

智能体平台

产品介绍

文档版本

01

发布日期

2025-09-16



版权所有 © 华为云计算技术有限公司 2025。保留一切权利。

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

商标声明



HUAWEI和其他华为商标均为华为技术有限公司的商标。

本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

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

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

华为云计算技术有限公司

地址：贵州省贵安新区黔中大道交兴功路华为云数据中心 邮编：550029

网址：<https://www.huaweicloud.com/>

目 录

- 1 什么是智能体平台? 1
- 2 产品优势.....4
- 3 应用场景.....6
- 4 产品功能.....11
 - 4.1 资产中心..... 11
 - 4.2 Versatile 空间..... 11
 - 4.3 开发中心..... 12
 - 4.4 模型中心..... 18
 - 4.5 运营运维..... 20
 - 4.6 空间管理..... 20
- 5 计费说明.....21
- 6 安全..... 22
 - 6.1 责任共担..... 22
 - 6.2 身份认证与访问控制..... 23
 - 6.3 数据保护技术.....24
- 7 权限管理.....25
- 8 约束与限制.....28
- 9 与其他服务的关系..... 30
- 10 基本概念..... 31
- 11 修订记录..... 33

1 什么是智能体平台？

什么是智能体平台？

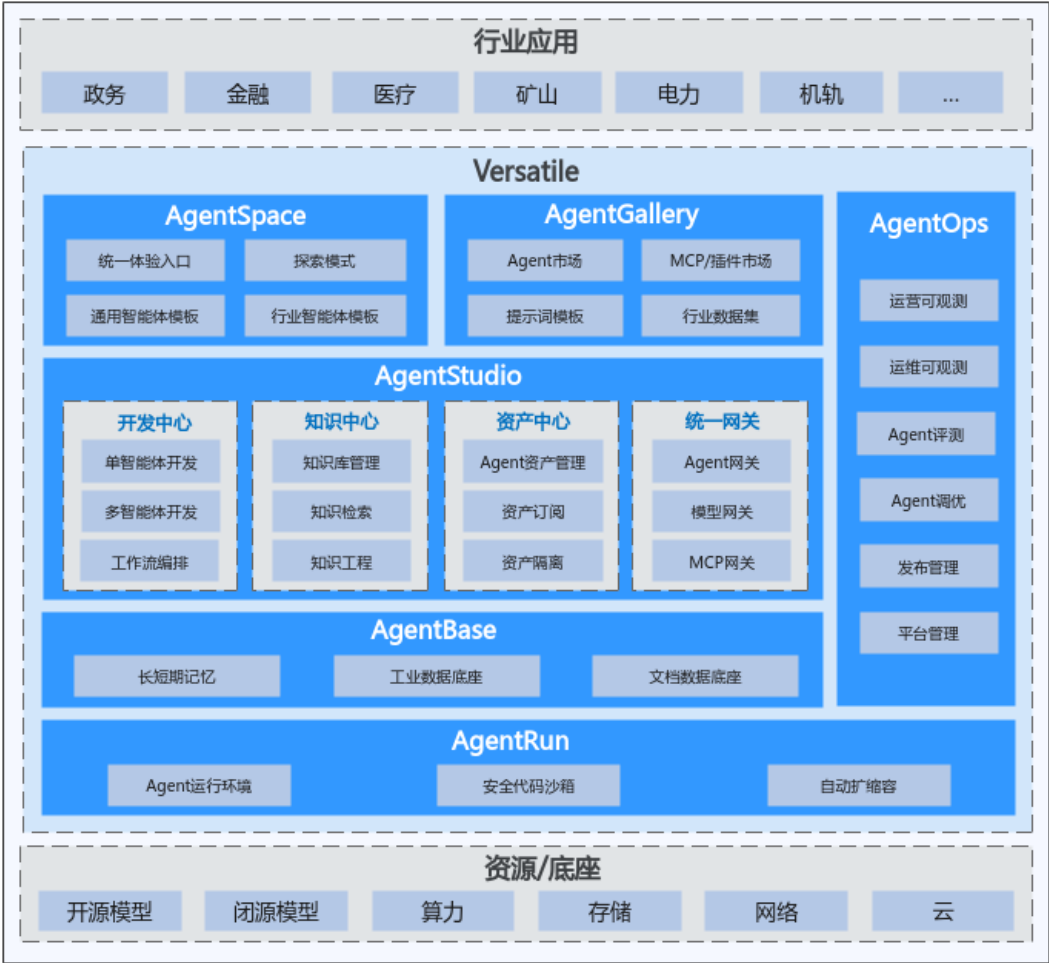
智能体平台（Versatile Agent Platform, Versatile）是一个一站式企业级智能体构建平台，无论是研发技术人员还是业务人员，都可通过简易的界面化操作，快速开发应用。

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。

产品架构

Versatile的产品架构如[图1-1](#)所示。

图 1-1 产品架构



Versatile 提供了一站式 AI 应用搭建能力，平台提供多种能力组合，支持构建不同复杂度的 Agent 开发场景。

- **AgentSpace**：即 Versatile 空间，为用户提供统一 Agent 入口，Agent 统一交互入口，人与 AI 协同的智能入口，统一的 Agent 指挥中心。
- **AgentGallery**：即资产中心，提供 Agent/MCP/插件等接入能力；提供全套工具链及托管环境，在用户开发应用时可选择各类资产，实现动态集成。
- **AgentStudio**：即 Agent 的开发工具，定位为 AI 应用工程，包含开发中心、知识中心、资产中心和统一网关，提供极低门槛，极简高效的开发方式，打造全开放、强运营的 AI 应用使能平台。
- **AgentBase**：为 Agent 提供极致 AI 亲和的文档数据底座和工业数据底座；同时提供数据建模、知识管理、上下文记忆功能。
- **AgentOps**：承载 Agent 从“构成”到“养成”使命；为 Agent 提供完整的追踪、评测、测试、部署和监控功能。
- **AgentRun**：为 Agent 提供安全可靠运行环境，实现极致伸缩和资源高效。支持三方 Agent 应用接入和托管；虚拟机级资源隔离；实例基于微虚拟机（microVM）实现 <1ms 冷启动。

访问方式

如果用户已注册公有云，可直接登录管理控制台，从主页选择"Versatile"。如果未注册，请参见[注册华为云并实名认证](#)。

2 产品优势

极高智能

- **自优化Agentic Model，越用越智能：**基于Agentic Model框架，通过RL反馈框架，构建数据飞轮，实现Agent迭代式优化，使Agent越用越智能。
- **精准的场景意图识别能力：**自动化识别和优化，可自定义无限扩充的场景识别技能。
- **端到端提示词优化能力：**基于客户评测集，自助生成效果最佳的提示词，提升10%~20%的精度效果。

极简开发

- **智能交互：**构建人与AI智能交互模式，实现智能体更加自主。
- **个性化：**构建短、中、长期融合的记忆系统，支撑客户提供个性化Agent。

极稳运行

- **高并发：**极速弹性沙箱（10ms级别启动/伸缩）。
- **低时延：**全链路优化，端到端时延1.5s，部分场景可到500ms。
- **安全：**细粒度安全隔离（Agent单次任务级的授权与安全隔离）、MA Guard内容安全、防攻击、防投毒。
- **高可靠：**支持跨机房主备/双活容灾方案，满足企业级高可用需求，业务SLA大于4个9。

极致开放

- **平台兼容：**兼容主流LangChain、Dify等Agent应用。
- **企业系统API快速转换为MCP工具：**RomaConnect帮助企业快速转换现有API为MCP Server（华为云自身实践）。
- **MCP工具构建：**自研、合作伙伴，华为云，三方云等MCP均可对接。

经验即服务

- **源自华为智能化实践：**将华为1000+Agent实践能力外溢，提供企业典型场景的Agent模板。

- **方案设计：**专业团队深入行业客户业务场景，与客户进行联创，结合业务流程创新。

3 应用场景

面向不同的企业需求，Versatile提供不同的功能服务。

例如，知识问答、智能助手、教育辅导等通用应用场景，可在Versatile体验各大模型推理云服务，并通过可视化画布流程编排进行业务集成。

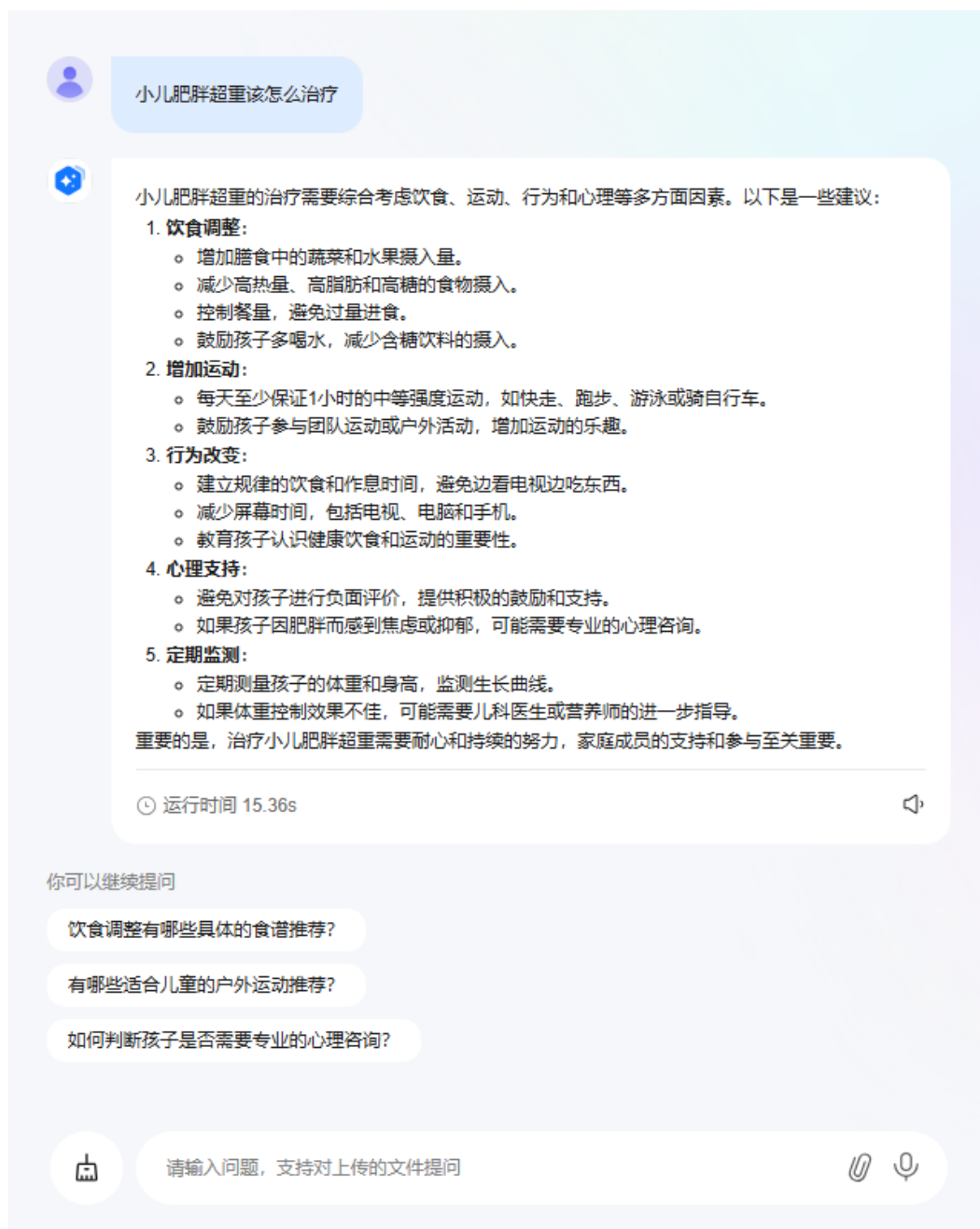
借助Versatile提供的功能服务，用户可以便捷地搭建各类智能体应用。

知识问答

通过对话沟通，快速理解并响应用户的需求，提供高效的解决方案或信息。对于涉及行业和领域的专业知识或技术，用户可以使用平台提供的知识库功能上传相关数据，实现数据的高效管理与存储。在Agent中添加知识库，选择“自动调用”模式，保证Agent的每轮对话回答内容均来自用户提供的专业数据库，确保所提供建议和指导的准确性、专业性。

举例：儿科知识问答Agent，不仅可以迅速响应用户问题，还可为患者提供专业且权威的儿科医学知识。

图 3-1 知识问答



智能助手

理解用户意图并进行动态响应，为用户提供个性化服务，通常需要访问外部资源（如实时数据、API服务等），或者需要进行复杂的逻辑处理。在单一模型的基础上添加技能（工具、工作流），对Agent进行能力拓展。

举例：旅游助手Agent，根据实时天气及火车票查询结果规划最优旅行方案，为用户提供全面的旅游辅助服务。

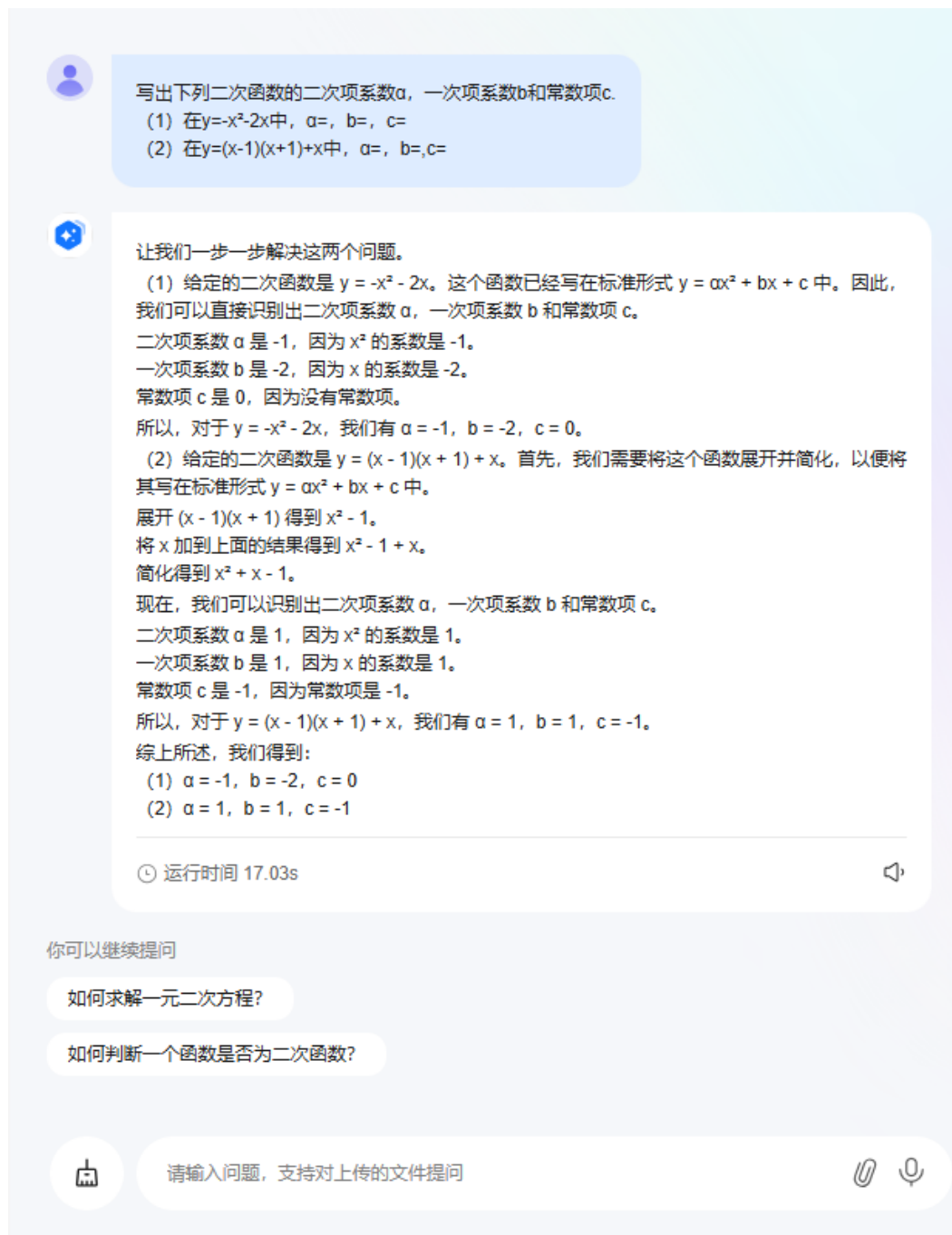
图 3-2 智能助手



教育辅导

个性化学习答疑、知识点讲解。结合学科知识库和大模型推理能力，学生通过对话提问（如数学题解析），模型实时生成解题步骤并反馈。零代码快速覆盖多学科需求，灵活适应不同提问方式。

图 3-3 教育辅导



其他场景

- 金融风控
 - 场景介绍：贷款审批、反欺诈审核。
 - workflow设计：通过节点串联实现多步骤审核（如调用征信接口→大模型分析收入证明→判断节点触发人工复核）。
 - 优势：流程透明可控，分支条件清晰（如不同额度触发不同审批路径），避免大模型幻觉干扰关键决策。

- 供应链管理
 - 场景介绍：库存预警与自动补货、物流异常处理。
 - workflow设计：通过代码节点对接ERP系统，判断节点识别库存阈值，触发补货插件或邮件通知节点。
 - 优势：复杂规则（如供应商优先级、物流延迟策略）可通过分支节点精准实现，稳定性高于纯模型驱动。
- 政务审批
 - 场景介绍：企业资质申报、惠民补贴发放。
 - workflow设计：串联意图识别节点（分类用户需求）、提问器节点（补全材料）、代码节点（验证数据合规性），最终由消息节点反馈结果。
 - 优势：标准化流程减少人为错误，分支节点可覆盖政策差异化要求（如不同地区补贴标准）。

4 产品功能

4.1 资产中心

Versatile资产中心提供应用、MCP、插件、提示词等AI资产的共享，包括系统预置资产、开发者上传资产等，加速AI应用的开发以及公共资产的沉淀。

表 4-1 Versatile 资产中心

功能	说明
应用广场	提供开箱即用的智能体和工作流，涵盖数据处理、通用、医疗、金融、政务五大场景，可参考模板快速定制开发，降低开发门槛。
MCP广场	提供MCP预置服务，满足不同场景预置开发，一键集成调用。
插件广场	提供预置插件，拓宽应用开发边界。
提示词模板	预置提示词模板，可直接在应用开发中一键引用，提升应用实际效果。

4.2 Versatile 空间

Versatile空间集成通用型AI助手以及各专家Agent，打造企业级智能协作统一入口，通过自主任务规划及多工具协同等，帮助用户完成复杂任务。

表 4-2 Versatile 空间

功能	说明
统一体验入口	空间内应用统一体验入口，支持一键选择空间内应用体验效果，支持对话任务创建。
预置AI Agent专家	预置市场研究分析、采购智能助手、合同审查助手、DeepResearch等智能体，可直接提问体验Agent能力。

4.3 开发中心

Versatile支持单智能体、工作流和多智能体三种应用开发方式，平台集成盘古大模型及DeepSeek等第三方模型，提供角色设定、插件扩展、工作流编排等功能，支持知识库管理、RAG检索和智能提示词优化，确保交互精准可靠。同时支持通过API、网页多渠道发布应用，助力开发者高效打造专业级智能体应用。

单智能体应用

表 4-3 Versatile Agent

功能	说明
预置Agent模板	平台预置开箱即用的应用及工作流模板，用户参考模板快速定制开发。
插件	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。
预览调试	- 使Agent开发全程可视化，支持用户整体或分节点查看运行日志，包含输入输出、tokens、耗时等关键信息。 - 支持调用链、时序图两种图览模式，高效识别执行关系和结果。 - 提供历史对话记录的查看与按时间筛选。
语音和安全	提供一句话识别、语音播报、关键字内容检查和安全护栏能力。

工作流应用

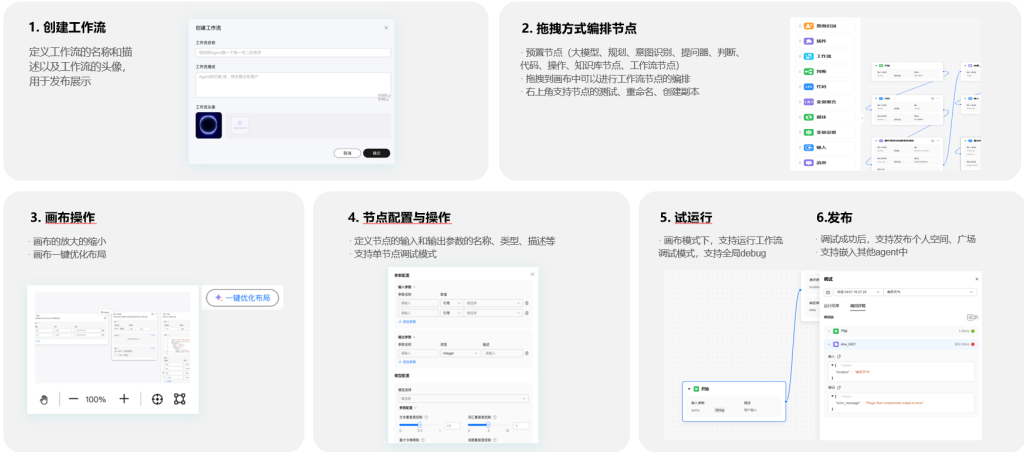
提供灵活的画布式节点编排，让Agent能够准确解决复杂的业务场景问题。

表 4-4 Versatile 编排能力

功能	说明
编排能力	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。 提供灵活的画布式节点编排，让Agent能够准确解决复杂的业务场景问题。
拖拽方式编排节点	平台预置节点，支持用户从工具栏拖拽节点到画布中进行工作流节点的编排。
画布操作与试运行	支持画布一键优化，支持运行工作流调试模式，支持全局Debug。
模型选择	平台提供盘古大模型，支持多款第三方模型，DeepSeek类第三方深度思考模型在Agent开发平台已完成适配。

功能	说明
语音和安全	提供一句话识别、语音播报、关键字内容检查和安全护栏能力。

图 4-1 workflows编排流程



工作流节点

表 4-5 workflows节点能力说明

节点名称	节点说明	节点配置
开始	流程开始节点。	输入参数（系统预置+用户自定义）。
大模型	大模型节点，可在节点中配置已部署的模型，旨在应用编排中使用大模型能力。	输入输出参数、模型选择、模型参数设置、多轮会话、提示词配置。
代码	支持安全代码沙箱以及调用 FunctionGraph函数（FunctionGraph函数支持选择 Python3.9、Node.js14.18），提供代码执行能力。	输入输出参数、Python代码编辑器。
意图识别	提供基于大模型的意图分类功能。	输入参数、意图分类配置、模型选择、提示词配置（支持意图数量>=2）。普通模式意图数量20个，高级模式意图数量200个。
判断	条件分支节点，支持多路条件判断。	输入输出参数。 条件分支配置（支持条件路数>=2）。
插件	插件调用节点，可以选择系统注册的插件，并进行工具插件调用。	插件选择、插件输入参数。

节点名称	节点说明	节点配置
提问者	提供通过多轮交互的方式提取必要的参数信息。	输入输出参数、问题配置、模型选择、提示词配置。
结束	流程结束节点，用于流程执行信息输出。	输出参数、信息输出格式。
消息	只负责流程执行过程中的信息呈现，输出结果对大模型对话上下文没有任何影响。	输出参数配置，消息展示信息配置。
输入	支持中间过程的信息输入，支持的参数类型同开始节点。	输入参数。
变量聚合	对多个分支的输出进行聚合处理。	分组参数配置。
知识检索	可以根据输入参数从指定知识库内召回匹配的信息。	输入输出参数、知识库、检索策略、仅查询指定标签的文档。
工作流	实现工作流引用子工作流的嵌套调用，降低工作量复杂度。	输入输出参数、子工作流。
循环	通过设定循环次数和逻辑，重复执行一系列任务。	循环设置、循环数组、中间变量、输出参数。
MCP服务	客户自定义接入的MCP服务，可以选择指定的工具进行调用。	输入输出参数、工具。

多智能体应用

随着任务复杂度增加，单Agent面临上下文过长、工具繁多难以选择，从而导致性能下降的问题。多Agent协作通过任务动态分解与分配，任务交接和协同工作等技术克服这一问题，支持用户复杂意图场景的准确识别，能够完成多工作流的嵌套调用，以实现更复杂场景的智能体协同，适用于解决企业内流程冗长、涉及多个专业领域的复杂任务。

表 4-6 Versatile 多智能体应用

功能	说明
全局意图控制	支持全局意图跳转，控制器多工作流间信息共享，有助于实现子工作流之间高效交互与协作。
任务智能调度	任务智能调度自动为每个任务使用LLM-based选择器分配最适合的子工作流。

智能补齐

将简单的配置工作交给AI完成，秒级生成高质量的prompt指令和开场白等文案，秒级一键匹配最适合的插件，降低应用开发难度，节省用户配置时间。

表 4-7 Versatile 智能补齐

功能	说明
prompt智能优化	基于智能prompt模板，快速扩写生成一个高质量的标准prompt。
插件智能添加	根据相似性算法，快速匹配已有内容。
智能对话内容配置	基于应用描述一键生成开场白、推荐问、追问。
智能添加名字	基于输入的名字，智能生成名字。
智能添加描述	基于输入的描述，智能生成一段描述。

MCP 服务

每个AI应用都需要通过单独封装API对接相同的外部系统，存在碎片化集成的问题，通过MCP可实现交互标准化接口。

表 4-8 Versatile MCP 服务

功能	说明
预置MCP服务广场	提供MCP预置服务，满足不同场景预置开发，一键集成调用。
MCP管理	提供开发者自定义注册的MCP服务。
MCP调用	一键在应用中配置MCP服务，实现MCP服务的调用兼容。

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 4-9 Versatile 知识库

功能	说明
知识库列表管理	支持展示知识库名称、知识库ID、知识库描述、更新时间，支持知识库列表操作（查看、删除、编辑、启用、停用、命中测试）。
创建知识库	支持填写知识库名称、描述，选择向量模型、精排模型、选择解析策略及拆分策略；支持选择新建知识库和引用知识（链接已有知识库）。
上传知识库文档	支持文档列表展示，支持导入新文档，支持知识库文档列表操作（查看、下载、删除）。

功能	说明
文档拆分入库	- 多模态RAG，支持文本、图片、表格等10+格式的内容解析入库、检索、多模态结果返回。 - 支持修改知识库配置的解析策略，识别文档版式内容，节点元素，通过TableRAG提升表格问数能力。 - 通过段落拆分策略和段落拆分模型将现有文档进行结构化处理。 - 将文档通过盘古搜索语义模型进行向量化后导入到向量数据库（CSS）中。
检索策略	搜索模块语义检索、关键词检索、混合检索、FAQ检索、相关度阈值、topk召回数量、FAQ直出阈值匹配最接近的文档或者文本块，并进行相关度排序。
命中测试与答案生成溯源	- 支持测试知识检索成功率。 - 将搜索的结果+原始查询语句通过大模型进行归纳总结后，将答案按结构化的方式呈现给用户。 - 多模态文档理解：融合文本/图片/表格等异构数据生成精准答案。 - 单智能体应用在生成结果中支持显示当前模型生成内容参考的搜索结果片段以及对生成的结果做二次校验确保结果符合客观事实和内容可信。
智能搜索增强	- 多模态意图解析：支持图文混合query的意图分类与图片语义改写。 - 跨模态检索优化：通过图片识别技术提升搜索结果相关性。

提示词工程

平台提供prompt模板与开发工具，任何人都可以开发出高质量的prompt指令，基于自适应模板生成高质量提示词提升大模型输出效果。

表 4-10 Versatile 提示词工程

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，选择最优的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。

功能	说明
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。

对话

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 4-11 Versatile 对话体验

功能	说明
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。
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。

配置管理

配置管理包含意图包和数据源管理。

表 4-12 Versatile 配置管理

功能	说明
意图管理	意图包配置，用于高级意图识别节点意图配置。
数据源管理	支持通过公网接入MYSQL类型数据库，在工作流数据库节点引用，支持数据库相关操作。

发布

提供企业级Agent多渠道发布、升级、环境转移服务。

表 4-13 Versatile 发布功能

功能	说明
发布渠道	支持多个主流发布渠道，包括网页端、API。
版本管理	在多个不同配置版本间进行还原，方便Agent升级与回退。
资产导入导出	支持应用、工作流、插件、知识库的一键关联导出，隔离生产与开发环境，跨环境导入后能自动校验异常配置，确保高可用可靠。

4.4 模型中心

模型中心提供标准化API接口，支持盘古大模型与业界主流模型的接入与管理，提供多种路由策略，实现模型无感切换和灵活调度，支持NLP与多模态理解模型的在线调测及参数配置。

模型服务

Versatile支持对接模型，用户可自由选择通过大模型开发平台或直接对接第三方模型，并提供模型接入管理。

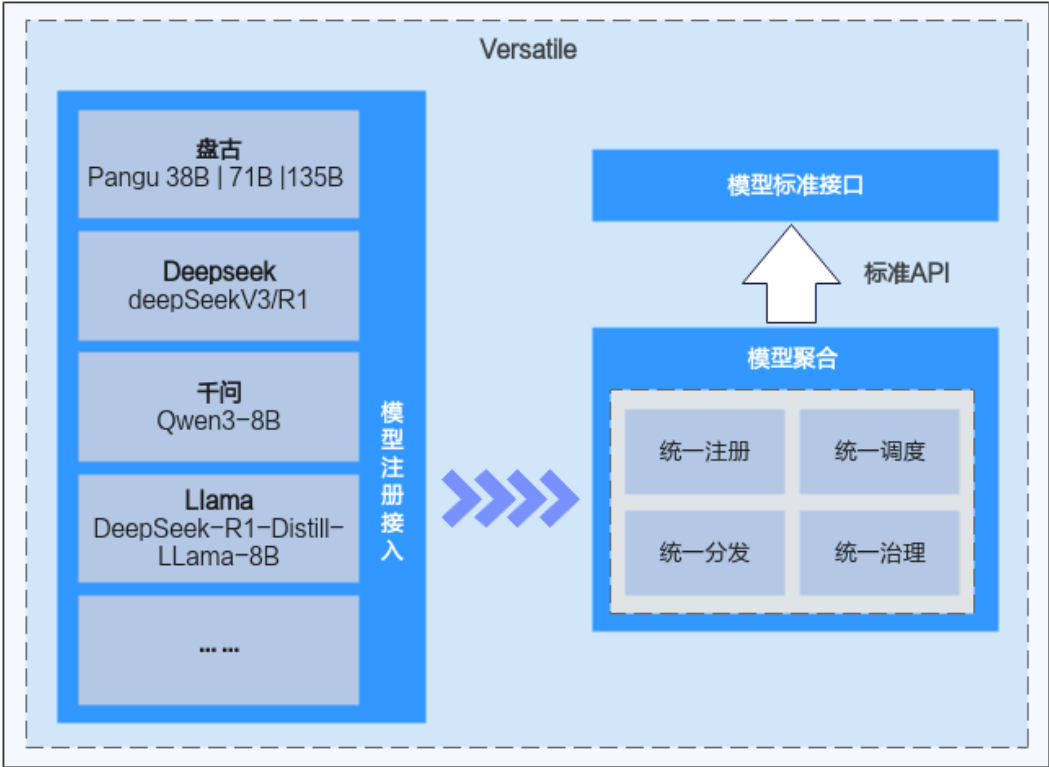
表 4-14 Versatile 模型服务

功能	说明
模型接入	客户可通过大模型开发平台或直接对接第三方模型推理服务两种方式接入。
模型管理	统一管理所有接入的模型，具备鉴权配置、修改和删除能力。

路由策略

多模型路由与开放，兼容业界主流大模型，模型推理API标准化，提供统一访问接口及一站式模型管理，使应用无感模型变化。支持创建和修改路由策略，配置专属模型调测。

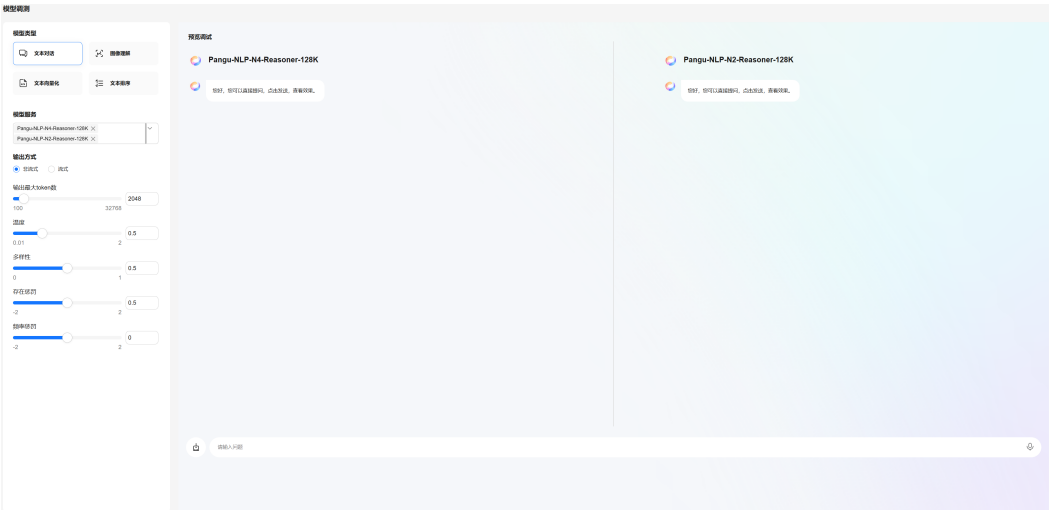
图 4-2 路由策略



模型调测

支持文本对话、图像理解、文本向量化和文本排序四种类型模型的调测，可选择非流式和流式输出模式。调测模型包括输出最大token数、温度、多样性、存在惩罚、频率惩罚参数，通过预览调试查看调试效果。

图 4-3 模型调测



4.5 运营运维

Agent 观测

全维度Agent观测，实现对Agent调用的多维指标监控运营。

表 4-15 Versatile Agent 观测

功能	说明
监控运营	支持多维度运营指标统计，实现对Agent的调用多维指标监控运营，提供运行分析报表页面，报表系统支撑海量用户调用数据和异常数据的统计能力。
全链路可视化	以会话串联用户交互问答过程，以Trace承载应用全链路交互节点，可视化关键内容。
链路追踪	针对异常请求，内部错误，安全攻击等场景，提供详细记录，支撑Agent优化，提升质量。
Agent监控	提供Agent数据分析看板、成本监控、指标定义、Agent追踪能力。

4.6 空间管理

空间管理助力团队高效便捷地协作开发应用，多重权限控制提升企业资产数据安全。

表 4-16 Versatile 工作空间

功能	说明
多空间管理	默认配备独立个人空间，同时可支持加入/创建多人协作空间。
团队协作	空间内资产可以在空间成员之间共享，应用支持有权限的跨空间复制，便于协同开发，使得应用开发灵活、便利、高效。
资产导入导出	支持应用、工作流、插件、知识库的一键关联导出，隔离生产与开发环境，跨环境导入后能自动校验异常配置，确保高可用可靠。
空间隔离	不同空间之间的资源隔离，便于企业内各组织独立管理资源。
角色权限隔离	角色权限细化，区分所有者、管理员、成员角色，保障空间应用开发和资产数据安全。

5 计费说明

通过阅读本文，您可以快速了解Versatile的计费模式、计费项等主要计费信息。

计费模式

Versatile限时免费，前往控制台订阅成功后即可使用，应用开发过程中所需的大模型资源和增值特性服务计费详情可参见[表5-1](#)。

计费项

表 5-1 模型推理和增值特性服务计费说明

服务	计费说明
模型推理	客户可通过大模型开发平台（ ModelArts Studio， Maas ）或直接对接第三方模型推理服务两种方式接入。 通过Maas部署的方式接入时，计费说明请参见 Maas部署计费说明 。
智能体API	限时免费体验。
一句话识别（ 语音交互-语音输入 ）	限时免费体验。
语音合成（ 语音交互-卡片消息朗读 ）	限时免费体验。
文本内容审核（ 安全防护 ）	限时免费体验。

6 安全

6.1 责任共担

华为云秉承“将公司对网络和业务安全性保障的责任置于公司的商业利益之上”。针对层出不穷的云安全挑战和无孔不入的云安全威胁与攻击，华为云在遵从法律法规业界标准的基础上，以安全生态圈为护城河，依托华为独有的软硬件优势，构建面向不同区域和行业的完善云服务安全保障体系。

与传统的本地数据中心相比，云计算的运营方和使用方分离，提供了更好的灵活性和控制力，有效降低了客户的运营负担。正因如此，云的安全性无法由一方完全承担，云安全工作需要华为云与您共同努力，如[图6-1](#)所示。

- **华为云：**无论在任何云服务类别下，华为云都会承担基础设施的安全责任，包括安全性、合规性。该基础设施由华为云提供的物理数据中心（计算、存储、网络等）、虚拟化平台及云服务组成。在PaaS、SaaS场景下，华为云也会基于控制原则承担所提供服务或组件的安全配置、漏洞修复、安全防护和入侵检测等职责。
- **客户：**无论在任何云服务类别下，客户数据资产的所有权和控制权都不会转移。在未经授权的情况，华为云承诺不触碰客户数据，客户的内容数据、身份和权限都需要客户自身看护，这包括确保云上内容的合法合规，使用安全的凭证（如强口令、多因子认证）并妥善管理，同时监控内容安全事件和账号异常行为并及时响应。

图 6-1 华为云安全责任共担模型



云安全责任基于控制权，以可见、可用作前提。在客户上云的过程中，资产（例如设备、硬件、软件、介质、虚拟机、操作系统、数据等）由客户完全控制向客户与华为云共同控制转变，这也就意味着客户需要承担的责任取决于客户所选取的云服务。如图6-1所示，客户可以基于自身的业务需求选择不同的云服务类别（例如IaaS、PaaS、SaaS服务）。不同的云服务类别中，每个组件的控制权不同，这也导致了华为云与客户的责任关系不同。

- 在On-prem场景下，由于客户享有对硬件、软件和数据等资产的全部控制权，因此客户应当对所有组件的安全性负责。
- 在IaaS场景下，客户控制着除基础设施外的所有组件，因此客户需要做好除基础设施外的所有组件的安全工作，例如应用自身的合法合规性、开发设计安全，以及相关组件（如中间件、数据库和操作系统）的漏洞修复、配置安全、安全防护方案等。
- 在PaaS场景下，客户除了对自身部署的应用负责，也要做好自身控制的中间件、数据库、网络控制的安全配置和策略工作。
- 在SaaS场景下，客户对客户内容、账号和权限具有控制权，客户需要做好自身内容的保护以及合法合规、账号和权限的配置和保护等。

6.2 身份认证与访问控制

用户访问Versatile的方式有多种，包括Versatile控制台、API、SDK，无论访问方式封装成何种形式，其本质都是通过Versatile提供的REST风格的API接口进行请求。

Versatile的接口均需要进行认证鉴权，以此判断是否通过身份认证。

通过控制台发出的请求需要通过Token认证鉴权，调用API接口认证鉴权支持Token认证和AK/SK认证两种方式。

6.3 数据保护技术

表 6-1 数据保护技术

数据保护手段	说明
传输中的数据保护	在Versatile中接口调用时，为保证数据传输的安全性，推荐使用HTTPS协议。

7 权限管理

Versatile 工作空间角色

Versatile的用户可以被赋予不同角色，实现对工作空间权限的精细化控制。

表 7-1 工作空间角色定义

模块	内容	所有者	管理员	成员
空间管理	添加成员	√	√	×
	移除成员	√	√	×
	修改角色	√	√	×
	修改空间名称、描述	√	√	×
	删除空间	√	√	×
	转让空间	√（离开前需指定下一位所有者）	√	×
	退出空间	√（离开前需指定下一位所有者）	√	√
应用管理	创建	√	√	√
	查看&调试	√	√	√
	复制	√	√	√
	修改	√	√	√
	发布	√	√	√
	导入&导出	√	√	√
	删除	√	√	√
组件库	创建	√	√	√

模块	内容	所有者	管理员	成员
	导入&导出	√	√	√
知识库	新建知识库	√	√	√
	接入知识库	√	√	√
	编辑知识库	√	√	√
	停用/启用知识库	√	√	√
	命中测试	√	√	√
提示词	创建提示词工程	√	√	√
	撰写提示词	√	√	√
	评估提示词	√	√	√
	编辑提示词工程	√	√	√
	删除提示词工程	√	√	√
	删除提示词数据集	√	√	√
配置管理	创建意图包	√	√	√
	导入&导出意图包	√	√	√
	接入数据源	√	√	√
	编辑数据源	√	√	√
	删除数据源	√	√	√
模型中心	鉴权配置	√	√	√
	新增模型供应商	√	√	√
	修改我的模型	√	√	√
	删除我的模型	√	√	√
	创建路由	√	√	√
	模型调测	√	√	√
	修改路由策略	√	√	√
	删除路由策略	√	√	√
	模型调测	√	√	√

模块	内容	所有者	管理员	成员
运营运维	查看	√	√	√

8 约束与限制

规格限制

表 8-1 规格说明

资源类型	规格	说明
计算资源	所有按需计费、包年/包月、套餐包中的计算资源规格，包括CPU、GPU和NPU。	购买的所有类型的计算资源均不支持跨Region使用。

免费资源限制

Versatile提供以下有限免费资源的服务能力。

表 8-2 约束与限制

模块	服务能力	资源（每租户）
模型及资源	模型推理体验资源	2个模型（R1\V3）分别1M tokens累计
	模型每分钟请求数RPM	不限
	智能体API调用量	累计500次
	三方模型接入	√
知识库	知识库上传数量	不限
	存储容量	1G
	三方知识库对接	√
业务功能	MCP FG托管（NPX/UVX限制）	10个
	Trace上报	存15天

模块	服务能力	资源（每租户）
	模型路由\负载均衡	√
	安全审查	√
	应用导入导出	√
	应用评测	√
	运营统计报表导出	√
企业管理与服务	团队空间数量	1
	成员数量	3
	应用数量	10
	成员权限控制	√
	跨空间复制	×
	应用版本控制	√

9 与其他服务的关系

表 9-1 Versatile 与其他服务的关系

服务名称	Versatile与其他服务的关系	主要交互功能
对象存储服务 OBS	Versatile使用OBS存储文档数据和Agent元数据信息。	• Versatile上传Agent元数据到OBS • 从OBS下载Agent元数据
AI开发平台 ModelArts	Versatile可以使用ModelArts Studio部署的三方模型推理服务进行Agent应用开发。	模型接入
云搜索服务 CSS	Versatile使用CSS加入检索模块，提高模型回复的准确性、解决内容过期问题。	• 创建知识库 • 知识库节点
函数工作流 FunctionGraph	Versatile使用FunctionGraph实现代码级别函数托管，实现函数调用功能。	代码节点创建函数
大模型安全护栏 ModelArts Guard	Versatile使用ModelArts Guard实现大模型级别与Agent应用级别的安全控制。	• 大模型节点，配置护栏 • 工作流配置护栏 • Agent配置护栏

10 基本概念

Agent

Agent，即智能体，能够根据用户指令或预设规则自主执行任务。它通常具备一定的智能和决策能力，可以模拟人类的思考和行为模式，自动完成诸如信息检索、数据分析、任务调度等复杂操作。广泛应用于客服、自动化办公、数据分析等场景，Agent通过学习和优化不断提升自身性能，为用户提供高效、便捷的服务。

Prompt

Prompt，即提示词，是用户输入给AI模型（如大模型、生成式AI等）的指令、问题、描述或上下文信息，是模型理解用户需求、确定输出方向与内容的核心依据，可通过文字、代码片段、格式说明等形式，引导模型生成符合预期的结果（如回答、文本创作、代码编写等）。

插件

插件是一种用于扩展大模型功能的工具集合，让大模型与外部系统进行交互，快速扩展智能体能力，平台提供了丰富的内置插件，开箱即用。

workflow

工作流是指利用大模型的能力，结合自动化任务编排技术，通过可视化或代码方式连接多个AI模型、数据处理模块、逻辑判断节点等，实现复杂任务自动化执行的端到端流程。

节点

节点是构成整个流程的基本功能单元，可承载数据输入输出、任务执行（如调用大模型/插件）、逻辑判断（如条件分支）等单一或特定组合操作，是流程中信息流转和任务推进的关键落点。

变量

用于存储应用运行中需动态调整的参数。

提示词工程

提示词工程是通过设计、优化输入给AI模型的指令（提示词），引导模型更精准理解需求、生成符合预期的输出，以提升模型应用效果的优化手段。

大语言模型

大语言模型（Large Language Model，简称LLM）是基于海量文本数据训练而成、具备强大自然语言理解与生成能力的人工智能模型，其核心通过复杂的深度学习架构（如Transformer）捕捉语言中的语法、语义、逻辑及知识关联，可实现文本生成、问答、翻译、摘要、代码编写等多样化自然语言任务，能理解上下文并生成连贯、符合人类语言习惯的输出，是当前自然语言处理领域的核心技术之一。

记忆

记忆（Memory）是指通过技术手段让大模型在交互过程中“记住”历史信息（如用户过往提问、对话上下文、特定任务参数等），而非每次交互都从零开始理解，从而实现连贯的多轮对话、个性化需求适配（如记住用户偏好）或复杂任务的持续推进（如逐步完成文档创作、问题拆解）的能力，核心是解决大模型“短期上下文有限”和“长期信息留存”的问题，提升应用的交互流畅性与实用性。

Token

Token是指文本被分割后的最小语义单元。它可以是一个单词、一个标点符号、一个子词或一个特殊符号。

向量化模型

向量化模型是将文本数据转换为数值向量的过程。常用于将文本转换为机器可以处理的形式，以便进行各种任务，如文本分类、情感分析、机器翻译等。

向量化检索

向量化检索是一种在高维向量空间中查找相似向量的技术，广泛应用于自然语言处理、图像识别、推荐系统等领域。通过将数据转换为向量表示、建立高效的索引结构和使用相似度度量方法，向量化检索能够快速找到与查询最相似的数据。

多模态模型

多模态模型是指能够处理多种类型数据（如文本、图像、音频等）的机器学习模型。这些模型可以将不同类型的数据进行融合和联合分析，从而实现更全面的理解和更准确的预测。多模态模型的应用非常广泛，例如，在图像识别中，可以将图像和文本信息结合起来，提高图像识别的准确性；在自然语言处理中，可以将文本和语音信息结合起来，提高文本语义理解的准确性。

11 修订记录

发布日期	修订记录
2025-09-16	第一次正式发布。