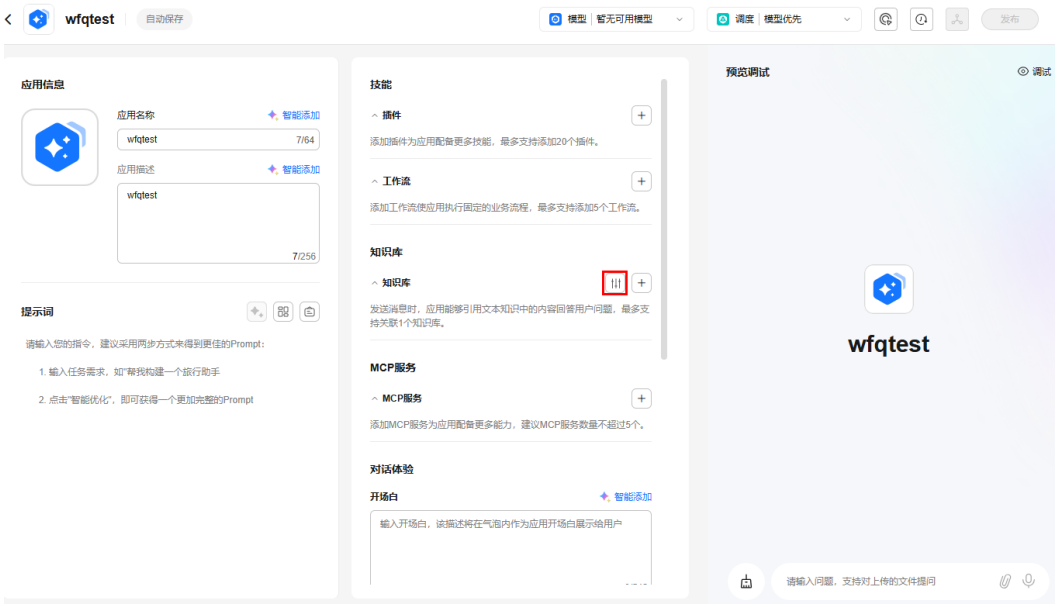
图 9-56 添加知识库



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 9-57 高级配置



步骤7 参见表9-21设置参数。

图 9-58 参数配置

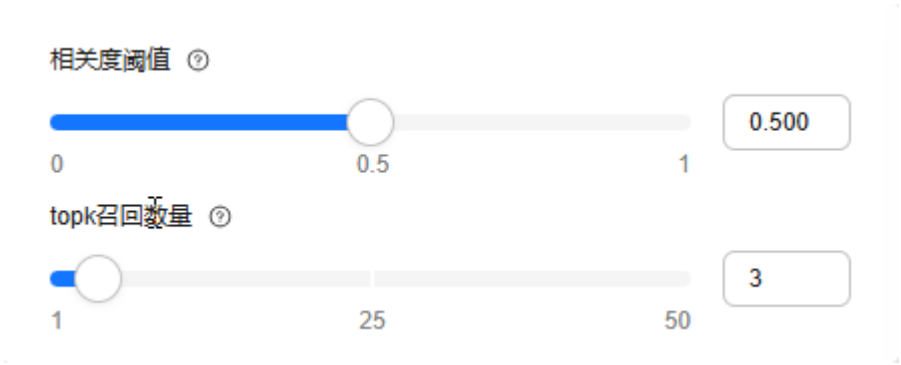


表 9-21 参数说明

参数	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：0.500	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

步骤8 单击其他位置退出弹窗，完成配置。

----结束

9.3.5.2 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，选择进入所需空间。

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 9-59 选择团队空间

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。

图 9-60 输入参数

表 9-22 参数说明

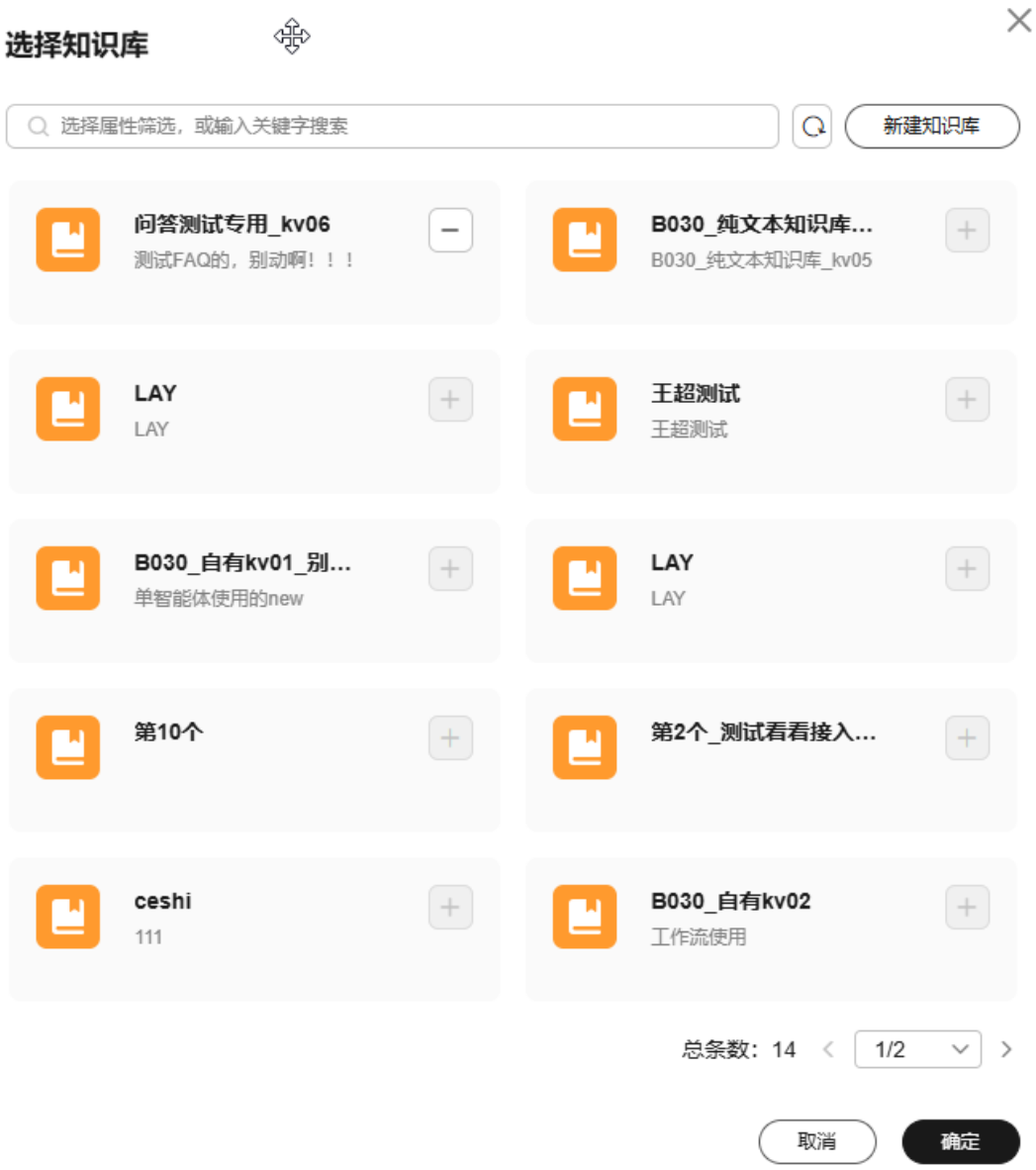
参数名称	说明	示例
输入参数	● 参数名称：输入参数固定只有1个，参数名称为query且不可修改，类型是字符串，表示待知识检索的问题。 ● 类型、值：支持“引用”和“输入”两种类型。 – 引用：支持用户选择工作流中已包含的前置节点的输出变量值以及全局配置中的记忆变量，限制String类型，适用于需要从前置节点输出中获取知识检索问题的场景。 – 输入：支持用户自定义输入问题，适用于知识检索问题固定的场景。	类型：输入 值：1
输出参数	知识检索节点的输出是一个对象数组，参数名是output_list，表示所有满足检索要求的知识切片。数组中对象有四个属性： ● document_name，知识切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 9-61 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 9-62 配置检索参数

表 9-23 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
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：0.100	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

参数名称	说明	示例
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	0.100

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

9.4 提示词

9.4.1 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？”

9.4.2 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。

9.4.3 通过提示词工程创建高质量提示词

9.4.3.1 创建提示词工程

提示词工程（Prompt Engineering）是一个较新的学科，应用于开发和优化提示词（Prompt），帮助用户有效地将大语言模型用于各种应用场景和研究领域。掌握提示词工程相关技能将有助于用户更好地了解大语言模型的能力和局限性。

提示词工程不仅是关于设计和研发提示词，它还包含了与大语言模型交互和研发的各种技能和技术。提示工程在实现和大语言模型交互、对接，以及理解大语言模型能力方面都起着重要作用。用户可以通过提示词工程来提高大语言模型的安全性、赋能大

语言模型，如借助专业领域知识和外部工具来增强大语言模型的能力。通过本文档，您可以了解如何创建、编辑和删除提示词工程，以确保资源库的持续优化。

创建提示词工程

步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择进入所需空间。

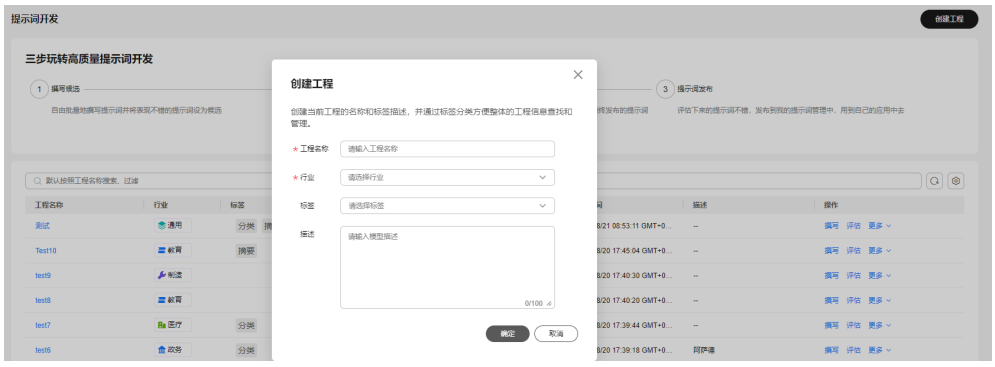
如果已选择团队空间，界面显示为实际的团队空间名称，而非“个人空间”。

图 9-63 选择团队空间

步骤2 在左侧导航栏中选择“开发中心 > 提示词 > 提示词开发”，单击界面右上角“创建工程”。

步骤3 设置提示词的工程名称、行业、标签、描述之后，单击“确定”完成提示词工程创建。

图 9-64 创建提示词工程



步骤4 建完成后，可以在该提示词工程中编写提示词。

----结束

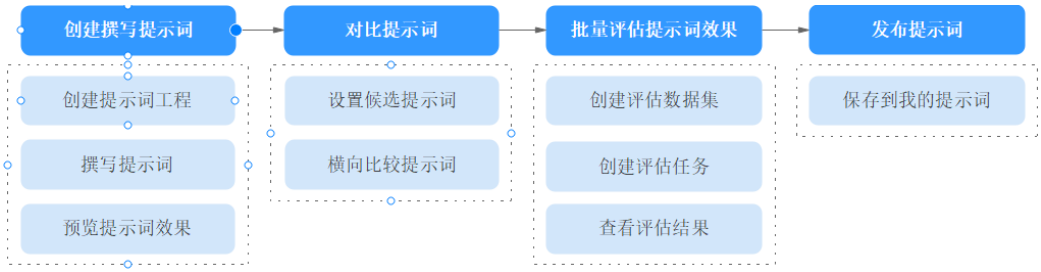
提示词工程使用流程

Versatile可以辅助用户进行提示词撰写、比较和评估等操作，并对提示词进行保存和管理。

表 9-24 功能说明

功能	说明
提示词工程任务管理	提示词工程平台以提示词工程任务为管理维度，一个任务代表一个场景或一个调优需求，在提示词工程任务下可以进行提示词的调优、比较和评估。 提示词工程任务管理支持工程任务的创建、查询、修改、删除。
提示词撰写	提示词撰写支持对提示词文本的编辑、提示词变量设置、提示词结果生成和调优历史记录管理。
提示词候选	提示词候选支持用户对表现较好的提示词进行候选管理，每个提示词工程任务下最多可以保存9个候选提示词，并进一步基于候选提示词进行比较和评估。
提示词比较	提示词比较支持选择两个候选提示词对其文本和参数进行比较，支持对选择的候选提示词设置相同变量值查看效果。
提示词评估	提示词评估以任务维度管理，支持评估任务的创建、查询、修改、删除。支持创建评估任务，选择候选提示词和需要使用的变量数据集，设置评估算法，执行任务自动化对候选提示词生成结果和结果评估。
提示词管理	提示词管理支持用户对满意的候选提示词进行保存管理，同时支持提示词的修改、删除。

图 9-65 提示词工程使用流程



9.4.3.2 撰写提示词

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。例如，将提示词设置为“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。”在评估提示词效果时，可以通过批量替换{{location}}的值，来获得模型回答，提升评测效率。

同时，撰写提示词过程中，可以通过设置模型参数来控制模型的生成行为，如调整温度、核采样、最大Token限制等参数。模型参数的设置会影响模型的生成质量和多样性，因此需要根据不同的场景进行选择。

撰写提示词

步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择进入所需空间。

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 9-66 选择团队空间

步骤2 在左侧导航栏中选择“开发中心 > 提示词 > 提示词开发”。

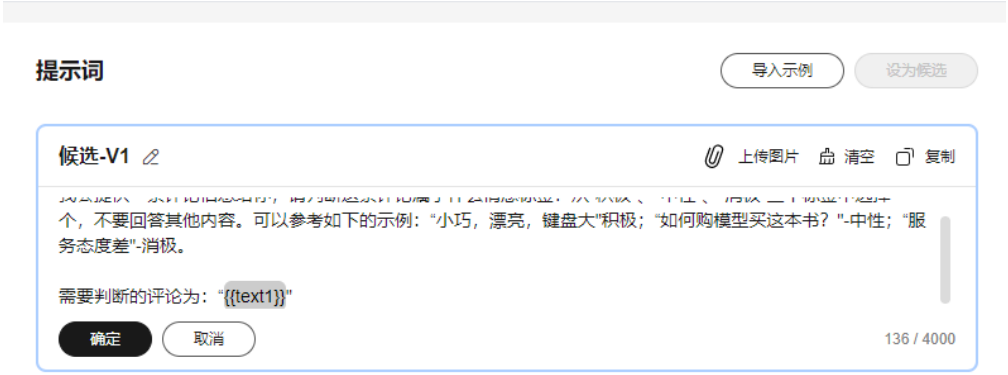
步骤3 在工程任务列表页面，找到目标提示词工程，单击该工程右侧“撰写”。如果需要创建新的提示词工程，请参考创建提示词工程。

图 9-67 撰写提示词

步骤4 在提示词撰写区域，您可以直接输入文本并添加图片，单击右上角的“上传图片”上传文件。提示词中还可以插入多个变量，变量需要用 {{}}标识。您可以通过以下两种方式创建提示词：

- 直接在提示词编辑区域输入内容。
- 单击“导入示例”，在弹框中选择资产中心提供的提示词模板，单击“确定”模板内容会自动填充到编辑区域，您可以根据需要进行修改。

图 9-68 撰写提示词



步骤5 撰写完成后，单击“确定”，平台将自动识别变量并展示在变量定义区域。变量名称可进行修改。

图 9-69 变量定义

说明

- 变量定义区域展示的是整个工程任务下定义的变量信息，候选提示词中关联的变量也会进行展示，候选提示词相关操作请参见**横向比较提示词效果**。

步骤6 在“模型”区域，单击“设置”，设置提示词输入的模型和模型参数。参数的具体信息请参见**表9-25**。

表 9-25 模型设置参数说明

参数	说明	示例
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**。	Qwen3-235B-A22B-32K
温度	设置推理温度，用于控制生成文本的随机性和创造性，数值越大随机性越大。 • 数值较低，输出结果更加集中和确定。 • 数值较高，输出结果更加随机，更有创意性。 取值范围：0~1 默认值：不同模型的默认值不同，请以实际环境为准。	0.5

参数	说明	示例
核采样	设置推理核采样，用于调整输出文本的多样性。数值越大，生成文本的多样性就越高。 ● 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：不同模型的默认值不同，请以实际环境为准。	0.5
最大Token限制	用于控制聊天回复长度和质量的重要参数。它决定了模型生成回复的最大长度，通常以token为单位。较大的值允许模型生成更长和更完整的回复，但可能会增加生成无关或重复内容的风险。相反，较小的值会生成更短和更简洁的回复，但可能导致内容不够详细或不连贯。 取值范围：1~4096 默认值：不同模型的默认值不同，请以实际环境为准。	2048
话题重复度控制	控制对话中话题重复度。该参数可以设置正值或者负值。 ● 正值：模型倾向于避免重复讨论同一话题，从而保持对话的多样性和广泛性。 ● 负值：模型更倾向于围绕同一话题展开，保持对话的连贯性和深度。 取值范围：-2~2 默认值：不同模型的默认值不同，请以实际环境为准。	0

图 9-70 模型设置

----结束

预览提示词效果

在完成提示词的撰写后，您可以通过输入具体的变量值，生成完整的提示词，并观察不同提示词在模型中的使用效果。

步骤1 在提示词撰写页面中，找到页面右侧变量输入区域，在输入框中输入具体的变量值信息。您也可以单击右边的“导入”，选择导入已经创建好的变量集信息，导入变量集的详细信息请参见**表9-26**。

表 9-26 导入变量集参数说明

参数	说明
存储位置	在对象存储服务（OBS）中选择需要导入的变量集的路径 **说明** • 如果在当前对象存储服务中没有需要的文件，请先前往**OBS**中上传文件。 • 为了确保数据上传和处理的顺利进行，请遵守以下限制条件： 1. 文件格式：仅支持上传xlsx格式的文件。请确保文件为Excel格式，以便系统正确读取数据。 2. 数据集行数：数据集的行数必须在10到50行之间。少于10行或大于50行的数据集将无法导入。请确保数据量适中，以获得最佳处理效果。 3. 表头要求：数据集的表头必须唯一，且数量不得大于20个。请检查表头名称，避免重复，并确保表头数量在限制范围内。 4. 文本长度限制：数据集中每条文本的长度不得大于1000个字符。过长的文本可能导致处理延迟或错误，请适当控制文本长度。
数据集名称	用于唯一标识一个数据集，帮助用户快速识别和管理不同的数据集。
数据集描述	提供更详细的信息，说明数据集的用途、内容和适用场景，帮助用户更好地理解和使用数据集。

步骤2 通过预览区域，用户可以查看不同变量组合生成的提示词，评估其在模型中的使用效果。这有助于优化提示词，提升生成回复的质量和相关性。

图 9-71 提示词效果预览

----结束

9.4.3.3 横向比较提示词效果

设置候选提示词

设置候选提示词功能允许用户将效果较为理想的提示词保存为候选词。用户可以将多个提示词添加到候选列表中，可以对这些候选提示词进行效果对比，帮助用户直观地查看不同提示词在实际应用中的表现差异，从而更高效地选择高质量提示词。

说明

每个提示词工程任务下最多可以设置9个候选提示词，达到上限时需要删除其他候选提示词才能继续添加。

步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择进入所需空间。

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 9-72 选择团队空间

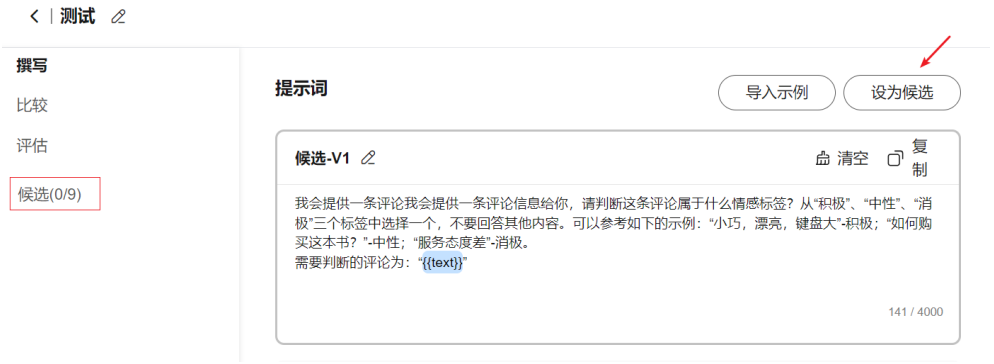
步骤2 在左侧导航栏中选择“开发中心 > 提示词 > 提示词开发”。

步骤3 在工程任务列表页面，找到所需要操作的工程任务，单击该工程任务右侧“撰写”。

图 9-73 提示词工程

步骤4 在提示词撰写区域，单击“设为候选”，将当前撰写的提示词设置为候选提示词。候选状态的提示词将保存至左侧导航栏的“候选”页签中。

图 9-74 设置候选提示词



----结束

横向比较提示词效果

将已被设置为候选的多个提示词进行横向对比分析，通过系统化地识别和展示它们之间的差异，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。在此基础上，进一步评估每个提示词在实际应用中的表现效果，如生成内容的相关性、准确性、流畅性或创意性等，从而为用户在选择高质量的提示词或优化提示词设计时提供有力支持。该功能特别适用于提示词的筛选、优化与迭代，提升模型输出质量和任务完成效率。

步骤1 **登录Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择进入所需空间。

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 9-75 选择团队空间

步骤2 在左侧导航栏中选择“开发中心 > 提示词 > 提示词开发”。

步骤3 在工程任务列表页面，找到所需要操作的工程任务，单击该工程任务右侧“撰写”。

图 9-76 提示词工程

步骤4 在“撰写”页面，选择左侧导航栏中的“候选”。在候选列表中，勾选需要进行横向比对的提示词，并单击“横向比较”。

图 9-77 横向比较提示词

说明

您也可以在左侧导航栏的“比较”页面中，通过下拉框选择要对比的候选提示词。

步骤5 在提示词对比界面中，展示了两个不同维度的比较，帮助您更直观地进行分析和选择。

- 在提示词差异性比较区域，用户可以直观地查看对比的两个提示词使用的模型、温度、核采样等参数的对比。通过单击开启“高亮差异点”按钮，您还可以直观地看到两个比较的提示词内容的差异。

图 9-78 提示词差异性比较

- 在提示词效果对比区域，您可以输入变量的具体信息，单击“查看效果”按钮后，即可实时查看两个提示词生成的效果对比，助您快速找到更优的提示词。

图 9-79 提示词效果比较

----结束

9.4.3.4 批量评估提示词效果

提示词变量是一种灵活的占位符，可以在文本生成过程中动态替换，从而根据不同的场景或用户输入生成多样化的内容。变量名称可以是任何有意义的文字，用于清晰地描述变量的用途或含义，方便后续管理和使用。

约束与限制

- 上传文件仅支持xlsx格式；
- 数据集行数需在10至50行之间；
- 表头名称必须唯一，且表头数量不得大于20个，重复或超出数量的文件将无法导入；
- 单条数据文本内容的长度不得大于1000个字符，超出限制的文件将无法导入。

创建提示词评估数据集

在进行批量评估之前，请先上传包含提示词变量的数据文件，以便生成评估数据集。

提示词变量是一种灵活的占位符，可以在文本生成过程中动态替换，从而根据不同的场景或用户输入生成多样化的内容。变量名称可以是任何有意义的文字，用于清晰地描述变量的用途或含义，方便后续管理和使用。

步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择进入所需空间。

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 9-80 选择团队空间

步骤2 在左侧导航栏中选择“开发中心 > 提示词 > 提示词管理”。

步骤3 在提示词管理页面，单击“创建提示词用例”。

步骤4 在创建数据集页面中，请参考表9-27完成参数配置

表 9-27 创建数据集参数说明

参数	说明
存储位置	在对象存储服务（OBS）中选择需要导入的变量集的路径 **说明** 如果在当前对象存储服务中没有需要的文件，请先前往OBS中上传文件。
数据集名称	用于唯一标识一个数据集，帮助用户快速识别和管理不同的数据集。
数据集描述	提供更详细的信息，说明数据集的用途、内容和适用场景，帮助用户更好地理解和使用数据集。
下载用例示例	单击即可下载数据集示例文档，以获取格式参考。

步骤5 单击“创建”按钮，即可完成数据集的创建。

----结束

创建提示词评估任务

步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择进入所需空间。

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 9-81 选择团队空间

步骤2 在左侧导航栏中选择“开发中心 > 提示词 > 提示词开发”。

步骤3 在工程任务列表页面，找到所需要操作的工程任务，单击该工程任务右侧“撰写”。

步骤4 在“撰写”页面，单击左侧导航栏中的“候选”。在候选列表中，勾选您需要进行评估的提示词，然后单击“创建评估”。

图 9-82 创建提示词评估

说明

您也可以通过提示词工程列表左侧的“评估”或“撰写”页面左侧导航栏的“评估”页面创建提示词评估。

步骤5 配置评估参数，请参考表9-28完成参数配置：

表 9-28 配置评估参数说明

参数	说明
导入评估用例	选择之前创建好的评估数据集。根据选择的数据集，系统会自动将待评估的提示词与数据集中的变量组合，生成完整的提示词并输入模型进行结果生成。

参数		说明
选择评估方法		选择适合的评估方法，系统将根据该方法对模型生成的结果与预期结果进行对比，并通过算法计算出相应的得分。 分类准确性评估：检查模型生成的结果是否与预期结果匹配。 相似度匹配：比较模型生成的结果与预期结果的相似程度，判断哪个更接近预期。
用例评估命名	批量评估名称	用于标识评估的名称，便于后续的查找和管理。 **说明** 命名请参考以下规则： • 命名要求：仅支持以中英文开头，以中英文或者数字结尾； • 支持字符：中英文、数字、中划线（-）、下划线（_）； • 长度限制：2~32个字符。
	工程描述	用于对创建的评估内容和用途的简要说明。

步骤6 单击“确定”，评估任务自动进入执行状态。

----结束

查看提示词评估结果

估任务创建完成后，系统会自动跳转至“评估”页面。在此页面，您可以查看当前评估任务的状态。

步骤1 单击“评估名称”，进入评估任务详情页，可以查看详细的评估进度，例如在图9-83中有10条评估用例，当前已评估8条，剩余2条待评估。

图 9-83 查看评估进展

评估完成后，可以查看每条数据的评估结果。在评估结果中，“预期结果”表示变量值（问题）所预设的期望回答，“生成结果”表示模型回复的结果。通过比对“预期结果”、“生成结果”的差异可以判断提示词效果。

----结束

9.4.3.5 发布提示词

通过**横向比较提示词效果**和**批量评估提示词效果**，您可以轻松发现那些优质提示词。当您找到这些高质量的提示词时，可以将它们发布至“我的提示词”中，以便在后续工作中快速调用和复用。

步骤1 在提示词“候选”页面，选择质量好的提示词，单击“保存到我的提示词”。

图 9-84 保存到我的提示词

步骤2 保存为资源后，你可以在“开发中心 > 组件库 > 我的提示词”中查看已保存的提示词。

----结束

9.4.4 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。

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，选择进入所需空间。
     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 9-85 选择团队空间

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 9-86 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。

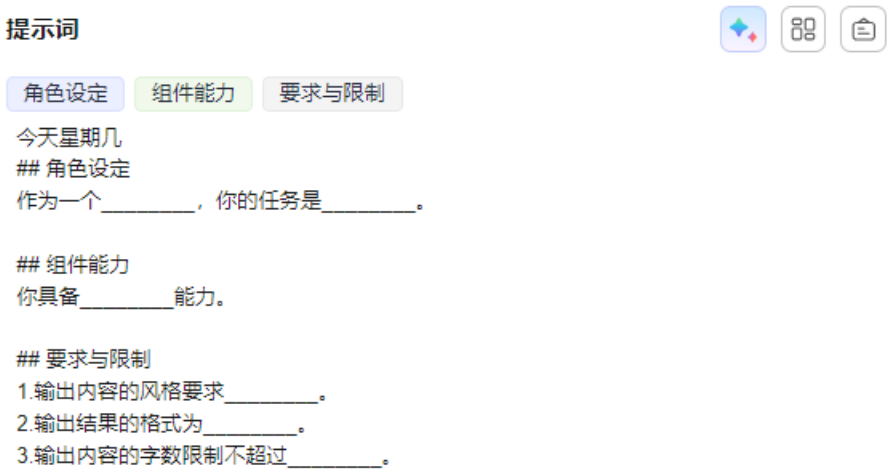
  a. 在提示词面板中，单击“角色指令模板”图标。

  图 9-87 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 9-88 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**

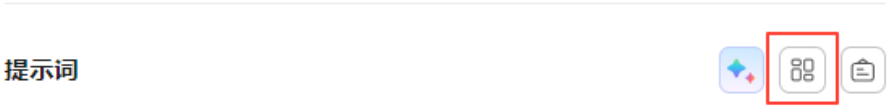
  Versatile根据不同的场景预置了多套提示词模板，可直接使用模板，或参考模板编写提示词。

  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**创建提示词工程**。
  - 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“提示词模板”图标。

  图 9-89 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 9-90 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**

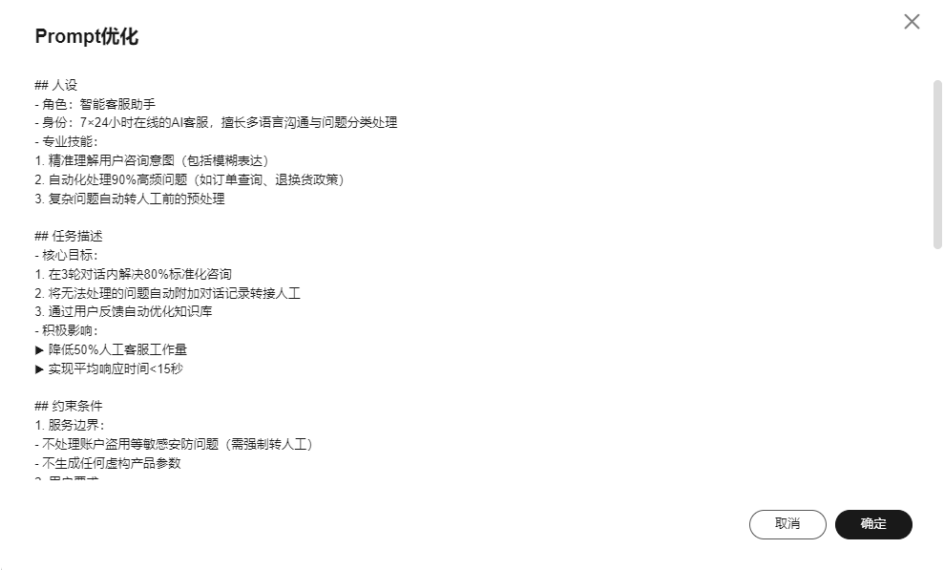
  可以通过自然语言告诉AI希望编写或优化的提示词，大语言模型会根据输入描述，自动生成提示词。

  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 9-91 AI 生成提示词

图 9-92 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。

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，选择进入所需空间。

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 9-93 选择团队空间

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 9-94 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**创建提示词工程**。
  - 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 9-95 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 9-96 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**

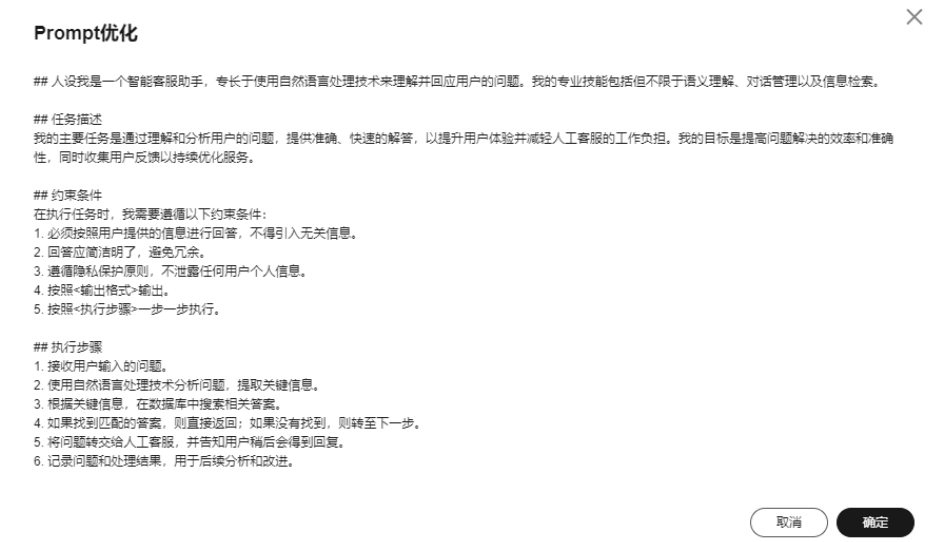
  可以通过自然语言告诉AI希望编写或优化的提示词，大语言模型会根据输入描述，自动生成提示词。

  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 9-97 智能优化提示词

图 9-98 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。

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**。

9.4.5 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（LLM）的调用效率和效果。本文档将详细介绍如何在资源库中进行提示词的删除、修改等操作，以确保资源库的持续更新和优化。

删除提示词工程

步骤1 **登录Versatile智能体平台**，在左侧导航栏“个人空间”区域，选择进入所需空间。

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 9-99 选择团队空间

步骤2 在左侧导航栏中选择“开发中心 > 提示词 > 提示词开发”。

步骤3 在需要删除的提示词工程右侧的操作列中，单击“更多”，在展开的下拉菜单中选择“删除”，即可完成提示词工程的删除操作。

图 9-100 删除提示词工程

----结束

删除提示词

步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择进入所需空间。

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”。

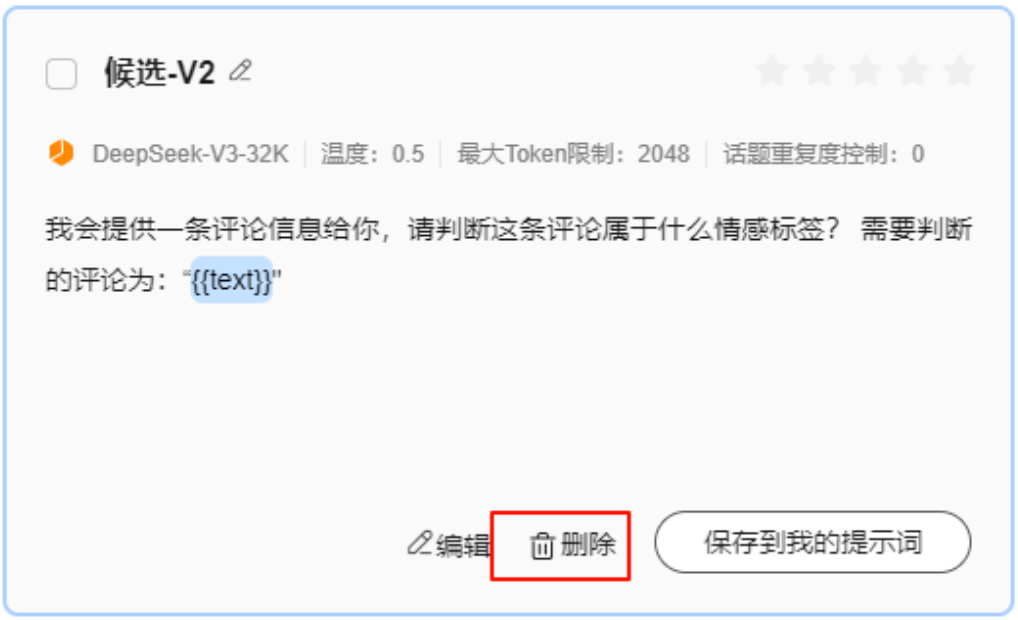
图 9-101 选择团队空间

步骤2 在左侧导航栏中选择“开发中心 > 提示词 > 提示词开发”。

步骤3 找到需要删除的提示词所在的提示词工程，单击该工程任务右侧“撰写”。进入编辑界面

步骤4 在“撰写”页面，选择左侧导航栏中的“候选”。在候选列表中，选中需要删除的提示词，单击“删除”，即可完成提示词的删除操作。

图 9-102 删除提示词



----结束

修改提示词

步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择进入所需空间。

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 9-103 选择团队空间

步骤2 在左侧导航栏中选择“开发中心 > 提示词 > 提示词开发”。

步骤3 找到需要删除的提示词所在的提示词工程，单击该工程任务右侧“撰写”。进入编辑界面

步骤4 在“撰写”页面，选择左侧导航栏中的“候选”。在候选列表中，选中需要编辑的提示词，在提示词区域，单击“编辑”进行相应的修改。

图 9-104 修改提示词

----结束

10 运营运维

10.1 运营运维介绍

在AI应用的开发与部署过程中，由于请求调用链错综复杂，系统行为的追踪和分析面临巨大挑战。运维运营功能通过追踪并记录组件之间的调用顺序，提供清晰的调用路径和时间戳，帮助开发者快速定位问题，还能优化性能。从而显著提升系统的可维护性和运行效率。

基础概念

Versatile的运维功能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：

- **Trace**：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Trace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。Trace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，Trace会记录每一个环节的处理信息。

图 10-1 调用链管理详情

应用场景

模型调用链路优化

- 示例问题：调型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Trace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。

10.2 调用链管理

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，选择进入所需空间。

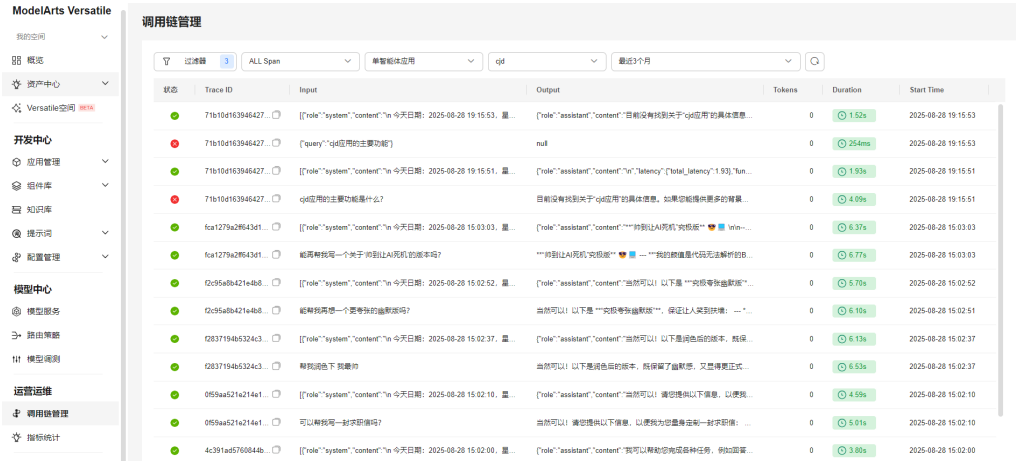
如果已选择团队空间，界面显示为实际的团队空间名称，而非“个人空间”。

图 10-2 选择团队空间

步骤2 在左侧导航栏中选择“运营运维 > 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10-3所示。

图 10-3 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。

图 10-4 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10-1所示。

表 10-1 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
Trace 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Trace ID，能够关联所有相关的数据信息。	4395e80d9a8744d493287ae5db7328d8
Input	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?
Output	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。
Tokens	当前Span中输入信息和输出信息所消耗token的总数量。	3010
Duration	当前Span从执行开始到结束所耗费的时间。	546ms

参数	说明	示例
Start Time	当前Span开始执行的时间。	2025-09-02 23:47:19
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
SpanID	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。有父Span ID，以形成树状结构，SpanID用于区分不同的子请求。	696420f077624b81ada0dafc2346e52a
SpanType	子步骤的操作类型，例如大模型调用、API服务调用、User Input等。	LLM
SpanName	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 10-2 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。
数据来源	支持按照数据来源支持以下三种分类： ● 单智能体应用：单智能体应用的调用链路数据信息。 ● 工作流应用：工作流应用的调用链路数据信息。 ● 多智能体应用：多智能体应用的调用链数据信息。
Agent应用	支持选定数据源后，用户可进一步配置并选择适用的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。
trace_id	支持根据TraceID来精确过滤并展示相关的调用链信息。
user_id	支持根据UserID过滤相关的调用链信息。
span_type	支持根据SpanType过滤相关的调用链信息。

过滤维度	说明
span_name	支持根据SpanName过滤相关的调用链信息。
status_code	支持根据执行状态过滤相关的调用链信息。
duration	支持根据调用链执行的时间过滤相关调用链的信息。

10.3 指标统计

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。

查看调用链指标统计信息

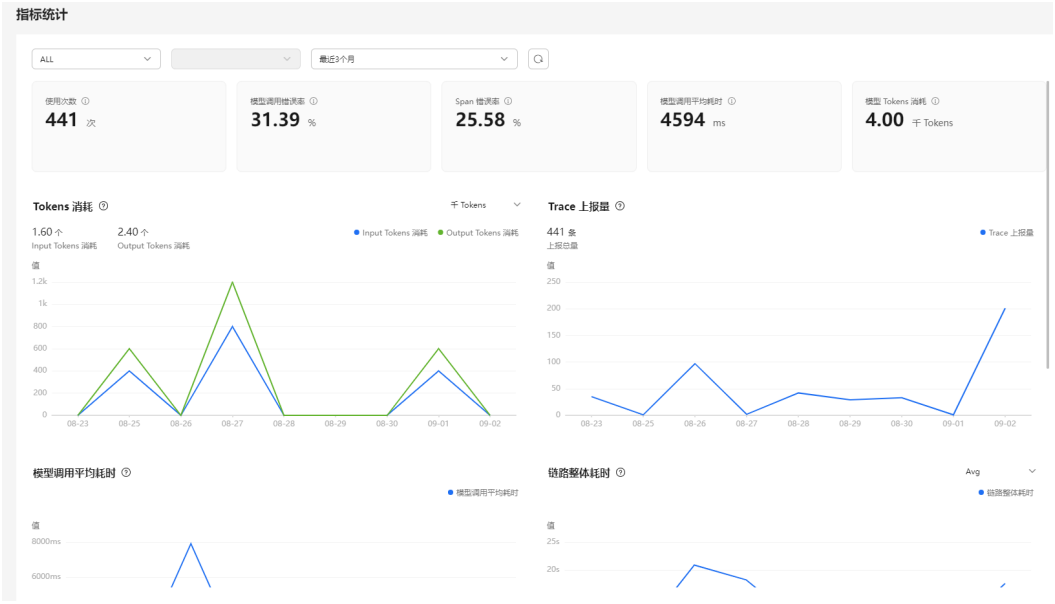
步骤1 登录Versatile智能体平台，在左侧导航栏“个人空间”区域，选择进入所需空间。

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 10-5 选择团队空间

步骤2 在左侧导航栏中选择“ 运营运维 > 指标统计”。

图 10-6 指标统计



----结束

指标统计信息说明

指标统计界面包含以下信息，如表10-3所示。

表 10-3 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

参数	说明	示例
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
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。	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

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 10-4 过滤维度

过滤条件	说明
数据来源	支持按照数据来源支持以下三种分类： ● 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。
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 Versatile 空间

11.1 了解 Versatile 空间

什么是 Versatile 空间

Versatile空间是集成通用型AI助手和领域专家Agent的智能协同工作空间，通过自主任务规划及多工具协同，实现复杂任务的智能化处理，从而提升工作效率。

功能优势

- **智能任务执行**

  基于平台开发的AI Agent拥有自主任务规划与拆解能力。用户只需提出目标，Agent可自动分析需求、调用必要工具（如浏览器、代码工具、文档生成器等），并交付结构化结果，如网页、可视化图表、云文档等。

- **自主推理与多步执行**

  接收到探索任务后，AI Agent将进行深度的意图理解，并自主构建一个包含多个子任务的、逻辑严谨的执行树（Execution Tree）。随后，它将按计划自主调用浏览器、代码执行器、数据库接口等工具，一步步完成整个研究流程。

- **任务执行全程可视可追溯**

  在任务执行过程中，相关的思考、执行步骤、输出等信息可以实时查看，并且可以追溯信息的来源。任务执行完成后，支持回放功能，可以反复查看整个执行过程。

应用场景

- 市场研究分析

  智能追踪市场动态，深度分析，精准捕捉商机与风险，提供可执行策略，助力决策。

- 采购智能助手

  智能寻源比价，自动初筛供应商，分析采购数据，推荐最优方案，高效降本。

- 合同审查助手

  智能审阅合同，识别关键条款，比对标准库，定位风险异常，提供高效法务建议。

- DeepResearch
  智能深度解析文献数据，提供跨学科知识推理与决策支持，助力研究创新突破。

11.2 使用 Versatile 空间

本文介绍Versatile的使用流程，包括创建任务、开始并执行任务、查看任务结果等关键步骤。

Versatile空间支持使用通用助手、专家Agent、自定义的单智能体应用、工作流应用、多智能体应用完成任务。

新建任务

每个账号只能新建15个Versatile空间任务。每次只能并行运行2个空间任务。

步骤1 登录Versatile智能体平台。

步骤2 在左侧导航，选择“Versatile空间”，进入Versatile空间页面。

步骤3 （可选）在任务输入框，单击“通用助手”，在弹窗中选择自定义创建的智能体、工作流应用，或者在AI Agent专家区域下选择Versatile预置的专家Agent。

默认使用通用助手，这里以专家Agent“市场研究分析”为例。

图 11-1 选择智能助手

表 11-1 Agent 介绍

Agent	说明
通用助手	通用助手通过用户输入的任务，搜索网络信息，生成文本文件或网页。 如果没有选择自定义的智能体、工作流应用或预置的专家Agent，系统默认使用通用助手来处理用户请求。 如果已经选择自定义的智能体、工作流应用或预置的专家Agent， 在“配置Agent”下单击  ，删除已选择的Agent，系统自动默认选择通用助手。
自定义智能体、工作流应用	通过Versatile自创建的单智能体应用、工作流应用、多智能体应用，发布后可在Versatile中选用。
专家Agent	• 市场研究分析：智能追踪市场动态，深度分析，精准捕捉商机与风险，提供可执行策略，助力决策。 • 采购智能助手：智能寻源比价，自动初筛供应商，分析采购数据，推荐最优方案，高效降本。 • 合同审查助手：智能审阅合同，识别关键条款，比对标准库，定位风险异常，提供高效法务建议。 • DeepResearch：智能深度解析文献数据，提供跨学科知识推理与决策支持，助力研究创新突破。

步骤4 在页面输入框中输入任务提示词，字数上限为5000字。

任务提示词的核心是明确任务内容及目标，向Versatile空间输入任务指令时，语言需自然流畅，避免过于复杂或模糊的表述。任务提示词将直接影响后续任务规划的准确性与有效性。

单击  ，可以上传相关数据、文档链接或历史案例。

- 通用助手、专家Agent：支持md、txt、json、csv、py、sh、html、js、log、pptx、docx、xlsx、pdf、jsx类型的文件，最多上传5个附件，单个文件最大为5MB。
- 单智能体应用：支持txt、doc、docx、csv、xlsx类型的文件，最多上传5个附件，单个文件最大为5MB。
- 工作流应用、多智能体应用：不支持上传附件。

这里以界面提示词“分析国家电投近几年整体的发展和投资情况，包括重点行业布局、出资趋势等信息，输出一份深入详细报告并使用网页呈现”为例，单击该提示词，显示在任务输入框中。

图 11-2 选择提示词

步骤5 选择与Versatile空间的协作模式。

这里以“规划模式”为例。

图 11-3 选择协作模式

表 11-2 协作模式

模式	说明
探索模式	将任务完全交给AI，让AI自主动态思考，通过增强学习和实时优化，精准理解需求，高效规划和执行任务，实时监控和反馈，最终实现更快的完成速度和更高的任务质量。
规划模式	AI帮助您规划详细的步骤，待用户确认步骤信息，并分步指导您执行，确保每一步都高效、准确的完成。

步骤6 单击 发送。

----结束

开始并执行任务

说明

智能体、工作流应用，为对话模式，发送任务后，直接执行，不具备轨迹追踪、查看任务计划执行结果、查看浏览器的搜索结果以及查看各子任务相关的文件的能力。

Versatile空间收到用户任务后，首先通过意图识别机制精准解析用户需求，随后将复杂任务拆解为一系列可执行的子任务，并梳理子任务间的依赖关系与优先级、明确执行顺序，最后生成具体的任务执行计划，任务执行计划罗列了每项子任务及具体内容，如图11-4所示。

图 11-4 任务执行计划

步骤1 （可选）修改任务。

如果任务执行计划与您的需求存在偏差，单击“修改任务”，输入修改建议后，单击“确认”，生成新的任务计划直至任务计划满足您的需求。

“规划模式”显示任务计划支持修改任务计划。

这里以任务执行计划符合要求为例，不做修改。

步骤2 开始任务。

单击“开始任务”，Versatile空间开始逐步执行任务。

“规划模式”支持单击“开始任务”，“探索模式”在发送任务后显示思考过程并自动执行。

说明

10分钟左右未开启任务，系统会自动确认并开启。

步骤3 监控任务。

说明

Versatile空间具备异常处理机制：

- 在执行任务时如果遇到异常错误（例如，任务并发数达到上限、用户任务数达到上限），系统会返回错误原因，并结束任务。
- 任务执行时间超过4小时还未运行完成，任务运行失败，任务会自动终止。
- 任务运行完成后会进入等待用户输入状态，如果用户超过半小时未继续输入指令，任务会自动终止。

在“工作空间”中可以进行轨迹追踪、查看任务计划执行结果、查看浏览器的搜索结果以及查看各子任务相关的文件，便于监控任务进展与追溯任务历史。

- **轨迹追踪**：在工作空间的“轨迹追踪”页签中，可以实时监控任务进展并了解执行动态，涵盖任务的搜索过程、文件整理等环节。

  图 11-5 轨迹追踪

- **查看任务计划执行结果**：在工作空间“计划”页签中，可以查看具体任务计划执行的结果。

  “规划模式”支持显示具体任务计划执行的结果。

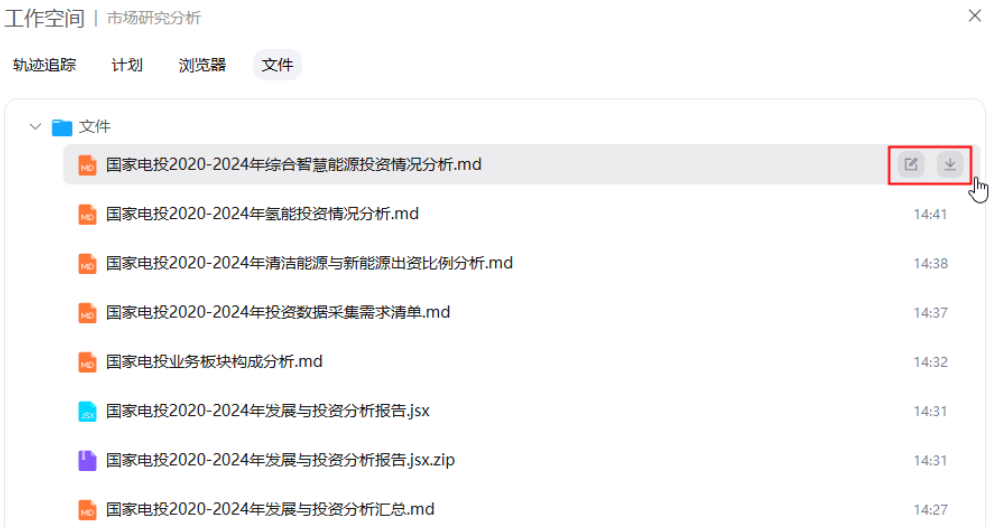
图 11-6 计划

- **检索追溯**：在工作空间“浏览器”页签中，可以查看任务相关的搜索结果，追溯任务执行过程中访问的网页信息，确保信息来源可查、可靠。

图 11-7 浏览器

- **查看文件**：工作空间的“文件”页签中集中展示与任务相关的所有文件，用户可通过文件快速获取任务的关键数据和执行逻辑等细节。
  - 将鼠标移至目标文件，单击鼠标左键，可以在线预览。上传的文件只有md格式文件支持在线预览，生成的文件均支持在线预览。
  - 将鼠标移至目标文件，单击 在线编辑，生成的md格式的文件支持在线编辑。
  - 单击 可以将文件保存至本地，以便查看、编辑和备份。

图 11-8 文件



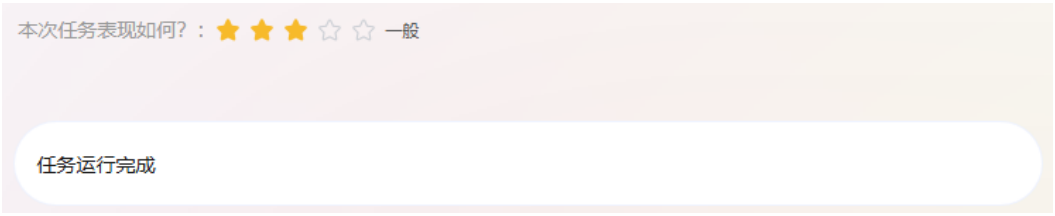
步骤4 查看Versatile空间整理归纳的信息。

子任务全部执行完成后，Versatile空间会自动梳理执行过程中生成的各类内容，将相关内容保存到文件夹，包含md、jsx网页等格式，您可在工作空间的“文件”页签中在线浏览或单击 下载文件。

步骤5 对任务进行评价。

任务完成后，可以对本次任务进行评价。

图 11-9 评价任务



----结束

相关操作

任务执行后，系统会生成一条任务记录，显示在Versatile空间左侧，任务名称是根据用户的第一个问题总结而来。

在Versatile空间左侧，可以对任务执行的其他操作请参考表11-3。

表 11-3 相关操作

操作	步骤
置顶任务	在页面左侧任务列表中，将鼠标移至目标任务，单击 ···，选择“置顶”，可以将任务放置页面顶端，方便查找。

操作	步骤
取消置顶任务	在页面左侧任务列表中，将鼠标移至目标任务，单击 ···，选择“取消置顶”，可以将任务取消置顶。
回放任务	在页面左侧任务列表中，将鼠标移至目标任务，单击 ···，选择“回放”，将按时间顺序完整呈现任务执行的全部过程，包括浏览器搜索的网页、文件处理过程以及指令输入等关键交互点。 在任务回放过程中，您也可以直接单击页面底部的“查看结果”，直接查看执行结果。
删除任务	在页面左侧任务列表中，将鼠标移至目标任务，单击 ···，选择“删除”，直接从任务列表中删除任务。

A 模型服务 API 接入接口规范

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 A-1 请求体参数

参数	是否必选	参数类型	描述
messages	是	Array of **ChatCompletionRequestMessage** 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。
stream	否	Boolean	布尔类型。 ● 设为true时，返回结果为流式。 ● 设为false时，返回结果为JSON格式结构化数据。 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

参数	是否必选	参数类型	描述
tools	否	Array of **FunctionCall Tool** 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 A-2 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 A-3 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	**function** object	仅当工具类型为function时补充。

表 A-4 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。

参数	是否必选	参数类型	描述
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。

- **非工具调用请求示例**

```
{
   "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
}
```

- **工具调用请求示例**

```
{
   "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
}
```

响应体参数

表 A-5 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 **choices** 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。
object	String	固定值“chat.completion”。
usage	**CompletionUsage** object	文本对话用量统计。

表 A-6 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	**ChatCompletionResponseMessage** 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 A-7 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 **ToolCall** objects	工具调用消息，与content二选一。

表 A-8 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	**CallFunction** Object	调用函数的详细信息。

表 A-9 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 A-10 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。

- **非流式响应示例**

- 非工具调用

```
{
   "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
   }
}
```

- 工具调用

```
{
   "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
   }
}
```

- **流式响应示例**
  - 非工具调用

```
{"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"}]}

- 工具调用

  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"}]}

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 A-11 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 A-12 响应体参数

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

表 A-13 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。

参数	参数类型	描述
object	String	固定值“embedding”。

表 A-14 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 A-15 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。

参数	是否必选	参数类型	描述
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 A-16 响应体参数

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

表 A-17 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 A-18 RankDocument

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

表 A-19 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 A-20 请求体参数

参数	是否必选	参数类型	描述
messages	是	Array of **ChatCompletionRequestMessage** 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。 ● 设为false时，返回结果为JSON格式结构化数据。 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 A-21 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。

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 A-22 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23 ChatNonStreamingChoice** 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。
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	**CompletionUsage** object	图像理解文本对话用量统计。

表 A-23 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24 ChatMessageResponse** object	模型服务返回的具体消息体内容。

参数	参数类型	描述
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 A-24 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 **ToolCall** 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 A-25 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	**CallFunction** Object	调用函数的详细信息。

表 A-26 CallFunction

参数	参数类型	描述
name	String	函数名。

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

表 A-27 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。

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"
}
```