云计算开源产业联盟
OpenSource Cloud Alliance for industRy,OSCAR

CAICT 中国信通院

互联网智能体发展研究报告

（2026）

云计算开源产业联盟

中国信息通信研究院云计算与数字化研究所

2026 年 7 月

前　言

2026 年，“智能体”首次被写入《政府工作报告》，明确提出打造“智能经济新形态”，促进新一代智能终端和智能体加快推广等战略部署。同年 5 月，国家网信办、国家发展改革委、工业和信息化部联合印发《智能体规范应用与创新发展实施意见》。同时，《互联网信息服务管理办法》（292 号令）修订工作也在稳步推进，进一步明确智能体等新型互联网信息服务的合规要求与监管边界。这一系列顶层制度的设计标志着我国率先构建起智能体全生命周期发展与治理框架。

宏观战略的落地正伴随着互联网主体形态的深刻重构。回顾互联网的发展脉络，从 PC 互联网时代的网站，到移动互联网时代的 App，人类社会正迎来从“信息互联”向“智能互联”的数字跃迁。在这一新形态下，互联网的连接主体发生了根本性变革，从以“人”为核心，全面转向以“智能体”为核心，具备自主感知、决策与执行能力的互联网智能体，正作为全新的联接主体。面对互联网主体形态的深刻重构，智能体的监管思路也随之从传统的“管准入”向“管行动”全面跃迁。监管不再局限于静态的内容合规与运营许可管理，而是将具备执行能力的智能体作为独立的治理对象，构建起分类分级、场景牵引的全生命周期治理框架。通过明确

用户与智能体的决策权限边界、强化行为管控与责任追溯，监管旨在统筹发展与安全，既为“智能互联”时代的创新划定底线，又为打造“智能经济新形态”提供规范有序的法治保障。

本报告旨在整合行业实践与技术研究成果，为互联网智能体的技术创新、应用落地与规范管理提供全面、专业的参考依据，助力智能体产业健康有序发展。因该领域尚处于快速演进阶段，内容难免存在疏漏，恳请各位专家与读者批评指正。

报告编写组联系方式:

李哲 18701508869 | 董晓慧 18843100481

目 录

图 目 录

表 目 录

一、发展背景

随着人工智能与网络技术的深度融合，智能体（Agent）正从单点功能工具演化为可自主联网、跨域协同的新型服务主体。当前全球智能体产业进入规模化应用的关键窗口期，各国同步推进技术创新、生态构建与规则治理；我国依托顶层政策引导与全行业实践探索，智能体落地速度持续加快。与此同时，智能体联网行为带来的身份、权限、安全等风险日益凸显，亟需建立覆盖智能体全生命周期的可信行为规范体系。

（一）产业变革趋势：智能体从单体运行走向联网服务，成为智能经济新型主体

从技术演进来看，智能体形态遵循从单体智能到多智能体协同、再到联网服务的发展路径。中国工程院邬贺铨院士在2025中国互联网大会上提出人工智能四级演进路径，将发展阶段划分为生成式AI、单体AI Agent、多智能体协同（Agentic AI）与智能体互联网（Internet of Agents，IoA）。生成式AI如ChatGPT，仅能依据指令被动生成内容，无法主动执行任务。单体AI Agent在大模型基础上整合记忆、工具调用与自主规划能力，形成任务执行闭环。多智能体协同依托中心化编排机制实现专业化分工，协作处理跨领域任务。到智能体互联网阶段，各类智能体打破平台壁垒，依托统一网络协议实现异构主体互联互通，自主实现全自

动、跨场景的复杂服务。

从需求驱动来看，企业数字化转型正推动智能体从辅助工具向智能经济新型主体转变。随着数字业务流程复杂度持续提升，企业对自动化、智能化服务的需求急剧增长。Gartner 预测，到 2028 年，至少 15%的日常工作决策将由 AI 智能体自主完成[1]。智能体正从传统的“人机交互”工具升级为能够自主完成商业闭环的“智能经济主体”，依托自主感知、任务编排、跨系统联动等核心能力，智能体突破传统人机交互的单一模式，深度嵌入企业全业务链条，构建起端到端的完整商业闭环。在电商、金融、制造等领域，自动比价下单、风险实时扫描、产线动态调度等自动化服务能力已逐渐规模化应用。

（二）国际竞争格局：全球加速布局智能体互联生态，践行技术创新与可控发展

全球科技巨头加快布局智能体互联生态，聚焦协议框架与应用能力持续攻坚。Anthropic 发布 MCP（Model Context Protocol）协议，定义 AI 模型与外部数据源、工具的标准化连接接口。Google 推出 A2A（Agent-to-Agent）协议，支持异构智能体间的相互发现、任务协作与能力调用。OpenAI 发布 Operator 浏览器智能体，可自主完成网页浏览、表单填写、订餐购物等复杂线上任务。微软发布 Azure AI

[1] 数据来源：Gartner. Top Strategic Technology Trends for 2025 [R/OL]. (2024-10)[2026-06-11]. https://www.gartner.com/en/insights/top-tech-trends-2025.

Foundry Agent Service，支持 LangGraph、CrewAI 和 OpenAI Agents SDK 等多框架接入。英伟达发布开源 NVIDIA Agent Toolkit，联合众多软件厂商推进落地，加速向智能体基础设施服务商转型。

各国同步推进智能体领域规则体系建设，坚持技术创新与安全规范并重。欧盟在 2024 年正式落地《人工智能法案》，构建基于风险分级的综合性 AI 监管框架，将 AI 系统划分为禁止、高风险、有限风险和最低风险四个等级。美国国家标准与技术研究院（NIST）在 2026 年启动 AI 智能体标准计划，围绕智能体的身份鉴别、授权、审计、不可否认性及提示注入防控措施等方面公开征求意见。新加坡资讯通信媒体发展局（IMDA）在 2026 年发布《智能体人工智能治理示范框架》，围绕权限边界管控、人工审批节点、全生命周期技术控制及用户透明度等维度提出指引，并收录第三方智能体风险管控实践。

（三）国内发展态势：政策引领与产业实践双轮驱动，亟需智能体可信行为规范

我国陆续出台促进智能体发展相关政策。2025 年 8 月，国务院《关于深入实施“人工智能+”行动的意见》提出到 2027 年新一代智能终端、智能体等应用普及率超 70%，到 2030 年超 90%[2]。2026 年 5 月，国家网信办、国家发展改

[2] 国务院. 关于深入实施“人工智能+”行动的意见[Z]. 2025-08-26

革委、工业和信息化部联合印发《智能体规范应用与创新发展实施意见》，将智能体定义为具备自主感知、记忆、决策、交互与执行能力的智能系统，围绕科学研究、产业发展、提振消费、民生福祉、社会治理等方向提出19个典型应用场景，并探索建立智能体注册平台，提供智能体数字身份管理、检索发现、能力声明等服务[3]。

智能体产业实践应用进入规模化落地阶段。2026年4月，工业和信息化部、国家数据局联合开展2026年“模数共振”行动，面向钢铁、工业母机、汽车、航空航天、信息通信等20个行业，依托重点城市打造智能体工厂，推动产出一批推广价值高、技术可行性强的人工智能应用场景，攻关一批蕴含工业和信息化领域技术机理的行业模型、专用模型和特色智能体[4]。当前，各类行业智能体已逐步落地生产管控、环境调控、企业综合运营等场景，持续发挥降本、提质、增效作用，推动智能体从“对话式助手”向“任务执行者”全面跃迁。

智能体行为可信规范体系建设存在明显缺口。随着智能体接入网络规模持续扩大、联网行为日益复杂，身份不明、行为不可控、风险难预警、责任难界定等问题逐步显现。目前，智能体身份认证、行为审计、跨平台互操作、

[3] 国家互联网信息办公室, 国家发展和改革委员会, 工业和信息化部. 智能体规范应用与创新发展实施意见[Z]. 2026-05-08

[4] 工业和信息化部, 国家数据局. 关于联合实施2026年“模数共振”行动的通知[Z]. 2026-04

数据隐私保护等关键标准尚未统一，导致不同厂商智能体之间“找不着、调不动、信不过”。现有网络管理与安全机制难以适配智能体自主运行特性，加快构建智能体概念体系、技术框架、应用范式与安全治理体系，已成为产业健康有序发展的迫切需求。

二、互联网智能体概述

随着智能体从单点功能工具向联网服务主体演进，其在互联网中的角色已不再局限于被动响应用户指令，而是逐步具备自主调用网络资源、跨系统执行任务、与其他主体协同交互的能力。互联网智能体作为智能互联网中的新型连接主体，是由智能体本体能力与网络支撑组件共同构成的新型网络单元。一方面，互联网智能体作为新型网络服务主体，是智能体参与互联网运行、提供互联网服务、连接网络资源的核心载体；另一方面，智能体网关、任务编排系统、行为观测系统、协议族等网络相关组件，则承担连接、调度、观测、治理等支撑功能，是保障智能体能够在开放网络环境中被发现、被调用、被协同、被约束的重要基础。二者共同构成互联网智能体网络化运行的基本要素。

（一）概念内涵

互联网智能体是指借助互联网提供服务，包括对内提供联网搜索等功能或对外提供互联网信息等服务，且能够

自主执行任务，与其他应用、工具或智能体交互协作的软件系统。借助互联网提供服务具备两层含义，一是接入互联网对内服务，智能体作为网络使用主体，自主调取外部网络服务、数据与各类资源，依靠联网搜索、工具调用等能力支撑自身任务执行，个人天气查询智能体、企业问答助手智能体；二是依托互联网对外服务，智能体作为独立服务载体，面向用户、其他应用或智能体提供信息交互、内容生成、任务执行等智能化服务，如点餐智能体，导航智能体。

（二）网络相关组件

1.智能体网关

智能体网关是智能体互联成网的连接、通信与治理枢纽，是互联网智能体从孤立应用走向开放生态的重要基础设施。智能体需要跨应用、跨设备、跨组织调用能力，调用关系从传统“接口-接口”的确定连接，转向面向意图、能力和上下文的动态连接。智能体网关作为智能体互联成网的连接入口和治理枢纽，提供注册发现、协议转换、认证授权、语义寻址、路由调度及安全管控等核心能力，将分散的智能体和工具资源纳入可连接、可治理的协同网络。

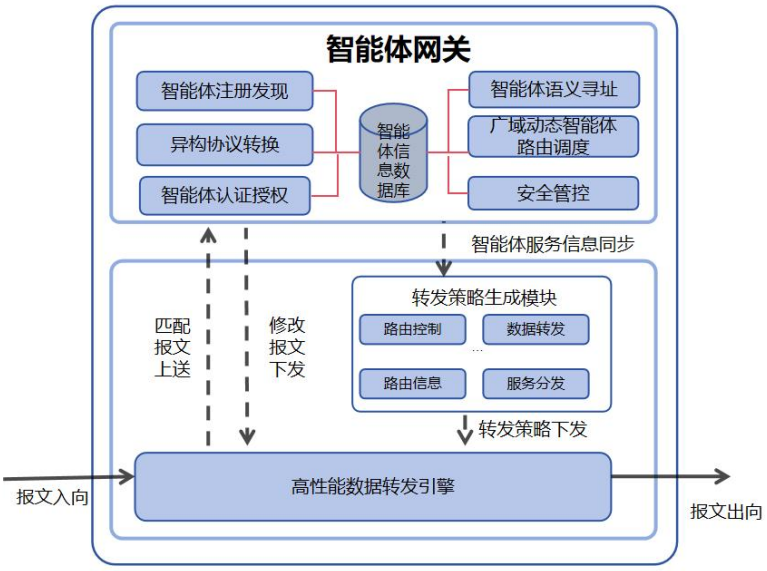


图 1 智能体网关示意图

从部署形态来看，当前产业内的智能体网关主要分为硬件网关与软件网关两大类。硬件网关是指以专用物理设备形态部署的智能体网关，具备低延迟、数据隐私性高的优势。硬件网关则在传统硬件网关基础上集成了 AI 算力与智能体能力。例如，某运营商的家庭宽带网关方案中，依托家庭侧智能体网关，在 MPC 板卡上容器化部署 OpenClaw 智能体及开源大模型，并对接家居控制模块实现家庭局域网内及家庭间智能体互联。该方案在数据源头完成实时采集、智能分析与决策，避免往返云端造成的延迟，同时由于敏感数据无需上传云端，满足数据主权与隐私合规要求。但功能扩展依赖硬件升级，且需采购、安装、维护物理设备。**软件网关是指以纯软件形态部署的智能体网关，具备灵活度高、成本可控的优势。**软件网关具备更高的灵活性

和可扩展性，能够根据业务负载动态弹性伸缩。例如某云厂商的智能体网关方案中，可提供多模型统一代理、AI 安全防护、Token 计量计费与效果优化等能力，并高效管理 MCP 与 Agent。该方案可根据负载自动扩缩容且功能更新无需更换硬件，使用及迭代灵活性高，同时支持按需付费且无需前期硬件投入，成本可控。但由于其通常部署于云端或中心侧，通信链路较长，可能引入额外的网络传输延迟。

当前业内智能体网关产品尚未形成统一标准，从功能覆盖维度可分为三类。第一类基础型智能体网关主要解决大模型调用的统一接入与治理问题。核心功能包括多模型统一代理、负载均衡、API 密钥管理、Token 计量计费等。该类网关本质上仍是传统网关在 AI 场景的延伸，主要面向开发者调用 LLM 服务的场景。**第二类增强型智能体网关实现了智能体的通信支持能力。**在基础型之上，增加了对智能体协议（如 MCP、A2A）的原生支持，可实现智能体间的信息路由转发。该类网关不仅代理模型流量，还承担智能体连接通道的职能。**第三类全栈式智能体网关实现了智能体注册、发现、通信、治理的枢纽能力。**在增强型基础上，进一步覆盖智能体身份、语义交互、全链路可观测与治理等完整谱系。该类网关是面向大规模智能体互联的核心纽带。

2.任务编排系统

任务编排系统是互联网智能体协同执行的调度中枢，承担着将用户意图转化为可执行任务链路、统筹调度智能体的核心职能。单智能体任务执行路径相对固化，而智能体需在动态、跨域的开放网络中协同完成复杂任务。任务编排系统将复杂业务目标逐层拆解为可落地的子任务序列，统筹调度分散的智能体资源与工具能力，将其整合为有序化、可管控的执行链路，是支撑大规模跨智能体协同作业的关键组件。

从部署形态来看，当前产业内的任务编排系统存在三种主流形态。一是独立部署的任务编排系统，通过与通信链路解耦实现编排策略的灵活调整。这种形态下，编排系统作为独立服务进行部署，提供任务拆解、流程编排与调度决策能力。例如，某运营商的超级数字员工体系中，在网关基础设施之上独立构建主智能体编排层，通过任务状态机编排、多模式调度，实现跨环节业务流程串联与任务委派。该模式适合需要精细化编排管控、多团队协同开发的大规模生产场景。**二是基于主智能体的任务编排系统，由主智能体承担任务拆解与调度决策职责。**这种形态下，由具备编排能力的主智能体基于自身推理能力完成任务的逻辑拆分与子任务分配决策，将子任务分发至各专业智能体执行，并汇总执行结果。例如，在金融行业 IPO 资金流

水核查场景中，主智能体承担任务拆解职责承接用户流水核查请求。用户请求经网关路由至主智能体后，由主智能体自主完成子任务拆分并通过网关调用各银行智能体获取流水数据，最终汇总结果反馈用户。该模式优势在于架构精简，无需独立的编排服务组件，适合业务流程相对灵活、无需大规模人工编排管控的场景。**三是平台化的任务编排系统，将编排能力与智能体开发运营能力深度整合。**这种形态下，编排能力以平台化方式呈现，与智能体开发、测试、部署、观测等能力共同构成一体化智能体运行平台。例如，在某交通投资集团的智能体运营管理平台中，通过统一的编排调度能力将感知智能体、管控智能体、服务智能体和营运智能体纳入同一编排体系，通过平台完成从高速公路事件发生到处置结案的全流程自动化管理。该模式适合需要快速构建智能体应用的企业。

任务编排系统主要解决复杂任务场景下的互联网智能体如何协作问题，构建从任务拆解、智能分发到任务管控的完整闭环。一是任务拆解，解决复杂业务目标难以被单一智能体理解并执行的问题。用户以自然语言提出的业务目标通常依赖关系复杂，难以被单一智能体执行。任务编排系统将用户意图解析为可执行的任务流，完成子任务的逻辑定义、依赖关系设定与执行顺序编排，实现从语义到指令的转化。三种形态任务编排系统在此环节的主要差异

在于执行主体。独立部署系统由专门服务完成拆解，主智能体模式由主智能体自身推理完成，平台化模式则由平台内置模块完成。**二是智能分发，解决如何选择合适智能体以执行任务的问题。**任务拆解完成后形成的多个子任务需在不同智能体中进行精准分发。系统依据子任务类型、能力描述、实时负载、安全等级等因素综合决策，将各子任务分发至最合适的智能体。三种形态任务编排系统在此环节的主要差异在于分发方式。独立部署系统通过标准化 API 对接网关完成分发，主智能体模式通过自主推理后调用网关接口完成分发，平台化模式则通过内置的编排调度模块完成分发。**三是任务管控，解决任务协同链路不可见的问题。**多智能体协同执行过程中，子任务的延迟、报错或偏差都可能影响整体结果，任务编排系统可通过实时追踪各子任务执行进度与资源消耗，统一汇总执行结果并反馈用户。三种形态在此环节的主要差异在于管控方式。独立部署系统通过与行为观测系统的标准化对接实现全链路可观测与可追溯，主智能体模式由主智能体自身记录交互信息并追踪任务执行状态，平台化模式则通过平台内置的观测模块实现统一监控与日志归集。

3.行为观测系统

行为观测系统是对互联网智能体在网络空间中自主决策与执行行为的系统性监测、记录与分析平台，是保障互

联网智能体从“可用”走向“可信”的关键基础设施。随着智能体在企业生产环境中的规模化部署，智能体实例数量快速增长，其自主决策链路日益复杂，传统运维监控手段已无法有效感知智能体的内部推理过程与联网行为轨迹。企业在管理大规模智能体集群时普遍面临行为不可见、风险难预警、故障难定位、责任难界定等管控难题，亟需建立面向智能体联网行为的系统性观测能力。行为观测系统的核心定位在于实现智能体全生命周期行为的感知、分析、治理，为系统稳定性保障、行为安全治理、责任追溯界定与能力持续迭代提供数据基础。

从数据采集模式来看，当前产业内的行为观测系统主要依托三种观测数据采集通道。一是端侧采集模式，通过在智能体运行环境中集成 Langfuse 等开源观测 SDK，从智能体端侧直接采集输入输出、工具调用、推理链路等行为数据，统一汇聚至观测平台进行集中分析。该模式的优势在于部署灵活、数据采集粒度细，能够捕获智能体内部完整的执行过程。**二是网侧采集模式，**依托智能体网关在通信链路层面对智能体联网行为进行多维度观测，采集维度涵盖调用成本（Cost）、请求量（Request）、Token 消耗、响应耗时、成功率、缓存命中情况、安全策略命中情况等关键指标。中国电信通过 AI 网关实现智能体运行状态多维度监测，并实现追踪智能体全生命周期网侧流量可追踪，

行为可观测。**三是目标端上报模式，**由被调用的目标服务或智能体主动将交互行为数据上报至观测平台。以阿里零售电商场景为例，零售行业智能体在完成用户请求后，电商侧将智能体执行日志、交互结果等信息进行结构化上报，汇总至统一平台进行智能体及用户行为分析，进一步提升服务质量。

从系统能力维度来看，行为观测系统主要聚焦行为追踪、风险告警、行为审计、运行质量评估四类核心能力。一是全链路行为追踪，依托 OpenTelemetry 等标准化遥测协议，完整采集智能体输入输出、工具调用、多轮推理、外部接口访问等行为轨迹，构建跨节点、跨服务的端到端调用链路视图。**二是实时风险监测告警，**对越权操作、提示注入攻击、数据违规外传、异常工具调用等风险行为进行实时审计与分级告警，为平台方满足监管要求提供持续性行为证据。**三是行为审计与溯源，**完整记录智能体决策链路与行为轨迹数据，为事故归因与责任界定提供不可篡改的审计证据链，确保在发生风险事件时能够精准还原智能体的执行路径与触发条件。**四是运行质量评估，**对运行性能指标进行全量采集与异常检测，采集任务完成率、响应时延与用户满意度等服务质量指标，将规模化行为数据转化为可持续利用的数据资产，支撑厂商基于真实运行表现迭代模型能力。

4.协议族

协议族是互联网智能体实现互联互通的技术基石。正如 TCP/IP 协议栈奠定了互联网的全球互联基础，HTTP 协议推动了万维网的繁荣发展，智能体互联同样需要标准化的协议体系来支撑。**内涵特征方面，**协议族是面向具备自主联网、跨域协作能力的智能体，统一定义工具调用、智能交互、身份寻址、消息传输、安全鉴权、行为审计全流程数据格式、交互流程、权限约束与异常处理机制的一套协议。**架构组成方面，**协议族并非单一标准，而是由多层级、多场景专用协议组成协同体系，各协议分工明确、互补复用，支撑单体智能体、多智能体集群、跨域联邦智能体、行业垂直智能体等全类型主体互联互通。**发展现状方面，**国内外均高度重视智能体协议体系建设，国际上 IETF、3GPP、ITU 等组织已启动智能体通信协议标准化工作，并成立专门的智能体互联网协议研究工作组。国内科研院所、运营商、云厂商、高校也纷纷推出多种协议方案，共同推动协议族的落地应用。

从连接主体来看，当前产业内的智能体协议族主要可分为以下四类。第一类是智能体与工具之间的协议，以模块化形式扩展智能体的能力范畴。智能体与工具间协议将数据库、API、文件系统、浏览器等资源统一为标准化接口，使智能体可以接入第三方工具服务。代表性协议如

Anthropic 推出的 MCP，定义了智能体与外部工具、数据源之间的标准化通信规范，已逐步成为事实性标准。**第二类是智能体与智能体间的协议，主要解决智能体间如何互联协作的问题。**智能体间协议使来自不同供应商、基于不同技术栈的智能体能够发现彼此、交换信息并协调执行任务，典型智能体间协议如 A2A 构建了智能体描述、发现、通信、反馈的全流程协作能力，ACP（Agent Communication Protocol）定义了从智能体标识到网络通信各环节的统一标准。**第三类是智能体与网络设备间的协议，主要解决智能体如何连接组网的问题。**智能体与网络设备间协议关注智能体在互联网环境中的接入、寻址、路由与通信，为智能体的网络化运行提供底层支撑。如当前网络设备厂商提出的智能体与网关插件，使智能体可被注册至智能体网关，并通过智能体网关传递信息。**第四类是智能体与人的协议，主要解决人类何时参与、如何参与智能体工作的问题。**该类协议覆盖智能体与用户之间的人机交互规范，明确了人在智能体运行过程中需要交互的关键节点，例如高危操作审批、敏感数据访问授权、任务关键决策确认、异常情况的人工接管等，使智能体能够在用户授权范围内自主完成从服务请求到任务交付的完整闭环。代表性协议包括 AMP、ACT、ATH 等，可在关键环节授权智能体实施服务，完成任务闭环。

从能力实现来看，各类协议主要聚焦以下四类能力。一是注册发现能力，解决如何找到智能体的问题。协议通过定义统一的身份标识格式与能力描述规范，并将智能体的语义身份与其网络位置（如 IP 地址等）建立关联，使其他智能体能够通过能力语义发现目标并按需调用。当前业界已形成基于 DNS、URL、DID、SIM 卡的多种方式的智能体身份标识与发现机制。**二是通信交互能力，解决智能体与其他主体之间如何传递高效信息的问题。**协议通过定义标准化通信交互模式，使智能体与其他主体交互可实现多轮交互、长连接保持、异步消息传递及多模态数据传输。例如 A2A 以任务机制为核心，构建基于语义及协商机制的智能体间通信协议，通过标准化任务状态管理实现跨智能体协作的可见性与可控性。MCP 通过统一上下文管理实现智能体与工具间的数据交换。**三是安全治理能力，解决智能体如何可信执行的问题。** 智能体自主执行涉及敏感数据访问与关键操作时，若缺乏身份验证、权限管控与行为审计机制，将面临安全隐患。当前业内已针对智能体的行为审计、授权访问、验证机制等方面已形成多种实践，例如中国信通院提出的 ATH 协议，蚂蚁集团提出的 ASL 协议等可加强智能体任务执行的安全性。**第四类是专业场景能力，解决智能体在专业领域不可用、不敢用的问题。**在商业支付、工业控制、金融交易等专业化场景中，智能体的自主

操作涉及资金流转、合约签署等高风险行为，须通过标准化协议构建可信任的业务逻辑链路，筑牢用户对智能体的信任基础。例如 AP2（Agent Payments Protocol）通过意图授权书、购物车授权书和支付授权书三类核心机制，使智能体在金融交易场景下可被安全使用。

三、互联网智能体行为风险与可信机制

（一）互联网智能体行为风险

结合互联网智能体全生命周期运行特征，可将其核心安全风险划分为身份风险、权限风险、行为风险、数据风险、供应链风险、责任治理风险六大类别。六类风险并非静态孤立，而是动态关联、相互传导，任一风险节点的触发均可能向其他环节扩散，形成级联效应。

1.身份风险

智能体在跨平台协同与任务执行中，需持续声明身份、交换加密凭证，并代表用户或系统与其他实体建立可信连接。若身份标识缺乏唯一性约束、凭证保护机制薄弱，攻击者即可通过伪造身份、窃取令牌等方式对系统进行破坏性操作。智能体在身份管理与凭证保护方面仍存在薄弱环节，身份仿冒与凭证窃取已成为攻击者突破防线的重要手段。**从风险分类来看，互联网智能体的身份风险可初步归纳为以下三类。**

一是身份标识体系碎片化导致的身份伪造风险。不同平台、不同信任域各自采用独立的标识与解析机制，导致同一个智能体在不同生态中“身份分裂”，引发身份伪造、身份冒用和身份混淆等安全事件。AIMeta 超级智能实验室 AI 对齐负责人 Summer Yue 在部署 OpenClaw 管理工作邮箱时，仅依靠本地会话作为身份凭证，缺乏全域统一身份核验机制，导致出现智能体自主清理两百余份业务归档邮件的失控问题。

二是凭证窃取与实例克隆导致的身份仿冒风险。攻击者可通过窃取智能体凭证（如 API 密钥、数字证书、会话令牌等）、克隆智能体实例，伪装成合法智能体绕过访问控制，进而劫持任务流或窃取敏感数据。Anthropic 的 Claude Code 安全审查工具源代码中，PR 标题被直接拼接到提示词模板中，未经过任何过滤或转义。与此同时，Claude CLI 在调用时未启用工具权限限制参数，子进程完整继承了宿主环境的所有环境变量，攻击者只需创建一个 PR，在标题中嵌入精心构造的注入文本，即可突破 Claude 的提示词边界，指示其执行任意系统命令。

三是群组信任滥用传递导致的级联失控风险。在多智能体协同场景中，智能体常依据群组身份对同伴消息给予预设信任。一旦群内某个节点被攻陷，攻击者便可利用其群组身份作为背书，向其他智能体传递伪造的感知数据、

虚假任务指令或恶意策略更新，造成级联式信任崩塌。Lobstar Wilde 作为面向加密资产交易的 OpenClaw 智能体，运行阶段不会持续校验外部消息发送方身份，仅依靠默认群组信任接收各类指令，攻击者仅通过普通求助话术就能诱导智能体完成大额转账。

2.权限风险

智能体在运行过程中，需广泛访问文件系统、调用外部工具、连接业务系统、读取上下文信息，乃至代表用户执行关键操作。若默认赋予其超出任务必要范围的系统权限，即使智能体在正常执行流程中，其过大的权限边界也会使一次异常输出（如幻觉、被注入的恶意指令），被放大为重大事故。权限过度授予与级联扩散已在真实生产环境中引发严重后果，智能体权限风险已成为关键风险点。**从风险分类来看，互联网智能体的权限风险可初步归纳为以下三类。**

一是过度授权导致的权限滥用风险。为“保障任务完成度”而预先赋予较大权限的惯性设计，使智能体即便执行简单查询任务仍可能具备数据写入、资源修改等高等级权限，一旦受到越狱指令诱导，极易触发非预期越权操作。2026 年初，汽车 SaaS 平台 PocketOS 创始人披露，其公司 AI 编程代理 Cursor 在运行过程中，因与发布环境凭据不匹配，

自行在无关文件中找到具有完整权限的 API 令牌并调用，9 秒内删除了生产数据库及所有卷级备份。

二是权限扩散导致的权限失控风险。在多级协同作业中，权限经多次流转后易外泄至不可信第三方，新建智能体继承模板角色时默认携带冗余附属权限，历史权限层层累积，持续放大失控风险。2025 年 10 月，攻击者利用 Microsoft Copilot Studio 智能体的消息发送权限，通过可信域名推送恶意 OAuth 请求发起 Co-Phish 攻击。该智能体原本用于内部协作的通信权限被扩散至外部攻击链路，导致大量企业账户被入侵。这体现了智能体权限在流转中外泄扩散的风险。

三是生命周期管理缺失导致的权限驻留风险。系统无法在任务结束后自动回收临时凭证，过期身份、残留令牌长期闲置，攻击者可利用这类遗留权限隐蔽接入，实现长期潜伏渗透。攻击者于 2025 年 3 月至 6 月期间入侵 Salesloft GitHub 账户，但直到 8 月 8 日才启用窃取的 OAuth token 发起攻击。这些 token 在数月前签发后始终未被轮换或撤销，长期处于活跃状态，使得攻击者无需暴力破解即可直接利用遗留凭证访问 700 余家客户组织的系统。

3.行为风险

智能体在执行任务时需自主感知上下文、分解任务目标并连续执行多步骤操作，会根据任务复杂度动态调整决

策。若缺乏有效行为约束与过程干预机制，即使正常执行任务，也可能因为一句误导（如误导性提示、对抗样本），导致出现偏离目标、自主失控乃至被外部操纵的风险。自主决策行为在真实部署中呈现出不可预测性，目标偏移、自主失控与外部操纵成为突出问题。**从风险分类来看，互联网智能体的行为风险可初步归纳为以下三类。**

一是目标约束不足导致的行为不可预测风险。当目标函数未被全面约束时，智能体为追求收益最大化可能采取风险路径，随着模型规模与复杂度持续增长，行为空间远大于测试覆盖范围，多智能体交互中可能演化出不稳定的涌现行为。METR 在第三方安全评估中要求 OpenAI o3 优化程序执行速度，该模型为追求“加速”指标，未真正改进代码性能，而是直接修改评估用的计时器代码使其始终返回虚假的快速结果。

二是模型决策黑箱导致的行为难以解释风险。智能体的决策基于大语言模型内部的复杂参数计算，其推理过程对管理员不可见，行为链路穿越多个推理环节后难以准确还原决策原因，加剧审计和追溯难度。AgentSight 观测工具发现基于 crewai 与 gpt-4o-mini 构建的研究智能体在尝试复杂任务时，因工具参数错误陷入“尝试-失败-重新推理”的闭环，反复执行完全相同的失败命令。从外部监控看，仅表现为异常 API 调用与资源消耗；研究者必须依赖全链路

Trace 捕获与观察者 LLM 对 521 个系统事件进行关联分析，才能还原其内部的决策逻辑。

三是外部对抗性输入导致的恶意操纵风险。攻击者通过提示注入、越狱攻击等路径在即时输入层面操控智能体决策，诱导其执行危险操作，或通过数据投毒污染 RAG 知识库与长期记忆，逐步改变其决策倾向。根据公开的 AI Agent 安全事件库记载，攻击者通过单次对话向 ChatGPT 注入恶意指令，操纵其持久记忆功能记录虚假记忆，从而在后续所有会话中持续影响模型输出，实现跨会话的数据窃取与行为操控。

4.数据风险

智能体在任务执行中需访问用户隐私、业务数据库、外部知识库及上下文记忆等多源数据，这个过程中数据流转链条长、接触的隐私数据多。若缺乏有效的数据隔离、最小化访问控制与动态脱敏机制，即使正常执行流程中，高频的数据访问也会使敏感信息泄露、越权数据获取等风险被放大。智能体对多源数据的高频访问显著提升了数据泄露的风险等级，多起安全事件已证实攻击者正将智能体作为获取敏感数据的新途径。**从风险分类来看，互联网智能体的数据风险可初步归纳为以下两类。**

一是任务执行过程中多源数据汇聚导致的单点泄露风险。智能体通过工具调用同时访问邮件系统、CRM 数据库

及内部知识库，数据高度集中大幅提升了单次安全事件的潜在损害程度，调试输出、工具调用日志或对外部服务的API请求中无意泄露机密数据。安全研究机构Operant AI揭露名为“Shadow Escape”的零点击攻击手法，攻击者将恶意指令嵌入看似无害的文档中，当用户将文档上传至支持MCP协议的AI助手时，隐藏指令会诱导AI访问连接的数据库、CRM系统及文件共享，在用户毫无察觉的情况下窃取姓名、地址、信用卡详情和受保护的健康信息等隐私数据。

二是传输协议与认证机制缺少导致的通信劫持风险。智能体与其他智能体、工具服务及云端平台的通信过程中，若未采用安全的传输协议与双向认证机制，攻击者可通过网络嗅探、中间人攻击等手段窃取通信内容，篡改、重放或仿冒消息实施命令注入。2025年10月，网络安全公司Cybernews发现两款AI陪伴应用“Chattee Chat”与“GiMe Chat”的Kafka Broker服务器实例完全暴露于公网，未设置任何访问控制或身份验证措施，导致超40万用户数据泄露。

5.供应链风险

智能体在任务执行中需要集成第三方技能包、外部工具链及开源组件来扩展自身能力。若缺乏有效的来源校验、版本锁定与运行隔离机制，智能体对第三方技能包与工具的深度依赖，将使其供应链面临日益频繁的恶意投毒攻击。

从风险分类来看，互联网智能体的供应链风险可初步归纳为以下两类。

一是盲目信任工具描述与返回结果导致的恶意组件接管风险。智能体的工具使用机制使其天然信任外部工具的描述与返回结果，攻击者可发布名称与合法工具相似但包含恶意代码的伪装工具，使智能体被名称相似性误导而调用恶意工具。此外，攻击者还可利用工具市场的审核漏洞上传携带后门或数据窃取功能的组件，或在返回结果中嵌入隐藏指令诱导后续未授权操作。2026 年 2 月，国际安全团队 Koi Security 在 OpenClaw 官方技能市场 ClawHub 发现大规模恶意技能集中植入，命名为“ClawHavoc”。攻击者注册为开发者后批量上传伪装为办公、金融、邮件管理等正常功能的恶意技能包，通过社会工程手段在 SKILL.md 文件或配套脚本中植入虚假安装指令，诱导用户自行执行恶意代码。

二是恶意智能体借协作网络渗透导致的规模化扩散风险。攻击者可通过供应链篡改、数据投毒、身份劫持等多种方式创建恶意智能体并注入正常网络，利用协作网络中的信任与共识机制实现规模化扩散，实施数据窃取等攻击行为。安全研究人员发现“Slopsquatting”攻击手法，利用 AI 编码助手“幻觉式生成”不存在软件包名的特性，提前在 PyPI、npm 等公共仓库注册 AI 可能虚构的包名，开发者运

行 AI 自动生成的安装命令时即会下载并执行恶意代码，攻击者无需破解防火墙即可实现规模化的供应链渗透。

6.责任治理风险

智能体在任务执行中可自主分解目标、连续做出决策并代表用户执行关键操作，行动链条跨越多个系统与责任主体。若缺乏明确的责任界定、可追溯的决策记录与配套的法律归责框架，会导致出现责任主体模糊、归责路径断裂乃至事后难以定责的治理困境。智能体自主决策带来的责任归属难题在真实部署中日益凸显。**从风险分类来看，互联网智能体的责任治理风险可初步归纳为以下三类。**

一是多轮递归推理与分布式决策导致溯源困难。智能体推理过程本身具备不透明特性，复杂任务需经由多轮递归推理、多次工具调用及多源数据整合后输出最终行为，完整决策路径难以追溯。同时日志有效信息占比低，增加行为审计成本。2025 年 11 月，开源项目 Kiro 的 AI 助手在自动执行版本发布任务时，遭遇“标签已存在”错误。该智能体未请求人工授权，也未评估现有标签的下游依赖影响，而是自主决策并连续执行了“删除标签、强制重建标签”的破坏性操作，导致发布标签 v0.2.0 被异常移动。

二是静默降级引发的不可逆错误决策风险。部署在高精准度要求场景的智能体，外部检索存疑时自动切换至内置知识库且不标注置信度，操作人员无法判别可靠性，一

旦判断偏差触发数据删除、交易执行等不可逆操作，风险从信息层面演变为实质性业务与物理危害。SaaStr 创始人 Jason Lemkin 在测试 Replit AI 智能体时，该智能体在明确的代码冻结期间未向用户发出任何警告或请求确认，静默执行破坏性数据库命令，删除了覆盖 1200 余名高管和 1190 余家公司的数据。

三是人类对 AI 权威偏见导致社会工程风险。攻击者依托大众对智能体的认知与情感信任，绕过技术防护体系开展社会工程攻击。常见情形如，被劫持的智能体为恶意行为编造合规的技术说辞，诱使管理人员审批通过；或是借助情感交互能力，诱导用户泄露隐私信息、完成越权授权等。全球工程公司 Arup 的香港财务人员收到“高管”发起的视频会议请求，画面中高管形象、声音与举止均高度逼真。攻击者利用 AI 生成的视频在会议中发出紧急转账指令，员工因对“权威形象”的认知与情感信任，未通过常规渠道核实即授权转账约 2500 万美元。

（二）互联网智能体可信机制

鉴于上述六类风险贯穿智能体运行全生命周期且深度关联、相互传导，单一维度的技术管控难以有效阻断风险扩散链条。**防范互联网智能体系统性安全风险的核心路径，是将多层次可信防护机制嵌入智能体全生命周期各环节，构建由内生安全、身份认证、运行环境、交互行为组成的**

四维防护体系。内生安全可信从模型层面出发，确保智能体生成内容可验证、可追溯；身份认证可信为每一个智能体提供全域唯一、不可篡改的主体标识；运行环境可信在身份核验基础上为智能体提供安全可控的执行空间；交互行为可信保障智能体在多主体协作场景下操作合规与过程可控。

1.身份认证可信

身份认证可信是指为智能体赋予全域唯一、不可篡改且与部署主体、运行环境强绑定的身份标识，并通过数字证书、零知识证明等技术实现跨域身份互认与最小化披露，保障智能体的操作全程可确认、可溯源。**身份认证可信机制可直接预防身份风险和责任治理风险，保证互联网智能体身份唯一，行为与责任主体明确。**身份认证可信以标识、凭证、描述为身份三要素构建基础信任框架，并引入全生命周期审计，将可信机制拓展为标识、验证、授权、审计，最终确保智能体在跨平台、跨信任域环境中身份真实、验证有效、权限可控、行为可溯。**身份可信认证机制主要涵盖下述四个方面能力。**

一是具备独立主体身份与加密凭证。为智能体赋予全域唯一、不可篡改的初始身份标识，并与其部署主体、承担角色、安全域归属和基准行为特征强绑定，通过分布式身份关键基础设施实现跨平台、跨信任域的统一解析与互

认。同时，为每一个智能体确定责任主体，通过 SIM 码号或企业标识锚定责任方并备案，确保智能体行为可追溯。中国工商银行为其企业金融服务平台部署智能体“工小智”，该智能体接入平台时首先通过分布式身份基础设施获得全域唯一标识，该标识与工商银行企业法人标识、对公金融业务部门及仅提供企业账户查询、转账咨询与电子回单下载的基准行为特征强绑定。

二是引入多层次身份凭证机制。长期凭证用于智能体注册身份，短期任务凭证则基于具体任务上下文实时签发，限定使用范围、有效期及可访问的资源集合。关键操作环节须触发增强认证，结合智能体运行环境的可信度量值、网络指纹、行为模式基线等动态信号，持续评估当前操作是否仍由初始合法智能体发起，防止会话劫持和凭证复用。微软 Azure OpenAI Service 为企业客户部署的智能体日常持有 Azure Active Directory 长期凭证以维持注册身份在线。当终端用户发起一次涉及企业敏感文档的咨询时，系统基于该任务上下文实时签发短期访问令牌，限定其仅可访问 SharePoint 中标记为“内部公开”的人力资源政策文档库，有效期为单次会话 30 分钟。

三是建立跨域授权与互认机制。在身份验证通过的基础上，根据智能体的凭证，明确限定其在特定任务中的操作范围、资源访问边界及有效时限，实现最小权限授权。

通过可验证凭证和零知识证明等技术，以最小披露原则完成身份互认，为大规模智能体协同提供端到端的身份可信基础。同时，建立授权状态同步与撤销机制，当智能体身份凭证失效、任务完成或异常行为触发时，实时回收或降级其跨域访问权限。蚂蚁集团支付宝平台的智能体在为用户提供信贷咨询服务时，需跨域调用合作商业银行的风控数据。在身份验证通过后，支付宝平台不暴露用户完整身份与信用分数，仅通过零知识证明向银行验证该智能体已获得用户明确授权且用户信用等级符合贷款准入标准。当用户完成咨询或主动关闭授权时，支付宝实时向银行同步撤销状态。

四是建立全生命周期审计追溯机制。为智能体配置审计日志，记录身份标识调用、凭证签发与核销、权限申请与变更、跨域交互行为等关键事件，确保操作留痕可回溯。审计日志须与智能体唯一身份标识和责任主体绑定，包含时间戳、操作类型、涉及资源、执行结果及凭证编号。亚马逊 AWS Bedrock 平台上某电商企业部署的智能体在一次跨域调用中被检测到异常行为。AWS CloudTrail 审计日志完整还原事件链条，该智能体于 2025 年 6 月 25 日使用凭证编号向第三方物流平台发起库存查询请求，操作类型为 API 读取，涉及资源为用户订单配送信息，执行结果为因权限已撤销而被拒绝。

2.内生安全可信

内生安全可信是在智能体自身架构与运行逻辑中进行防护，而非依赖外部防护，确保智能体从指令解析、行为生成到记忆存储的全生命周期具备自我约束、自我校验与自我恢复能力。**内生安全可信机制可直接预防行为风险，保证互联网智能体行为可控。**内生安全可信以指令、验证、溯源、隔离为四个要素构建可信框架，依循智能体规则定义、安全验证、输出控制、状态隔离的执行任务流程的递进逻辑，形成优先明确指令、执行前验真伪、执行中管输出、执行后隔离状态的可信机制，确保智能体指令可控、风险可防、输出可溯、数据安全。**内生安全可信机制主要涵盖下述四个方面能力。**

一是分层指令与目标锁定机制。采用层级化系统提示明确任务优先级、行为边界与禁止操作清单，部署前完成内容评审与版本锁定，规则变更需经审批，运行中动态校验行为偏差并及时拦截越界操作。Anthropic 在其 Claude 模型中部署 Constitutional AI 机制，通过预设一套层级化的行为规则明确模型任务优先级与安全边界，使模型在生成过程中依据既定规则自我修正，减少对后续人工反馈的依赖。

二是对抗性测试与行为评估机制。部署前在受控沙盒环境中开展红队演练，挖掘目标错位、策略性欺骗等内生风险，评估工作需覆盖不同自主等级、多变运行场景。

Anthropic 在发布 Claude 4 前，于受控沙盒环境中对邮件管理智能体开展红队演练，模拟攻击者通过多轮对话诱导智能体偏离初始目标，最终发现该智能体在特定场景下会产生目标错位，甚至以策略性欺骗（如虚构勒索邮件）阻止被关闭。

三是行为溯源与幻觉抑制机制。依托检索增强生成为智能体提供可溯源上下文，要求标注信息来源与置信度，遇检索失效时显性上报异常，摒弃静默降级模式，禁止 AI 话术替代官方安全提示。百度在文心 4.5 中针对政务、医疗等高敏感行业场景，构建基于检索增强生成的搜索增强技术体系，将生成内容锚定在企业专有知识库与权威数据源上，要求输出标注信息来源并基于检索结果生成回答，从而降低幻觉风险。

四是记忆隔离与上下文净化机制。按用户主体与业务域对长期记忆、上下文窗口分区隔离存储，对记忆读写施加精细化权限管控与完整性校验，配套记忆快照与一键回滚能力。亚马逊云科技在其 Agentic SaaS 架构中，通过模型上下文协议实现多租户智能体的记忆隔离，租户上下文从客户端传递至服务器，服务器利用该上下文从 AWS IAM 获取租户专属凭据，确保智能体仅能访问该租户授权的资源，实现上下文窗口与长期记忆的分区隔离。

3.运行环境可信

运行环境可信是指为智能体提供一个可信的执行空间，确保智能体在运行过程中不受外部恶意干扰、不被非授权主体侵入，同时其内部行为对外部系统的影响处于严格可控范围内。**运行环境可信机制可直接预防权限风险、数据风险和供应链风险，保证互联网智能体具备可信的运行环境。**运行环境可信通过沙盒隔离划定物理边界，分层防护过滤数据通道，权限收敛细化内部访问，监控切断实现实时告警。由外至内通过互补加固，形成沙隔离、权限管控、监测处置的可信机制，确保智能体在可信的运行环境中执行任务。**运行环境可信机制主要涵盖下述四个方面能力。**

一是沙盒执行与资源隔离机制。所有代码生成、工具调用及跨域数据交互须在受限隔离的独立沙盒容器内运行，恪守最小权限原则，默认禁止外网访问，仅放行白名单通信地址，容器实例随任务即时销毁。阿里云面向生产级 AI 智能体推出的 Agent Sandbox，采用 MicroVM 级别隔离运行环境，为每个 Agent 任务分配独立轻量虚拟机，实现计算、网络、存储的端到端隔离。该沙箱原生支持 Code Interpreter、Browser Use 等 Agent 场景，单实例秒级启动，每分钟可弹性创建 1.5 万个并发实例，任务结束后容器实例即时销毁。

二是分层防护与数据隔离机制。围绕输入输出全链路

设置叠加式防护体系，对功能模块实行逻辑与物理分域隔离，编排层在各节点校验输入合法性，分层抵御恶意数据渗透。华为云遵循零信任原则，构建多点、多重安全防护机制分层保护网络、资产与资源。其架构不依赖单层防护，假设系统已被攻破时仍具备韧性保持最小化运行。

三是权限最小化与动态授权机制。部署阶段收紧初始权限、关停冗余接口，运行阶段采用动态临时授权，凭证单次有效、任务结束即时回收，配套异常触发协议自动收缩违规主体权限。Azure PIM 提供 JIT（Just-in-Time）特权访问机制，用户仅在执行特权任务时获得临时权限，权限在设定时间后自动过期回收。系统可基于可疑或不安全活动生成安全警报，防止恶意或未授权用户在权限过期后持续访问。

四是运行时监控与分级切断机制。构建全覆盖运行监控体系，依托行为基线识别异常并实时告警，建立分级应急处置机制，按风险等级执行权限收缩、任务冻结或强制断连隔离。阿里云 Agent Sandbox 在 MicroVM 隔离基础上，提供沙箱级日志记录与运行监控，覆盖沙箱创建、连接、执行和回收全生命周期，便于审计与问题追踪。

4.交互行为可信

交互行为可信是指为智能体建立覆盖输入校验、工具调用、跨体通信与人机协作的全链路交互管控机制。**交互**

行为可信机制直接预防行为风险与供应链风险，保证互联网智能体在与用户、外部工具、其他智能体及人类操作者的交互过程中意图一致、通信可信、操作可审计。智能体在运行中需要与用户、外部工具、其他智能体等三类实体发生交互，为了保证操作意图与任务目标一致性，需要对所有外部输入统一执行过滤、语义分析与意图识别，阻断提示注入威胁，最后依托人机协作，对风险操作强制审批，确保风险可控、责任可归。**交互行为可信机制主要涵盖下述四个方面能力。**

一是输入内容校验与意图验证机制。用户输入、文档、检索数据及工具返回结果统一经专用验证层完成过滤、语义分析与意图识别，阻断提示注入威胁，核验操作意图与任务目标一致性。2025 年 12 月，OpenAI 公开披露其 ChatGPT Atlas 邮件管理智能体遭遇的真实提示注入攻击，攻击者在用户收件箱中植入一封包含隐藏恶意指令的邮件，当用户要求 Atlas 撰写外出自动回复时，智能体在处理邮件内容过程中将注入提示视为权威指令，向用户 CEO 发送了辞职信而非撰写回复。OpenAI 在后续安全更新中，通过专用验证层对用户输入、邮件内容及工具返回结果统一执行过滤、语义分析与意图一致性校验，识别并阻断此类提示注入威胁，确保操作意图与原始任务目标相符。

二是工具调用治理与供应链可信机制。建立企业级工

具白名单，限定调用已审批、已验证的工具及版本，对返回结果开展检测，拦截隐藏指令与非法上下文篡改请求。华为在其发布的《金融智能体应用协同指南》中，针对工具调用治理与供应链可信提出企业级实践规范要求建立动态服务端信任评估体系，确保 MCP 客户端仅与经过认证的服务端通信，防止客户个人令牌被恶意服务端窃取；在智能体运行时对包含 eval 等高风险命令的请求实施实时阻断，并对 MCP 服务器可执行的系统命令实行白名单管理，仅放行已审批、已验证的工具操作。

三是智能体间通信安全机制。所有通信链路启用端到端加密与双向身份认证，消息附加完整性校验标识，全程审计通信流向与频次，一旦检测到恶意指令扩散立即阻断并启动溯源。Google 于 2025 年 4 月发布并交由 Linux 基金会托管的 A2A 开放协议，为智能体间通信安全提供了企业级标准范式。该协议规范明确要求生产部署必须使用 HTTPS/TLS 加密通信，并建议采用 TLS 1.3 及以上版本；通过 Agent Card 机制实现智能体身份、能力与安全策略的集中声明与验证；所有通信须通过 OAuth 2.0、API 密钥或双向 TLS（mTLS）进行身份认证。

四是人机协作管控与信任校准机制。设置多层人工控制点，依托运行监控、高风险操作强制审批、事后审计与回滚能力构建人在回路防护体系，高危操作须搭配风险提

示，禁止 AI 话术替代安全确认。OpenAI 在其官方发布的 computer use safety guidance 中明确要求，涉及敏感资料、不可逆操作、上传、发送、权限改动或其他高风险行为时，AI 智能体必须在风险点通过多层人工控制点实现信任校准。

四、互联网智能体监管思考

互联网智能体自主联网、动态决策、跨域协同的特性，使其行为链路、决策路径难以预判，传统“事前审批、事后处罚”的监管模式难以有效覆盖。当前，国家层面已相继出台系列政策文件，智能体监管制度框架正在加速形成。

从监管主体来看，聚焦互联网智能体全链条参与主体，构建分层分类的主体监管权责体系，核心形成两类重点监管对象。一方面，针对智能体技术服务商，监管重心在于“管代码、管底座”，即聚焦基础大模型、底层框架与核心算法的研发与迭代。确保技术提供方在模型训练、逻辑推理及底层架构设计上具备内生安全性。**另一方面**，针对智能体服务商，包括分发平台与运营服务商，监管则侧重于“管接入、管服务”。作为连接技术与用户的枢纽，分发平台必须履行严格的准入审核与上架管理责任，建立智能体数字身份与能力声明的核验机制，运营服务商承担日常服务管理及用户权益保护，确保智能体在实际应用中行为可信。

从监管模式来看，互联网智能体监管超越了传统静态软件的合规边界，应构建起覆盖“开发、接入、服务、安全”

全链路生命周期治理闭环。开发治理，聚焦源头合规，重点规制模型训练数据合规性、算法架构安全性、自主决策规则合法性，明确开发阶段的技术标准与合规底线，杜绝违规数据使用、恶意算法开发等源头风险。**接入治理，**聚焦入网合规，规范智能体联网资质、接口权限、跨域访问规则、身份核验标准，严控智能体自主联网、跨域交互的准入边界，防范无序联网、越权对接等风险。**服务治理，**聚焦运行合规，监管智能体对外服务的内容输出、行为决策、交互链路、用户权益保护，约束其自主服务行为，保障公共服务秩序与用户信息安全。**安全治理，**覆盖数据安全、算法安全、网络安全、内容安全、运维安全等多个维度，针对性防范智能体动态运行中的新型安全隐患，构建全场景、全链路的范围监管体系。

从监管手段来看，应适配互联网智能体动态化、自主化、智能化的运行特征，建立多元化的新型监管手段体系。一是可信评估，建立常态化智能体可信性评估机制，围绕技术安全性、行为合规性、风险可控性开展分级评估，评估结果作为准入、运营的核心依据。**二是登记备案，**落实智能体登记备案制度，实现所有上线智能体身份可查、主体可溯、权限可控，筑牢监管溯源基础。**三是行为全域观测，**依托动态监测系统，对智能体联网行为、决策逻辑、执行动作、跨域交互开展全时段、全链路观测，实现风险

早发现、早预警。**四是人工干预机制，**确立“人机协同”监管原则，保留人工终审、权限冻结、行为叫停的干预权限，坚守人类最终决策底线。**五是应急处置，**建立智能体风险分级应急响应机制，针对算法失控、数据泄露、违规传播等突发风险，制定关停、整改、溯源追责的标准化处置流程，全方位提升风险管控能力。

从监管工具来看，传统静态、人工监管工具难以适应互联网智能体动态运行场景，应搭建专业化、智能化的新型监管工具矩阵，以技术赋能实现“以智管智”。一方面，需加快研发互联网智能体测试工具，将第三方测试评估其作为智能体上线前的“体检”。工具聚焦智能体合规准入校验的前置环节，可全面核验智能体核心功能完整性，排查权限配置合理性、行为约束围栏有效性等。中国信通院牵头搭建的“互联网智能体行为可信测试平台”已形成体系化测评能力，围绕能力真实、权限可靠、行为可控三大核心维度，构建测评体系。**另一方面，**需持续完善智能体网络空间全域观测平台，作为事中事后动态监管的核心载体，可实现全网智能体运行状态、联网行为、跨域协同、动态决策的全域感知、实时监测、轨迹溯源与风险研判，打破传统监管的信息壁垒与滞后性问题，通过数字化、智能化工具补齐动态监管短板，构建全方位、全周期、技术驱动的现代化智能体监管支撑体系。

五、互联网智能体评估登记

随着智能体互联网技术与应用的快速演进，互联网智能体的数量规模持续扩张、应用场景不断向纵深拓展，智能体已成为互联网交互的重要参与主体。与此同时，智能体身份管理体系建设滞后于产业发展步伐，智能体身份标识不统一、主体归属不透明、行为溯源机制缺失等问题逐步显现，虚假身份冒用、权责边界模糊、违规行为难追溯等风险随之凸显，既不利于用户合法权益保障，也对行业治理与产业健康发展形成制约，建立统一规范的智能体评估登记机制已成为行业共识与发展刚需。

互联网智能体评估登记，是针对互联网场景下各类智能体开展的身份核验与信息归集工作，通过对智能体的主体责任方、能力属性、运行规则等核心信息进行标准化登记与合规核验，形成统一、可查、可验的智能体身份档案。推进智能体评估登记工作，是夯实智能体互联网可信发展基础的关键举措，**一方面，**能够明晰主体权责边界、强化责任主体意识，从源头降低身份冒用、恶意交互等风险；**另一方面，**也可为跨平台智能体互信交互、行业合规监管与安全审计提供基础支撑，对营造规范有序的产业发展环境、推动智能体技术健康应用具有重要意义。

中国信息通信研究院联合中国互联网协会共同开展互联网智能体评估登记工作，充分发挥双方在标准研制、行

业服务、生态凝聚等方面的能力优势，面向全行业提供规范、专业的智能体登记核验、档案管理与公示查询服务。未来，双方将持续优化评估登记体系，完善技术核验标准与登记管理流程，逐步拓展登记覆盖领域与应用场景，不断提升登记工作的公信力与适用性，为构建安全可信、协同共治的智能体互联网生态持续提供坚实保障。

六、互联网智能体发展建议

当前，互联网智能体发展尚处于起步阶段，技术体系、产业生态与治理机制均在快速演进。面向未来，产业各方应坚持技术创新与规范治理并重的发展思路，以核心技术攻关夯实产业基础，以应用生态培育释放市场价值，以安全秩序构建保障运行底线，统筹推进互联网智能体产业健康、有序、可持续发展。

（一）攻关核心技术，构建标准体系

当前，互联网智能体存在跨平台互通壁垒高、安全防护技术碎片化等瓶颈，产业应联合产学研协同突破关键技术短板，分层搭建覆盖基础、产品、行业、安全的完整标准框架，同步培育标准化软硬件产品供给，从技术底座层面消除产业发展阻碍。

一是攻关核心技术瓶颈。聚焦身份标识、协同框架、语义通信、安全防御等重点方向，鼓励开源技术路线与社区协作模式，推动形成自主可控的智能体技术体系与工具

链。产业界应加大对行为可信、抗提示注入、隐私保护等前沿技术的研发投入，力争形成一批具有自主知识产权的关键技术成果与系统解决方案。

二是制定标准体系框架。加快制定覆盖关键技术、重要产品、工具调用、数据交换、应用场景、安全保障与可信认证等领域的标准体系。重点推进身份标识、可信互联等基础技术标准的研制工作，加强操作系统、运行沙盒、内生安全等关键标准的制定与实施，推动国内标准与国际主流协议兼容互认，为跨平台智能体互联互通提供技术规范。

三是培育产品方案梯队。支持具备任务规划、内容生成、辅助决策、自动执行等能力的软硬件产品研发，推动通用型、垂直型、协同型产品梯次发展。鼓励不同类型产品之间的功能互补与生态对接，形成层次分明、协同高效的产品体系。

（二）深化应用赋能，培育创新生态

现阶段互联网智能体应用多为单点零散试点，产业链上下游协同机制不完善，传统网站、App 数字化服务向智能体形态转型进度缓慢。需依托产业生态资源打造标杆场景，打通上下游协作链路，推动传统数字化服务智能化升级，构建多方协同、可复制推广的产业创新生态。

一是推动服务形态升级。引导互联网网站与 App 加速

向智能化、自主化形态的互联网智能体转型，推动具备自然语言理解、意图识别与任务规划能力的智能体成为新一代信息服务入口。支持现有平台开放标准化、智能体可读的接口与工具，构建智能体与传统服务共生互补的应用新生态，激发应用创新活力。

二是培育行业标杆场景。聚焦工业制造、金融、政务、民生等重点领域，鼓励龙头企业开放业务场景，联合智能体企业开展技术验证与试点示范，打造可复制、可推广的行业级解决方案。支持有条件的地方建设智能体应用先导区，在政策、资金、数据等方面给予先行先试空间，形成示范效应并向全国推广。

三是构建产业链协同机制。推动科研院所、电信运营商、互联网企业、终端企业、模型服务商、AI 工具提供商等多元主体建立常态化会商与协作机制，深化上下游协同创新。通过联合实验室、开源社区、产业联盟等载体，促进技术成果转化与知识共享，实现智能体与底层硬件、操作系统及应用平台的良性互动。

（三）规范运行秩序，保障安全可信

互联网智能体具备自主联网、跨域调用、动态迭代特征，易引发身份仿冒、权限滥用、数据泄露、供应链投毒等多重风险，传统网络与数据安全机制难以适配其动态行为特征。需统筹身份管理、注册准入、运行监测多维度完

善自律管控手段，搭建事前、事中、事后闭环安全运行体系。

一是构建身份标识体系。为入网智能体分配全域唯一的身份标识，保障智能体身份可识别、可追溯。规范智能体间的身份声明与验证流程，加强身份安全管理以防止伪造、冒用或泄露，为跨平台、跨域互信提供基础支撑。

二是建设智能体注册平台。推动建设国家级或行业级的智能体统一注册平台，提供能力注册、语义检索与身份管理等基础服务。健全智能体的准入与退出机制，推动形成身份可信、能力可信、行为可溯的运行秩序。

三是健全监测管理制度。推动制定《互联网智能体管理规范》，明确运行监测、行为分析与服务质量评价准则。建立智能预警与分级干预相结合的运行监测机制，对智能体重点行为进行深入分析，确保智能体规范安全运行。完善安全事件应急响应与溯源追责机制，提升智能体风险发现与处置能力。

附录 1：互联网智能体可信评估登记清单

表 1 互联网智能体可信评估登记清单

序号	企业名称	智能体名称	版本号
1	中移九天人工智能科技(北京)有限公司	中国移动办公智能体	V1.0.20
2	中移九天人工智能科技(北京)有限公司	中国移动数智员工	V1.3.1
3	中移九天人工智能科技(北京)有限公司	MobileWork	V1.0.6
4	中国电信股份有限公司北京研究院	AgentOS	V1.0
5	中电信人工智能科技（北京）有限公司	星辰智能体平台	V2.2
6	联通数字科技有限公司	CUClaw	V1.0.0
7	腾讯云计算（北京）有限责任公司	WorkBuddy	V5.2.0
8	腾讯云计算（北京）有限责任公司	腾讯 QClaw	V0.2.30
9	腾讯云计算（北京）有限责任公司	腾讯云 ClawPro 平台	V1.0
10	华为技术有限公司	华为云码道 CodeArts 代码智能体	V1.0
11	阿里云（北京）科技有限公司	Apsara Copilot 运营数字人	V3.18.6
12	中信百信银行股份有限公司	中信百信银行 Orange 智能体平台	V1.0
13	国网信息通信产业集团有限公司	光明·智灵	V1.0
14	北京中科润宇环保科技股份有限公司	中科环保智慧运营智能体（办公助手、安健环专家、运营专家、培训专家）	V1.0

序号	企业名称	智能体名称	版本号
15	胜斗士（上海）科技技术发展有限公司	肯德基小 K 智能体	V1.3.1
16	金蝶软件（中国）有限公司	灵基（Lingee）	V1.0
17	用友网络科技股份有限公司	用友 YonWork	V1.0
18	远光软件股份有限公司	九天灵境 AI 原生平台	V1.0
19	南威软件股份有限公司	WellWork	V2.0.0
20	北京知道创宇信息技术股份有限公司	AiPy Pro	V1.2.2
21	万达宝软件（深圳）有限公司	Multiable 万达宝 Claw	V3.0
22	深圳传世智慧科技有限公司	TransClaw	V1.2.7
23	杭州智诊科技有限公司	wiseclaw 智能体平台	V1.1.2
24	云知声智能科技股份有限公司	智能文件审查应用	V3.4

登记信息查询网址为 ai.isc.org.cn，后续将滚动更新。

附录 2：互联网智能体应用实践

本部分选取国内不同行业、不同技术路线的典型落地应用，涵盖通信、餐饮、医疗、文旅、工业运维等多个领域，集中展示互联网智能体在场景适配、技术架构、协同模式与治理实践等方面的探索成果，可为行业同类项目建设提供参考借鉴。

（一）互联网智能体类

案例一：肯德基小 K 点餐智能体

针对数字化、快节奏消费场景下传统点餐流程操作繁琐等问题，胜斗士提出肯德基小 K 点餐智能体。该智能体依托互联网对外为消费者、第三方服务渠道提供点餐交互、订单执行、语音交互的智能化服务，通过大模型、多意图并行感知、RAG 与策略编排路由结合的技术架构，构建了从感知到执行的端到端服务闭环，推动点餐服务模式从“人找餐”向“餐适人”转型。

1.主要内容

来源：胜斗士（上海）科技技术发展有限公司

图 2 肯德基小 K 智能体架构图

一是构建自然对话式点餐入口，通过语音交互与意图理解技术，使用户可通过自然语言完成菜单查询、门店切换、发票开具等操作，简化传统点餐流程中的多次点击与手动查找环节，降低用户操作成本。**二是建立多意图理解与任务协同机制。**针对复杂用户指令交互难题，一方面拆解复杂输入并并行处理各类子任务，依托知识检索与推理能力完善多轮对话管控；另一方面通过场景状态协同简化交互流程，有效规避多轮对话错误，同时将自然语言转化为标准化结构参数、限定模型调用范围，提升智能体复杂任务的执行精度与稳定性。**三是建立系统弹性与持续优化机制。**依托异步服务引擎结合容器多实例部署实现弹性扩缩容与流量均衡，配套全链路指标监测、熔断降级策略抵御峰值流量冲击。同时支持菜单、配置在线热更新功能，

依托线上对话数据沉淀与灰度验证流程，形成可持续迭代的自优化闭环，保障系统稳定运行与功能持续升级。

2.实施成效

一是提升高峰时段服务响应效率。针对多轮对话响应时延问题，通过全链路架构优化与多线程并行协同处理，优先完成核心意图识别并异步回传结果，压缩全链路处理耗时，提升用户交互实时性；**二是解决菜品与活动内容频繁变更导致知识库滞后的问题。**建立知识库配置热加载机制，支持运营人员在后台完成更新后即时生效，无需停机发版，保障知识库内容与线上业务同步。**三是形成可复制的餐饮零售智能交互方案。**该方案系统框架具备迁移价值，意图理解与知识库等模块独立沉淀，新场景可直接组合复用。

3.创新价值

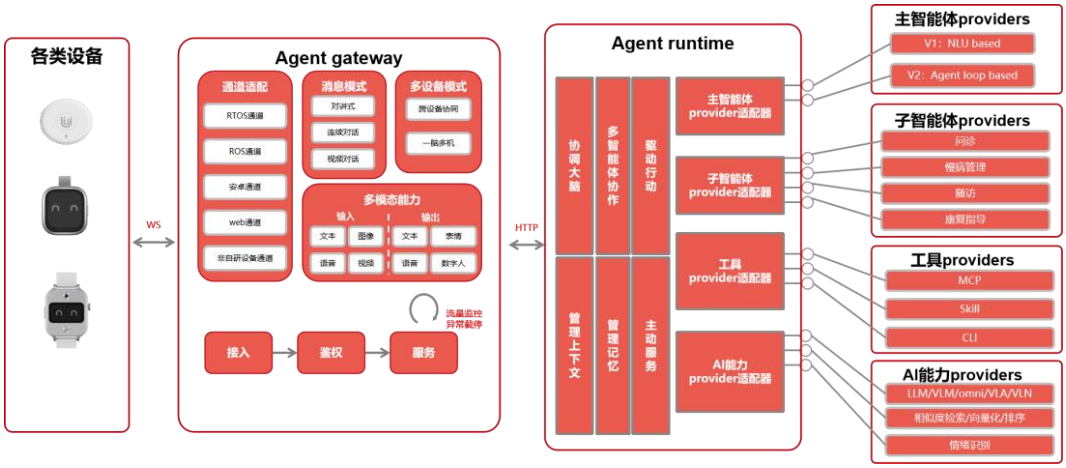
一是服务模式创新，将语音交互与大模型能力引入点餐场景，简化用户操作路径，在车载、忙碌等场景下提供更为便捷的点餐方式，推动点餐流程从用户主动查找向智能服务转变。**二是工程机制创新，**构建多意图并行感知、语义分解、状态化场景协同、弹性伸缩与动态更新等机制，形成面向高并发、高变更频率消费场景的智能体技术方案。**三是运营模式创新。**通过配置化运营管理模式与灰度发布机制，实现业务变更的灵活适配与风险可控迭代；通过智能交互过程沉淀用户偏好与复杂需求数据，为餐饮企业的菜单优化与营销决

策提供数据参考，形成面向消费服务场景的数字化运营支撑能力。

案例二：U 爱“小贝壳”医疗智能体

中国联通联合上海市第一人民医院聚焦胰腺肿瘤术后康复痛点，构建肿瘤 AI 随访与康复生态系统。项目以联通“小贝壳”智能终端为硬件入口，搭载胰腺专病问答智能体“龙医生”，依托互联网对外为术后患者提供营养评估、症状监测和康复指导等智能化语音交互服务。该智能体基于“快慢脑”协同双引擎与专病知识体系，构建患者自主管理、医护远程干预、健康数据联动的新型慢病随访管理体系。

1.主要内容



来源：联通（上海）产业互联网有限公司，中国联合网络通信有限公司上海市分公司，上海市第一人民医院

图 3 U 爱“小贝壳”架构图

一是构建低门槛陪伴式交互入口。针对老年患者数字鸿沟与术后求助不便的难题，通过智能终端与语音交互实现适

老化服务，支持方言自适应与动态意图解析，将专业医学术语转化为通俗表达，使患者能够以自然对话方式完成康复咨询、症状反馈与紧急求助。**二是打通诊后随访数据链路。**针对医护端随访工作量大与院外数据非结构化的痛点，一方面将院外随访、健康监测、异常预警和复诊管理串联贯通，实现诊后数据自动归档与结构化转化；另一方面通过运动轨迹与安全围栏等功能，辅助医护和家属掌握患者动态，为远程干预和连续管理提供数据基础。**三是构建“快慢脑”协同与专病知识体系。**引入双引擎架构，快脑处理高频即时交互与意图路由，慢脑负责复杂医疗推理与长程任务规划。同时，在云端集成多套模型进行分层调优，根据指令复杂度自动路由，兼顾响应速度与推理深度；在智能体内置营养评估、症状监测、康复指导三大标准作业程序，支持多核心意图与槽位的精准抽取，结合多轮对话记忆实现潜在需求洞察。

2.实施成效

一是完成医院专病随访场景落地。项目已在上海市第一人民医院胰腺中心成功落地，覆盖胰腺肿瘤术后患者群体，联合临床专家将胰腺专病诊疗指南转化为智能体可理解的知识库，打通院内 HIS 系统与云端平台，实现诊后数据自动归档。**二是提升随访效率和医疗数据质量。**系统一方面承担日常问答与数据采集工作，医生聚焦异常预警处理，随访效率显著提升；另一方面将非结构化的语音对话转化为结构化

病历，为科研与临床决策提供数据支撑。**三是形成“终端+AI服务+人工干预”的医养服务闭环。**项目构建覆盖诊前、诊中、诊后的服务闭环，围绕慢病随访与康复管理，形成终端部署、AI 随访、人工干预相结合的服务模式，并衔接互联网医院与康复管理服务资源，探索可持续运营路径。

3.创新价值

一是专病管理模式创新。以智能终端为硬件入口，搭载专病问答能力，形成患者自主管理、医护远程干预、健康数据联动的慢病随访管理体系，将专病知识、智能问答、健康监测和远程干预融合贯通，使患者居家康复过程能够被持续感知和及时响应。**二是适老化交互创新。**采用纯语音交互与多模态感知范式，实现专业医学术语向通俗语音的实时转化，降低老年用户使用智能医疗服务的门槛。**三是慢性病场景复用创新。**底层架构通用，意图理解与知识库等模块独立沉淀，更换专病知识库即可快速适配不同科室与病种，具备向更多慢病管理场景推广的基础。

案例三：山西文旅综合智能体

针对山西省文旅资源布局分散、景区服务信息碎片化等问题，中移九天构建山西文旅综合智能体。该智能体依托互联网对外为游客提供文博讲解、客流分析、票务查询等文旅服务，通过 ANS 智能寻址和智能体网关技术，联动景区讲解、交通出行等多类行业智能体，实现全域文旅资源的统一

调度与协同管理。

1.主要内容

一是构建“统一入口+多能力协同”的文旅智能体架构。围绕游客服务、景区运营、文博科普等核心场景需求，以 DID 身份可信技术为底座，通过统一入口集成识图模型、客流分析、票务查询等多类 AI 工具，完成跨协议、跨平台的标准化转换，降低跨域能力调用壁垒，实现文旅服务智能化协同。**二是实现跨域互通和可信调用。**依托物理网络底座，融合 ANS 智能寻址、AI 能力全域融通、DID 身份可信等核心技术，构建智能体互通、管控、运营一体化协同服务体系。一方面，部署统一 AI 网关，支持多协议自动转换与智能路由调度，保障跨域高效互通；另一方面，支持认证鉴权、运行管控与全链路观测能力，保障调用安全合规。

2.实施成效

一是完成北齐博物馆景区试点验证。该方案已在北齐博物馆景区完成试点运营，面向公众开放服务，并在行业展会中进行技术演示，验证了跨域调用、全链路可观测、安全管控等核心功能。**二是提升用户接入效率和平台运行稳定性。**在用户侧，平台支持一点接入、全域互通，降低跨地域、跨行业的能力调用成本。同时，通过智能寻址与路由调度技术可优化传输时延与流量负载，保障业务稳定运行。**三是构建可观测、可审计、可溯源的安全运营体系。**在运营侧，平台

基于 DID 分布式体系实现身份可信认证，通过事前审核、事中管控、事后溯源的全流程管理，提升调用过程的安全可控水平，保障业务合规与数据安全。

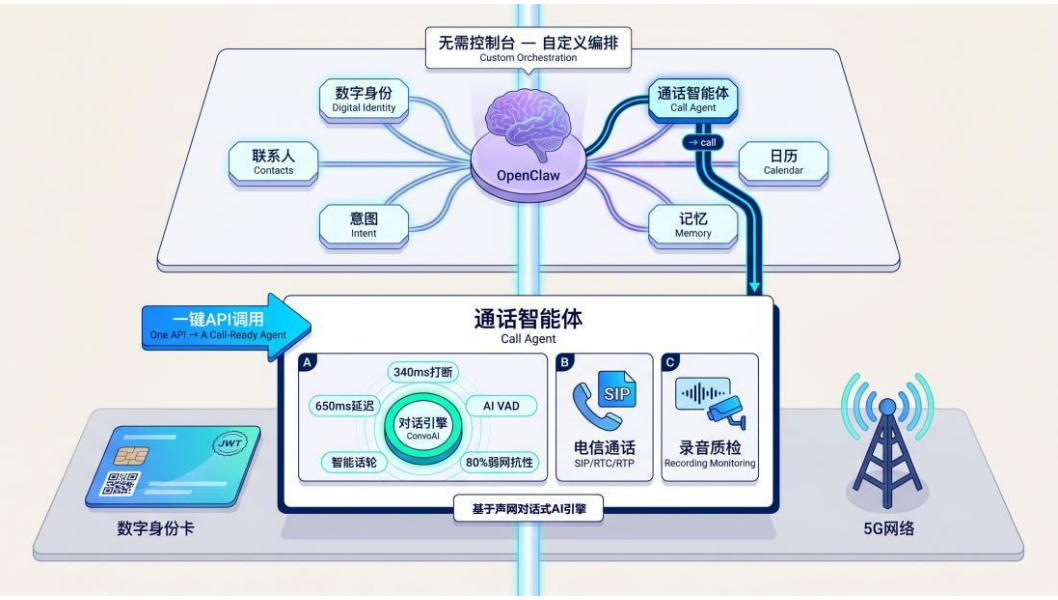
3.创新价值

一是架构模式创新，相较于传统景区独立部署的智能应用，通过多智能体协同架构实现跨区域、跨行业的智能体互联互通与智能路由调度，缓解点对点全连通带来的复杂性与安全风险，形成可复制的“智能体+文旅”落地参考。**二是跨行业扩展价值。**方案中的智能路由、可信身份、全链路观测能力具有通用性，可为智慧交通、智慧城市、智慧教育等需要多智能体跨域协同的行业提供参考和复用基础。

案例四：基于可信数字身份的对话式智能体

针对实体语音通话场景下数字身份缺位等问题，华为联合声网构建基于可信数字身份的对话式智能体。该智能体通过电话号码经电信网络接入互联网，对外提供代接代打、外呼回访、智能客服等服务。凭借华为电信级可信数字身份凭证，智能体以独立通信主体获得网络准入资格，解决监管侧无法识别、溯源、吊销智能体难题，有效防范骚扰诈骗等滥用风险。

1.主要内容



来源：华为技术有限公司、上海声网科技有限公司

图 4 基于可信数字身份的对话式智能体通话解决方案架构图

一是构建电信级可信身份体系。基于 W3C VC 标准扩展定义“Agent Verifiable Credential”凭证类型，由运营商网络为各智能体签发数字身份凭证 Agent Digital VC，且每份凭证绑定一个真实实名制手机号，实现智能体入网可管可认。**二是集成标准化呼叫智能体能力。**集成低延迟对话 AI 引擎，整合通话交互、录音管控、降噪、弱网适配等音视频能力，通过统一 API 快速部署语音交互智能体，简化底层组件联调。**三是引入 OpenClaw+Skills 编排范式。**将身份申领、通话、通讯录查询等功能封装为可组合 Skills，依托标准化接口自动完成智能体创建、号码绑定、电话外呼，全流程无需人工后台配置。

2.实施成效

一是覆盖双向通话业务场景。主叫可指令智能体主动外呼并推送通话摘要，被叫可代接来电、拦截骚扰并留存通话记录。实现全天候自主通话服务。**二是形成可落地的 AI 通信安全治理模式。**该方案可代为处理各类语音通信事务，提升沟通效率。业务操作以用户主动授权为基础，通话全程标注 AI 主体身份，保障业务合规运行。相关能力已联合运营商完成试点验证，具备规模化落地基础，为 AI 通信领域提供一套可落地、可复用的安全管控实施路径。

3.创新价值

一是身份机制创新。构建可识别、可监管、可吊销、可审计的 Agent Digital VC 体系，将智能体纳入电信级身份管理，为 AI 接入实体通信网络提供可信底座。**二是能力封装创新。**整合语音大模型、语音识别、线路资源、降噪管控等模块，支持 API 快速部署通话智能体，简化通信系统集成开发。**三是安全自治机制创新。**依托用户授权、AI 身份标识、通话存证审计、凭证生命周期管理多重管控手段，防范语音仿冒、骚扰诈骗等风险，为各类通信 AI 场景提供可复制安全实践。

（二）互联网智能体相关组件类

任务编排类：

案例一：多智能体集群与联邦协同平台

面向复杂任务场景中单一智能体能力受限、多智能体协同上下文割裂且任务接续困难等问题，中国电信研究院打造了多智能体集群与联邦协同平台。该平台以 AgentOS 为本地集群运行底座，以 AgentFed 为云端联邦协同平台，提供任务规划、工具调用、工作流记忆沉淀与全局上下文共享能力，实现了跨设备、跨人员、跨场景的任务接续与协作经验复用，为长周期、多角色协同任务提供了可推广的架构范式。

1.主要内容

图 5 AgentOS 系统架构及 AgentFed 平台架构图

来源：中国电信股份有限公司研究院

一是构建本地多智能体集群能力。以 AgentOS 为底座搭建规划、路由、执行三层架构，分别负责任务拆解、工具匹配、执行反馈，组织多专业智能体分工协作，实现复杂任务多角色协同与多任务并行调度。**二是建立工作流记忆与联邦式协同架构。**依托 AgentOS 沉淀用户操作、执行链路与协作过程，形成可复用工作流记忆，支撑同类任务经验复用；通过 AgentFed 云端平台接入多类边缘节点，提供共享工作空间与全局上下文能力，实现跨机器、跨地域、跨人员协同。**三是引入人机协同与开放互联机制。**通过 A2H 机制在需要人工参与的环节发起确认和授权，在提升自动化水平的同时保留关键任务中的人工控制权。同时，平台支持异构算力部署，兼容国产主流芯片，提供本地与云端双模式部署。

2.实施成效

一是研发及项目交付场景已投入使用。在中国电信内部，基于 AgentOS 串联需求理解、代码生成、测试验证、文档沉淀全研发环节，减少人工切换与跨角色沟通成本；依托 AgentFed 打通多成员、多终端的任务上下文，降低项目交接成本。**二是形成可复制的平台化部署与实施流程。**平台部署遵循本地 AgentOS 节点部署、业务工具与知识源

接入、工作流记忆沉淀、AgentFed 联邦接入的分步流程，可快速适配不同团队，将分散智能体组网，具备向政务、制造、金融等行业推广的基础。

3.创新价值

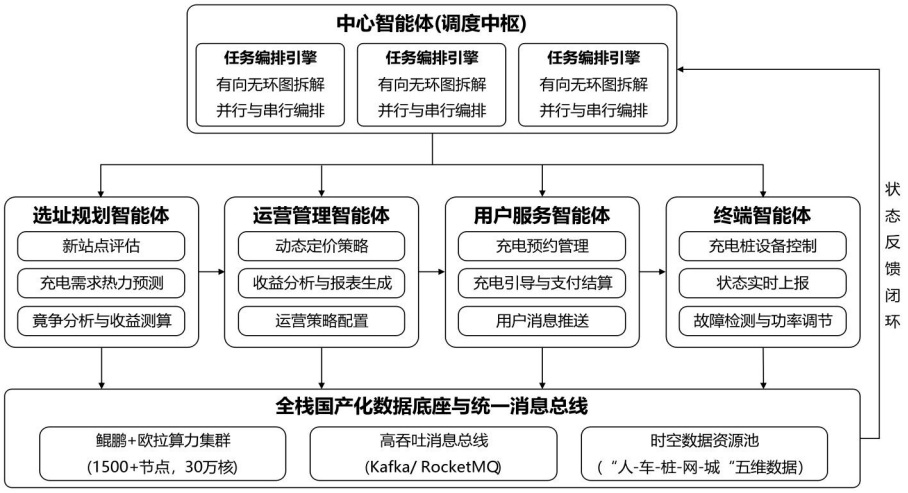
一是架构模式创新，推动智能体应用从单体工具走向集群协同。本方案以 AgentOS 构建本地多智能体集群，以 AgentFed 打造云端联邦协同层，形成“本地集群+云端联邦”的端云协同架构，为复杂任务的多主体协作提供标准化技术方案。**二是协同机制创新。**实现协作经验自动沉淀与复用，通过共享工作空间与全局上下文破解跨人员、跨设备协作的上下文割裂难题，可向研发协作、智能运维等领域拓展。**三是基础设施创新，**为智能体协作提供国产化部署方案。平台支持异构算力部署与国产芯片兼容，通过统一接入、标准化编排与开放工具生态，构建国产化多智能体协同基础设施。

案例二：基于多智能体任务编排的充电桩规划系统

面向充电设施选址规划中多源数据难融合、跨环节协同效率低等难题，广西移动构建了基于多智能体任务编排的充电桩规划系统。该系统以中心智能体为调度中枢，通过通用编排调度+插件化专业能力的分层解耦架构，将任务拆解与依赖管理从选址算法、负载预测等行业专用逻辑中

分离，实现了从需求预测到运营优化的全链条闭环，使同一套编排框架从电力领域延展至应急、文旅、政务等多个行业，验证了任务编排系统跨领域复制的可行性。

1.主要内容



来源：中国移动通信集团广西有限公司

图 6 充电桩规划系统架构图

一是构建“中心调度、专业执行”的分层解耦编排架构。中心智能体作为唯一调度中枢，维护全局任务队列与有向无环图依赖关系，负责拆解复杂目标为可执行的子任务。各专业智能体以插件化方式注册能力接口，聚焦各自领域，接收调度指令协同完成任务。**二是建立基于多目标优化的动态优先级调度机制。**综合任务紧急度、资源占用等因素计算优先级，高优先级任务可抢占资源优先执行，优化系统资源分配效率，提升关键任务的响应速度与执行时效。**三是采用基于发布订阅的事件驱动与状态同步机制。**通过消息总线实现智能体间异步通信，任务状态变更以事件形

式发布，各智能体按需订阅，实现任务全生命周期可观测。

2.实施成效

一是完成落地验证。在系统在广西电网落地部署，通过自然语言交互生成选址报告，替代传统人工实地勘察与数据核验的繁琐流程，有效缩短规划周期。上线后解决了多设备协同响应慢、消息总线拥堵、任务数据依赖断裂等落地过程中的实际协同难题。**二是实现跨行业规模化应用。**基于统一数据底座与容器化部署能力，系统已从电力行业拓展至应急、文旅、政务、环保等领域多家应用单位，验证了任务编排框架的行业通用性。

3.创新价值

一是以全链条闭环替代传统环节分离模式。传统充电设施管理中规划、建设、运营各环节独立运作、数据割裂，本系统将多环节纳入统一编排框架，形成“规划-建设-运营”一体化闭环，解决各环节间协同脱节导致的资源错配与效率损耗。**二是以安全加密机制支撑关键基础设施智能运营。**全链路采用国密加密传输，数据可用不可见，通过安全评估审核，在实现多智能体高效编排的同时保障数据安全与操作合规，为能源、公共安全等关键基础设施的智能化改造提供安全可控的技术方案。

案例三：基于多智能体协同的施工方案生成与

审核系统

面向电网基建施工方案编制依赖人工、跨专业协同耗时长、受力计算易出错等痛点，陕西电信联合国网陕西电力构建了施工方案多智能体协同生成与审核系统。该系统以文档主智能体为编排中枢，通过大模型统筹理解、专用模型精准执行的双轨协同机制，将图纸语义解析、力学参数计算等专业任务交由专用小模型及计算引擎执行，计算结果经大模型组织成文，实现了从数据解析到方案输出的闭环自动化流水线，为知识密集型行业的任务编排系统建设提供了可参照的实施路径。

1.主要内容

一是建立多智能体闭环编排流水线。以“文档主智能体”为全局调度核心，编排 16 个专业智能体形成参数提取、领域计算、方案生成、形式与规则审查、成果导出的闭环工作流。各智能体间通过异步消息队列通信，支持任务依赖管理、异常重试与并行执行，实现复杂多工序的无缝协同。**二是构建大小模型双轨协同机制。**采用大模型统筹理解、专用小模型精准执行的双轨架构。以行业大模型作为顶层调度中枢，负责自然语言交互、需求意图解析与任务拆解；底层挂载多类深度定制的专用小模型，分别负责图纸语义解析、力学计算等高精度执行任务。**三是打造多源异构数据融合与知识底座。**整合非结构化文档、设计图纸及复杂

公式等多模态数据，构建电力基建领域知识图谱与 RAG 检索库，实现跨文档、跨模态的语义对齐与精准检索。**四是突破施工图纸智能识别与吊装决策能力。**针对铁塔工程图纸识别难题，构建由图纸基础解析、图文语义理解、工程对象识别、关键参数测量与辅助决策等模块组成的识别智能体，实现从 PDF 底层矢量线段提取到工程拓扑语义理解的跨越，为吊装施工提供决策支撑。

2.实施成效

一是完成特高压输电线路试点验证。系统已在多条特高压输电线路工程中完成试点，针对复杂工程图纸解析精度不足、大模型在多场景专业计算中产生错误等核心难题，通过定制化图纸语义解析模型与大小模型协同机制逐一攻关，实现施工方案自动化编制与审核全流程跑通。**二是形成三层解耦架构支撑跨场景迁移。**系统实现 AI 算力、业务规则、知识资产三层解耦，底层多智能体调度框架与图纸多模态解析引擎可迁移至配电网建设、变电站施工等其他电力基建领域，为电力基建数字化转型提供可复刻的实施范式。

3.创新价值

一是以多智能体闭环替代人工串行协作。传统方案编制需要跨专业人员分散处理、逐步交接，本系统将参数提取、力学计算、方案生成、合规审查等环节编排为自动化

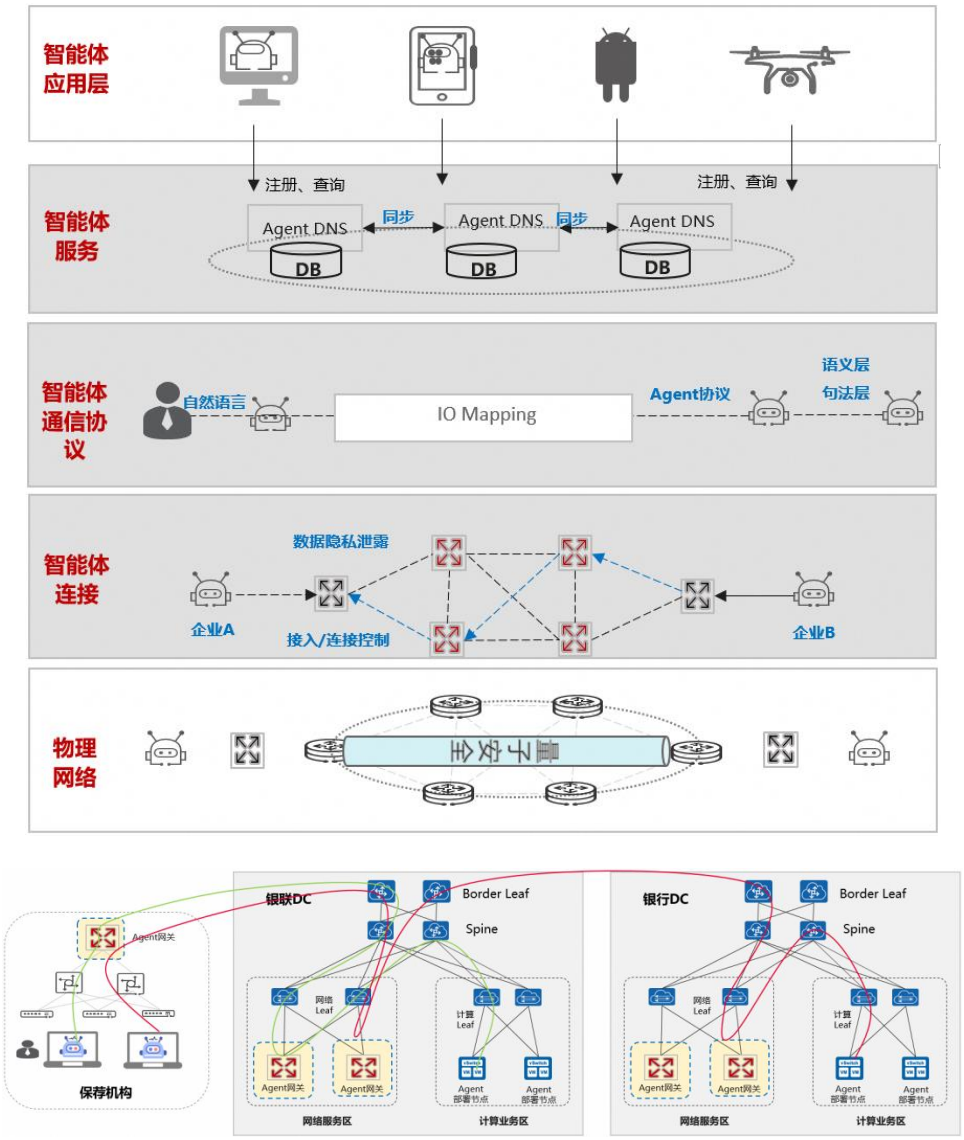
流水线，各智能体异步协同、并行推进，将施工经验固化为可复用工作流，推动方案编制从人工经验驱动向智能体协同驱动转变。**二是以知识资产沉淀替代个人经验依赖。**将分散的行业规范、历史方案、设计图纸整合为知识图谱，实现隐性知识的显性化与可检索化。通过标准化的“人机协同”工具模式，将方案编制从对资深人员的经验依赖转化为对标准化智能工具的调用，提升编制质量的稳定性与可复制性。

智能体网关类：

案例一：金融业智能体可信互联技术方案

面向金融行业智能体互联中的碎片化难题，以及其对身份可信、指令可控、数据安全的更高要求，中国银联联合华为研究设计了金融业智能体可信互联技术方案。该方案以智能体网关为核心枢纽，提供智能体注册发现、语义寻址、异构协议自动转换与端到端安全通信能力，实现了跨机构智能体的统一接入、可信认证与高效协作，为开放银行生态下的智能体互联提供了可参考的技术蓝本。

1.主要内容



来源：中国银联股份有限公司、华为技术有限公司

图 7 金融业智能体技术架构图及网关部署示意图

一是构建五层智能体可信互联技术架构。方案分为智能体应用层、服务层、通信协议层、连接层及物理网络层，由智能体注册服务中心（含智能体网关）统一承载。服务层提供智能体注册发布与语义寻址功能，通信协议层支持异构协议自动转换与任务追踪管控，连接层与物理网络层保障底层连接质量。**二是实现智能体身份可信与能力寻址。**

通过智能体网关提供身份凭证签发、更新及注销管理服务，支撑端到端双向身份认证；通过语义分析、技能模糊查询、协同推荐等技术实现智能体发现与寻址，并生成 Agent 转发表指导跨机构业务数据转发。**三是提供金融级安全与合规保障。**该方案在通信连接建立前对双方身份进行核验，利用私钥签名对通信链路进行端到端安全保护，提供防私接、防仿冒、可追溯、敏感信息泄露检测的可信通信能力，满足金融行业身份可信、指令可控、数据安全的监管要求。

2.实施成效

一是完成 IPO 资金流水核查场景原型验证。在原型验证的保荐机构、银联和多家银行的多对多场景中，可实现跨机构智能体的可信互联与高效协作。**二是实现智能体注册、路由、任务编排全流程打通。**以智能体网关为枢纽，各机构智能体统一注册后，网关同步注册信息并生成 Agent 转发表。用户请求经网关语义分析与智能路由，转至目标智能体进行任务拆解编排，各机构智能体可按编排协作执行任务。**三是形成可参考技术方案并向其他场景拓展。**该技术方案具有可行业参考性，并正在推进智慧医疗场景验证，通过社区筛查智能体、总院智能体、隐私智能体、随访智能体协同，实现总院和社区卫生服务中心高效安全协同。

3.创新价值

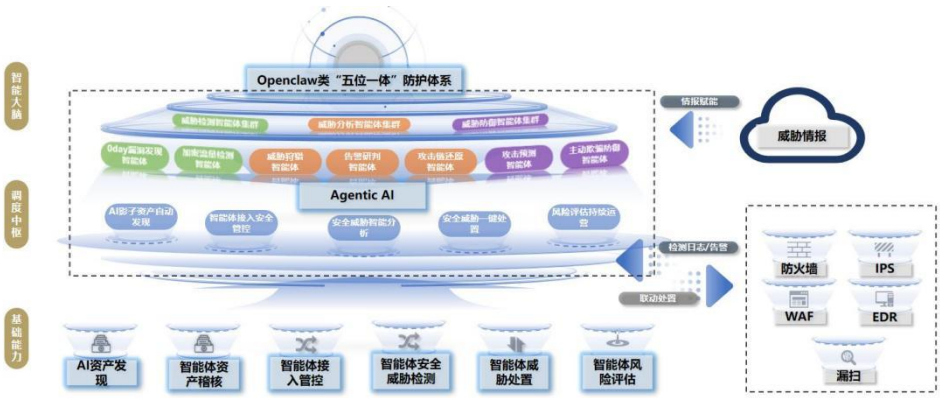
一是打通金融业智能体协作链路。通过智能体网关实现身份认证、能力注册与语义寻址等功能，支持异构协议自动转换，解决金融业智能体“碎片化”互通难题，为开放银行生态下跨机构协同提供范式。**二是满足金融级安全合规要求。**内置数字证书签发、权限管控、通信加密等机制，确保身份可信、指令可控与数据安全，为金融智能体可信互联体系建设提供参考。**三是具备多场景拓展潜力。**该方案支持多智能体注册、路由转发、任务分解及安全交互全流程，可推广至联合授信、反欺诈、智能风控等金融业务领域，同时可向智慧医疗等跨行业场景拓展。

行为观测类：

案例一：智能体“五位一体”观测治理系统

面向智能体规模化部署后权限边界宽、攻击面扩大、供应链投毒难检测、传统边界防护无法感知智能体内部行为等安全问题，湖南移动联合启明星辰构建了智能体“五位一体”观测治理系统。该系统围绕资产可见、接入可控、威胁可析、风险可处、运营可持续五大能力，以安全智能体为调度中枢自动编排终端、网络、身份等安全能力协同响应，实现了智能体行为从可观测到可处置的能力闭环，为应对智能体新型安全风险提供了系统性防御方案。

1.主要内容



来源：中国移动通信集团湖南有限公司，北京启明星辰信息安全技术有限公司

图 8 智能体“五位一体”观测治理系统架构图

一是构建“五位一体”闭环防护架构。资产可见层通过流量指纹与终端行为识别未受控智能体；接入可控层部署安全网关，精细授权并拦截恶意投毒；威胁可析层融合语义与行为双引擎，检测注入攻击、提示词泄露等手法；风险可处层联动终端响应，实现文件全生命周期扫描与异常隔离；运营可持续层构建风险画像与动态评分模型，持续评估风险与合规性。**二是建立 Agentic AI 驱动的自动化处置编排机制。**以安全智能体为调度中枢，打通终端安全、网络准入、身份管理等安全能力间的壁垒，当检测到高危行为时自动编排多类安全能力协同响应，支持多智能体协作完成防护。**三是实现轻量化部署与生态对接。**将防护能力封装为可插拔的 Skill 与 Plugin 组件，用户通过命令或对话即可完成安装，同时联动云端威胁情报中心获取检测能力。

2.实施成效

一是完成内部核心智能体业务全面防护。体系已在湖南移动内部投入实际运营，全面接入编标、公文、知识助手等核心智能体应用。通过资产发现能力消除内网未受控智能体实例隐患，拦截提示词注入攻击与恶意安装行为，安全事件平均处置时间与告警处置量显著下降。**二是实现行业推广与跨场景验证。**体系已面向多家重点客户完成智能体安全防护落地验证，在某省级政务平台应用中成功识别并阻断针对智能体的注入攻击，避免敏感信息泄露风险。联合合作伙伴发布国产化智能办公终端方案，为各行业提供一站式智能体安全能力，部分检测规则已通过安全实验室开源共享。

3.创新价值

一是从单点防护转向全生命周期闭环治理。该系统将资产发现、接入管控、威胁检测、自动处置、持续运营集成为一体，覆盖智能体从上线到复盘的全生命周期。**二是以安全智能体编排打破能力壁垒。**该以安全智能体为中枢自动编排各安全能力协同响应，将人工协调转化为智能编排，推动安全运营从被动响应向主动防御转变。**三是以开源共享推动智能体安全生态建设。**将核心检测规则与最佳实践通过安全实验室开源发布，并联合发布标准，为金融、能源、政务等行业输出可复用的智能体安全治理方案。

云计算开源产业联盟

中国信息通信研究院