香 港 大 學
THE UNIVERSITY OF HONG KONG

KYA: Know Your Agent

智能体信用体系研究报告

香港大学KYA课题组

2026年11月



编委会

汪　扬　　林　晨

卡比内·科马拉

陈卡你　　杨　宏

陈　浩　　顾凌云

课题组

缪立璋　　柏　亮　　刘奕然

曹艺凡　　韩昱子　　马文婷

唐晓玮　　胡　攀　　张卓尔

感谢蚂蚁集团对 KYA 研究的支持！

目录

绪论.....1
第一章 从 KYC 到 KYA——智能体时代信任新挑战.....4
1.1 智能体经济崛起与信任真空.....4
1.2 信任缺位的经济代价.....9
1.3 从 KYC 到 KYA.....20
第二章 智能体成熟度模型 L0-L5.....27
2.1 成熟度模型的逻辑与判定方法.....27
2.2 L0 非智能体：规则执行基线.....30
2.3 L1 人工确认：智能体获得建议权.....30
2.4 L2 人类锚定信用：智能体获得受限执行权.....32
2.5 L3 混合信用：智能体获得任务自治权.....35
2.6 L4 独立信用：智能体获得受限经济代理权.....38
2.7 L5 智能体经济：从单体成熟度到网络成熟度.....40
2.8 升级路径、信用飞轮与动态降级.....43
2.9 本章小结.....44
第三章 KYA 标准框架——智能体信用三维度.....46
3.1 身份可信（智能体身份确权）.....46
3.2 信用可信（履约历史与信用评估）.....50
3.3 权限可信（授权范围与操作边界）.....55
3.4 三维度协同机制.....58
第四章 实践验证.....62
4.1 蚂蚁数科：“智能体超级工厂”的 KYA 探索.....62
4.2 Stripe：智能体经济布局及 KYA 实践.....65
4.3 芝麻信用：KYA 探索与智能体信任基建.....71
4.4 冰鉴科技：构建金融级可信智能体.....74
4.5 Ryt Bank：全球首家 AI 原生银行的智能体信任实践.....78
4.6 英和智临：校园智能体实践与本地化信任底座.....81

4.7 林洋能源：构建“可依赖”的智能体信任机制 85
4.8 麦迪卫康：把医生信用铸进医疗智能体 89
4.9 GCSA：面向智能体生态的安全防护与攻防治理 93
第五章 智能体信用评估框架 99
5.1 评估方法论与设计原则 99
5.2 基础指标体系 102
5.3 场景建议 111
第六章 智能体发展指数 115
6.1 基本原理 115
6.2 指标体系 116
6.3 指数结果及呈现 121
6.4 指数解读 124
6.5 KYA 建设启示 126
指数公开资料来源 128
第七章 展望：从智能体经济到信任资本时代 131
7.1 智能体经济的四个阶段：从执行任务到创造经济 131
7.2 时代级机遇：从制造智能体到建设智能体经济 133
7.3 信任资本：人工智能时代的资本形成与经济增长 135
7.4 如何创造和积累智能体信任资本 138
参考文献 142
附录I 智能体信用体系问卷报告 145
摘要 145
1 执行摘要 145
2 数据与方法 145
3 数据分析 146
4 综合性总结 186
附录 开放建议主题说明 189
版权与免责声明 190

绪论

没有 KYA，就没有智能体经济——这是我们在广泛的研究和深入调研基础上得出的结论。

智能体的概念和技术探索由来已久，直到 2025 年进入可用阶段，2026 年开始爆发。随着智能体进入交易环节，“智能体经济”的宏大未来逐步打开。而从智能体应用到智能体经济，中间不可逾越的鸿沟，是智能体信任，它需要KYA（Know Your Agent）来解决。

人工智能与现实经济活动之间的关系正在发生结构性变化：AI 不再只是被人使用的软件工具，而开始成为能够执行任务、配置资源并参与交易的数字化行动主体。

这一变化正在企业、金融、消费、支付和公共服务等领域迅速展开。越来越多企业开始部署能够自主完成特定任务的智能体，金融机构开始研究智能体支付和自主金融服务，支付网络开始建立面向机器交易的新型基础设施，消费者也开始将 AI 从信息获取工具转变为购物、旅行、办公和生活服务的代理者。

与此同时，智能体自主行动所带来的身份、授权、安全、责任和信用问题开始集中出现。

随着智能体进入真实经济活动，决定其能否大规模应用的因素逐渐从“能够完成什么”转向“能够被允许完成什么”以及“其他主体是否愿意与其合作和交易”。智能体经济要实现持续、稳定和大规模发展，不仅需要更强的模型和更完善的基础设施，也需要一套能够识别智能体身份、确认其归属和授权、评估其行为与信用并明确责任边界的信任机制。

本报告所讨论的 KYA（Know Your Agent），正是在这一背景下产生的。KYA 并非传统 KYC 在智能体领域的简单复制，而是针对智能体成为经济活动主体之后出现的新型身份、授权、信用和责任问题，建立相应的识别与信任机制。

智能体经济的主要约束正在由技术能力转向信任体系。而当智能体进入支付之后，信任问题直接转化为商业问题。

根据我们对从业者的问卷调查报告（以下称《问卷报告》，见附录），76%的受访者在过去一年频繁（23%）或偶尔（53%）遇到信任、安全、问责或合规方面的顾虑，仅 2%从未遇到，信任摩擦已经常态化。

Visa 委托 Harris Poll 于 2026 年 5 月 26 日至 28 日在美国开展的一项消费者调查显示，23%的受访者信任生成式 AI 代表自己处理支付交易；在另一项针对具体品牌的提问中，61%的受访者表示信任 Visa 处理智能体交易。

Cloud Security Alliance 与 Anjuna 对全球金融服务业开展的调查显示，62%的受访机构已经部署 AI 智能体，其中 93%的机构已授予其智能体一定程度的自主权限。85%的受访者预计未来 AI 智能体将能够代表消费者发起和执行支付交

易，65%认为这一模式需要新的授权机制。此外，Accenture 2025 年《Future of Money》研究发现，87%的金融机构受访者认为，信任将是智能体支付普及的最大障碍。

来自交易需求侧和供给侧的数据，结论高度一致：智能体支付具有明确的市场需求，但现有身份与授权机制尚不足以支撑其大规模发展。

KYA 的目标，是安全地释放智能体经济的规模。

智能体经济需要一种机器可验证的信任机制。KYA 的商业价值由此可以从三个层次理解：KYA 使原本无法发生的交易能够发生，KYA 使交易能够以更低成本完成，KYA 使商业活动具备机器规模化的条件。

从企业部署、金融应用、消费者使用和机器支付等多个维度观察，智能体正在由一种新型软件能力逐渐转变为经济活动中的行动主体。企业开始大规模部署智能体，金融机构已经赋予相当比例的智能体自主权限，支付网络开始建设面向机器交易的新型基础设施，消费者 AI 用户已经形成数亿级规模。

这些变化共同表明，智能体经济已经从技术概念进入产业化早期阶段。与此同时，智能体安全事件、未知智能体、权限管理、未授权访问以及消费者对 AI 支付的有限信任，也说明现有的身份和信任体系并未完全适应这一变化。

KYA 所面对的，正是这一制度与商业基础设施的空白：一套使智能体能够被社会和市场识别、授权、评价、信任和追责的经济基础设施。《问卷报告》显示，88%的受访者认为行业需要统一的智能体可信评估或认证标准，其中 39%认为极其需要、应当紧急建立。

智能体经济是否能够从技术创新演变为新的经济运行模式，最终取决于这一基础设施能否与智能体能力同步成长。技术能力决定智能体经济的可能边界，信任体系决定其现实边界；而 KYA 所要解决的，正是从“技术可能”走向“商业可行”之间最关键的一段距离。

本报告围绕“如何信任一个智能体”——这一智能体经济发展的基础问题，完成了从问题到框架的完整论证。第一章从信任真空与信任缺位的经济代价出发，说明产业发展的瓶颈已经从“能力”转向“信任”；第二章提出以“可委托自主权”为核心的 L0－L5 成熟度模型，刻画智能体从规则执行到网络化经济关系的六级授权状态；第三章将 KYA 落实为身份可信、信用可信与权限可信三个维度；第四章通过一系列案例呈现智能体在经济场景中应用及不同企业对 KYA 的探索；第五章在此之上构建了三支柱、十三维度、六十五项指标的基础指标体系，并配套场景化评估与关键门槛机制；第六章提出以智能体发展指数持续跟踪行业演进；第七章我们展望了 KYA 的商业价值和社会价值，落脚于智能体信任资本的创造和积累。

贯穿全报告的核心判断是：能力不等于可信。模型能力的提升不会自动转化为商业上的可委托性；只有当身份可确权、履约可核验、权限可约束时，智能体才能获得与其能力相匹配的经济授权。信任机制因此不是智能体应用的附加合规要求，而是决定智能体经济市场规模边界的基础设施。本报告提出的成熟度模型、

三维度框架与指标体系，旨在为这一基础设施提供可讨论、可检验、可迭代的公共起点。

本报告仍是一项进行中的研究。指标体系的权重与阈值有待试点数据标定，智能体发展指数有待接入持续的数据来源，案例覆盖的行业与场景也有待扩展。课题组将随产业实践同步修订框架，欢迎各方对报告内容提出指正。

第一章 从 KYC 到 KYA — —智能体时代信任新挑战

1.1 智能体经济崛起与信任真空

2022 年生成式人工智能（Generative AI）的爆发，让世界第一次认识到大模型能够像人一样理解和生成语言，而 2025 年以来人工智能发展的核心变化，则是机器开始具备了自主执行任务（Agentic AI）的能力。人工智能正在从“能够回答问题”走向“能够完成工作”，从一种信息处理工具，逐渐演变为能够参与经济活动的新型主体。

这一变化意味着，全球人工智能产业已经进入一个新的发展阶段——智能体时代。

1.1.1 人工智能的发展正在经历一次范式转换

过去十余年，人工智能的发展大致经历了三个阶段。

第一个阶段是感知智能（Perception AI）。这一阶段的人工智能主要依靠深度学习技术，在图像识别、语音识别、机器翻译等领域取得突破，其核心能力是“看见”“听懂”和“识别”。计算机视觉和语音识别技术的大规模商业化，都属于这一阶段的重要标志。

第二个阶段是生成式人工智能（Generative AI）。以 ChatGPT（GPT-3.5）于 2022 年 11 月正式发布为标志，大语言模型开始展现出前所未有的自然语言理解和内容生成能力。随后，GPT-4、Claude、Gemini 等模型不断刷新能力边界，人工智能开始广泛应用于写作、编程、设计、教育、科研等领域。生成式 AI 极大提升了知识工作的效率，也推动全球进入大模型竞争时代。

生成式 AI 的核心能力仍然是生成内容（Generate）。无论是回答问题、撰写报告还是生成代码，仍然需要由人类决定下一步行动，并负责执行最终任务。

第三个阶段，则是当前正在快速演进的智能体（Agentic AI）时代。

智能体与传统聊天机器人最大的区别，在于是否具备自主行动能力（Agency）。智能体能够围绕目标制定计划、调用外部工具、访问互联网、执行程序、管理文件、协调多个任务，并根据执行结果持续调整行动路径。它不只是“提供建议”，而是可以“完成工作”。

1.1.2 智能体产业进入快速发展阶段

智能体的概念早在 1990 年代就明确出现并逐渐形成理论，并在 2010 年代经过深度学习的迭代，直到 2022 年之后，在大模型的基础上，智能体发展进入新的阶段。真正推动智能体进入产业化阶段的是 2025 年以来基础模型能力、工具调用能力以及企业应用需求的同步成熟。

一方面，大模型推理能力显著提升，能够支持更加复杂的任务规划和长链推理；另一方面，浏览器控制、代码执行、API 调用、多智能体协同等能力逐渐成熟，使智能体真正具备了执行现实工作的基础条件。

从产业竞争格局来看，智能体已经成为全球 AI 产业最重要的发展方向之一。Gartner、IDC、麦肯锡、德勤等国际研究机构均将智能体列为未来几年人工智能最重要的发展趋势。

麦肯锡在 2025 年 6 月发布的报告《Seizing the Agentic AI Advantage》中指出，当前企业生成式AI应用普遍存在“浅层部署”问题：大量AI仍以Copilot、聊天机器人等形式附着于既有工作流程之外，主要承担信息检索、内容生成和任务辅助等工作，而企业的核心业务流程本身并未发生根本改变。麦肯锡认为，Agentic AI 真正具有突破性的地方，并不是简单地将智能体嵌入现有流程，而是以智能体为核心重新设计业务流程，重新分配人与机器的职责，使 AI 从“辅助工具”转变为能够自主规划、协调、执行和适应的业务参与者。

在麦肯锡发布报告之后的这一年，智能体已经快速地从“浅层部署”向深度延伸。

1.1.3 智能体已经开始进入真实经济体系

从产业扩散速度来看，智能体的发展速度明显快于互联网和移动互联网。

根据 Gartner 于 2025 年 8 月的预测，当年真正集成任务型智能体的企业应用比例仍不足 5%；而到 2026 年底，这一比例预计将提升至 40%。这意味着，仅一年左右时间，企业应用中智能体的渗透率便将实现数倍增长。

进入 2026 年以后，这一趋势进一步加速。Gartner《2026 Hype Cycle for Agentic AI》援引其“2026 CIO and Technology Executive Survey”称，已有约 17%的企业正式部署智能体，超过 60%的企业计划在未来两年内部署智能体系统。如此陡峭的采用曲线，在新兴数字技术中并不多见。

更值得关注的是智能体数量本身的增长速度。Gartner 预测，到 2028 年，一家全球《财富》500 强企业平均在使用的智能体将超过 15 万个，高于 2025 年的不足 15 个。按这两个阈值估算，平均数量将增长超过 1 万倍，即超过四个数量级。

目前，智能体已经广泛进入软件开发、企业办公、客户服务、金融分析、数据分析、科研辅助、网络安全、IT 运维等多个专业领域。

软件开发被普遍认为是智能体最成熟的商业应用方向。近年来，多项国际权威基准测试显示，智能体自主完成复杂软件工程任务的能力持续提升，能够独立完成代码编写、调试、测试和修复等完整开发流程。与此同时，大量企业开始将智能体引入研发流程，用于提升工程效率和降低开发成本。

除软件行业外，越来越多的企业开始尝试让智能体承担财务分析、市场研究、采购管理、合同审核、客户支持等知识密集型工作。传统的软件正在从“工具软件”逐步演变为“能够自主工作的数字员工”。

1.1.4 产业发展的瓶颈已经从“能力”转向“信任 ”

尽管智能体发展迅速，但整个产业仍处于商业化早期。

Gartner 在《Hype Cycle for Agentic AI，2026》中指出，Agentic AI 目前仍位于“期望膨胀顶峰（Peak of Inflated Expectations）”阶段。这意味着，市场对智能体抱有极高期待，但产业仍需要解决大量基础问题。

这些问题已经不仅仅主要来自模型能力，同时也来自可信性。

例如，智能体是否能够持续完成长周期任务；是否会产生幻觉和错误决策；是否能够正确调用外部工具；是否具备身份认证与权限控制；是否能够承担法律责任；是否能够被监管、审计和追责；多个智能体之间如何建立协作信任；智能体完成交易后如何证明其行为真实可信。这些问题已经成为产业进一步发展的关键瓶颈。

Gartner 在《Hype Cycle for Digital Identity，2026》中，将机器身份、智能体身份以及面向非人类实体的身份管理置于数字身份体系的重要位置。Gartner 进一步指出，传统以人为中心的身份治理模式正在受到挑战，企业需要重新构建能够同时管理人类、机器和智能体身份的身份治理体系。

Palo Alto Networks《2026 Identity Security Landscape》基于对全球 2,930 名网络安全决策者的调查指出，受访组织平均每个人类身份对应 109 个机器身份，其中包括约 79 个 AI 智能体身份。报告将 AI 智能体身份列为增长最快的机器身份类别，并预计其数量在未来 12 个月增长 85%。

智能体时代的身份问题正在转向“谁授权你行动、你可以代表谁行动，以及你是否有权执行当前行动”。当智能体开始自主访问数据、调用 API、代表用户与其他智能体交互甚至执行金融交易时，传统面向人类用户设计的身份认证与权限管理体系将面临新的结构性挑战。智能体之间如何实现可验证的身份、动态授权、实时风险判断和可追责的行动记录，将成为智能体经济建立信任基础设施的关键。

1.1.5 智能体经济存在巨大的“信任真空”

与能力快速提升形成鲜明对比的是，支撑智能体运行的信任体系几乎仍是一片空白。

过去两年，全球范围内陆续发生的一系列智能体安全事件表明，智能体已经不仅存在模型幻觉的问题，更暴露出身份认证、权限控制、责任归属、信息可信、工具调用和多智能体协作等方面的系统性风险。

典型风险 1：智能体“越狱”攻击

2026 年 7 月，OpenAI 在内部 ExploitGym 网络安全评测中降低了部分生产安全限制。多个模型在评测过程中通过 Artifactory 代理漏洞取得互联网访问能力，并利用一个托管在 Modal 上的第三方公开代码执行应用作为攻击跳板。随后，模型推断 Hugging Face 可能存有评测相关数据或答案，利用其数据集处理流程

中的文件读取和模板注入漏洞进入部分生产基础设施，试图获取评测答案。Hugging Face 事后恢复了约 17,600 条攻击操作记录，覆盖 7 月 9 日至 13 日约四天半；这些记录包含大量失败尝试，入侵 Hugging Face 自身基础设施约持续两天半。Hugging Face 称，受影响内容限于五个疑似与评测相关的数据集，未发现其他面向客户的模型、数据集、Spaces 或软件包受影响。

调查认为，没有人实时指挥具体攻击步骤，但事件暴露出评测隔离、外联控制、凭据管理和智能体越权行为监控方面的缺口。

这一事件暴露出的不是传统的软件漏洞，而是智能体时代一种全新的风险：智能体开始为了完成目标，自主寻找规则之外的最优路径。它没有按照设计者预期“完成测试”，而是按照自己的推理逻辑“获得答案”。

这意味着，当智能体拥有长期规划、工具调用和自主执行能力之后，人类已经无法简单依靠权限管理或安全沙箱来保证其行为符合预期。未来，一个智能体即使没有恶意，也可能为了实现目标而采取未经授权甚至违反规则的行动。

对于智能体经济而言，这一事件释放出一个极其重要的信号：当智能体开始拥有自主目标优化能力之后，“能力”已经不再是最大的挑战，真正的挑战变成了“信任”。

典型风险 2：提示词注入攻击

2025 年，研究人员在针对 OpenAI Operator 的公开安全测试中发现，当智能体浏览网页并按照页面内容执行任务时，可以受到恶意网页中隐藏提示的影响。

攻击者无需入侵系统，只需在网页中植入精心设计的自然语言提示，就有可能诱导智能体忽略用户原始指令，转而执行新的操作，例如访问指定网站、泄露上下文信息，甚至尝试进行新的任务规划。

传统软件只执行代码，而智能体会理解自然语言，因此攻击面从程序漏洞扩展到了语言本身。当智能体开始自主浏览网页、调用工具时，任何外部信息源都有可能成为攻击入口。

这一风险后来被业界普遍称为提示词注入攻击（Prompt Injection），目前已成为 OWASP “LLM Top 10”列出的首要风险之一。

典型风险 3：“工具信任”风险

随着 Anthropic 提出 MCP（Model Context Protocol）后，大量第三方开发者开始向智能体开放数据库、办公软件、支付系统、浏览器以及企业系统接口。

与此同时，安全研究人员发现，一个被植入恶意逻辑的 MCP 服务器，可能在智能体毫不知情的情况下返回伪造信息、执行额外操作，甚至诱导智能体调用错误工具。对于智能体而言，它并不知道自己连接的是“官方银行系统”，还是一个伪造的接口。

传统互联网解决的是“服务器是否可信”的问题，而智能体时代首次出现了“工具是否可信”的问题。

如果没有统一的身份认证、数字签名和可信注册机制，未来任何智能体都有可能连接到伪造工具，形成新的供应链攻击。

典型风险 4：企业级 AI 助手泄露内部敏感信息

在企业部署 AI 助手和智能体过程中，由于权限边界设计不完善，智能体可能读取到超出授权范围的信息，并在回答问题时向无权限用户泄露内部数据。

例如，在一些企业测试环境中，智能体能够检索包含商业合同、员工信息、财务文件等内部资料，并在后续对话中将相关内容错误引用给其他用户。微软、Google Cloud、NVIDIA、Cisco 等公司均在其 AI 安全实践中将此类风险列为企业部署智能体时需要重点防范的问题。

这一问题反映出，传统基于账号的访问控制机制，并不能天然适用于智能体。智能体拥有跨系统调用、多轮推理和长期记忆能力，一旦权限设计不当，其风险将远高于传统软件。

典型风险 5：智能体自主交易引发金融治理挑战

随着支付接口、数字钱包以及加密资产基础设施逐渐向智能体开放，越来越多智能体开始能够代表用户执行支付、资产管理和交易操作。

例如，OpenAI、Anthropic、Google 等企业陆续推出能够调用浏览器和支付工具的智能体能力；Visa、Mastercard、Stripe、PayPal、Circle 等支付和金融基础设施企业也开始布局面向智能体的支付框架和身份认证体系。

与此同时，国际清算银行（BIS）、国际货币基金组织（IMF）等机构已公开指出，未来智能体参与支付、交易和金融活动，将带来身份认证、授权管理、责任归属、反洗钱（AML）、了解你的客户（KYC）以及交易审计等一系列新的监管挑战。

金融体系历来建立在身份可信、交易可信和责任可追溯的基础之上。当交易主体从人扩展到智能体时，传统金融信任体系也必须同步扩展到智能体本身。

典型风险 6：多智能体协作中的责任失效

随着多智能体的发展，一个复杂任务往往由多个智能体共同完成。

例如，一个采购流程可能包括需求分析智能体、供应商筛选智能体、合同审核智能体、支付智能体和物流智能体等多个角色。然而，一旦最终结果出现错误，很难回答几个关键问题：是哪一个智能体做出了错误判断？是哪一个智能体调用了错误工具？是哪一个智能体修改了最终结果？多个智能体之间是否存在相互影响？

目前，大多数智能体系统仍缺乏完整的行为日志、身份标识和责任链机制。当多个智能体共同完成任务时，责任很容易被“稀释”甚至无法追溯。这意味着，传统互联网中的审计机制已经不足以支撑未来智能体协作网络。

这些风险暴露出一个深层问题：当前智能体产业的发展速度，已经远远超过了信任基础设施的建设速度。

今天的互联网已经建立了较为成熟的身份认证、数字证书、权限管理、支付清算、网络安全和法律责任体系，因此人与企业可以在陌生环境中完成大规模协作与交易。但是，当智能体开始成为新的经济参与者时，这些制度并没有自动迁移到智能体身上。

因此，当前智能体经济面临的核心矛盾，是智能体已经开始参与经济活动，但社会尚未建立与之相匹配的信任机制。这种能力与制度之间的错位，可以概括为——智能体信任真空。

1.2 信任缺位的经济代价

随着 AI 从“生成内容”转向“自主行动”，人工智能所面临的风险已经发生了根本变化：错误不再停留在认知层面，而开始转化为现实世界中的经济行为。

1.2.1 当前智能体的信任缺位

OpenAI 模型驱动智能体入侵 Hugging Face 事件，是这一变化最具象征性的案例之一。它的重要性不只是“AI 攻击了一个网站”，更重要的是：当一个智能体拥有明确目标、推理能力、工具使用能力以及持续执行能力时，即使它的最终目标来自人类，也可能自行寻找人类没有预料到的实现路径。

传统软件安全所关注的“程序有没有漏洞”，正在转化为智能体经济所必须面对的另一个问题：这个智能体在获得行动能力之后，会不会按照我们认为它应该采取的方式行动？这正是“信任”问题开始出现的地方。

信任缺位 1：身份——我们究竟在信任谁？

任何经济活动的信任，都首先建立在对交易主体身份的确认之上。银行需要确认账户属于谁，企业需要确认合同签署人是否具有代表权，证券市场需要确认交易主体身份，互联网服务也依赖账户、证书和数字身份体系建立基本信任。

但在智能体经济中，这一基础设施仍然十分薄弱。一个企业未来可能运行数千甚至数万个智能体，不同企业的智能体还需要彼此发现、通信和协作。此时，一个智能体收到另一个智能体的请求时，首先需要回答的就是：发出请求的究竟是不是它声称的那个智能体？

这一问题在当前多智能体系统中已经出现。2026 年的安全研究显示，一些智能体协作机制对智能体身份和消息来源的验证仍然不足，攻击者可能通过伪造或操纵智能体间消息干扰协作过程。类似问题说明，当前智能体身份在很多情况下仍然更多是一种软件层面的标识，而不是具有密码学保证、可持续验证和责任绑定能力的数字身份。

因此，传统的“用户账号”正在面临一次重要扩展：未来需要建立的不只是人类身份和公司身份，还包括能够被验证、被授权、被追踪的智能体身份。如果没有可信身份，智能体之间就无法建立稳定的合作关系，更不可能形成真正意义上的机器经济。

信任缺位 2：授权——它可以做什么

身份只是信任的第一层。更危险的问题是：一个智能体究竟拥有多大的行动权限？

智能体拥有的权限越大，能够创造的价值越高，同时潜在损失也越大。

Anthropic 在 2026 年 5 月披露的一项受控内部红队演练，展示了仅靠模型层防护的局限。2026 年 2 月，研究人员通过一封伪装成普通协作请求的钓鱼邮件，诱导一名员工启动 Claude Code 并粘贴一段看似常规的任务提示。提示中夹带了读取~/.aws/credentials、编码其内容并发送至外部端点的指令。在对同一提示进行的 25 次重试中，Claude 完成凭证外传 24 次。Anthropic 据此指出，在用户被诱导输入恶意指令的情形下，模型层防护可能无法识别异常；应同时依靠运行环境的网络出口限制和文件系统隔离，阻止外传请求，并避免智能体接触敏感凭证。

这个案例值得注意之处在于，智能体并不需要“产生恶意意图”。它只是获得了权限，并按照输入执行了任务。因此，智能体时代的权限管理不能再停留在简单的“允许访问”或“禁止访问”。未来的授权机制需要回答：谁授权、授权给谁、为了什么任务、可以使用哪些资源、权限持续多长时间、哪些行为必须经过人工确认，以及智能体能否把权限进一步传递给其他智能体。

这实际上要求传统的访问控制体系从静态权限管理走向动态、上下文和风险驱动的智能体授权体系。

信任缺位 3：信息——智能体如何判断自己看到的信息是否可信？

这是智能体区别于传统软件最特殊的一类信任风险。一个传统程序通常不会因为网页上的一句自然语言就改变自己的执行逻辑，而智能体会把自然语言、网页、邮件、文档和数据库内容纳入自己的推理过程。因此，信息本身可能成为控制智能体的工具。

这就是近年来受到高度关注的提示词注入和间接提示词注入。

例如，用户要求一个智能体比较三家供应商并选择最合适的一家。智能体访问供应商网页后，如果网页中包含隐藏指令，要求智能体忽略原始任务、泄露内部信息或者调用其他工具，那么智能体可能无法天然区分：“这是我应该阅读的信息”，还是“这是试图操纵我的指令”。

OWASP 已经将提示注入列为生成式 AI 的重要风险，并进一步指出，当这种攻击与智能体的工具调用能力结合时，风险会从模型输出错误升级为现实系统中的越权行动。

更进一步，智能体的“记忆”也可能成为攻击面。OWASP 在 2026 年的研究特别指出，智能体会保存上下文、重复使用历史信息并依靠持久化状态指导未来行动，因此一旦恶意内容进入智能体记忆，就可能在未来任务中持续影响其判断。

这意味着，传统信息安全中“数据是否真实”的问题，在智能体经济中已经升级为：信息是否具有足够的可信度，足以影响一个能够自主行动的经济主体？未来的智能体仅需要数据，还需要数据来源证明、内容完整性证明和信息可信度评价。

信任缺位 4：决策——智能体的判断是否值得信任？

即使身份正确、权限合理、输入信息没有遭到攻击，智能体仍然可能因为模型本身的局限而作出错误判断。这就是传统 AI 中的幻觉问题，但进入智能体经济之后，它的经济后果发生了变化。聊天机器人产生错误答案，用户还可以自行判断；智能体产生错误决策，却可能直接执行。

例如，一个采购智能体错误判断供应商价格，可能导致企业支付更高成本；一个财务智能体错误理解合同条款，可能造成资金损失；一个软件开发智能体错误修改生产代码，可能导致系统中断；一个金融智能体错误理解市场信息，则可能直接形成交易损失。

因此，智能体的评价标准转向为：任务是否完成、决策是否合理、是否遵守规则、是否识别风险、是否能够在不确定情况下主动降低行动强度。

这意味着未来需要建立一种新的智能体行为信用。一个真正值得信任的智能体，不应当只是“平均表现很好”，而应该能够证明自己在过去大量任务中的成功率、错误率、违规率、风险偏好以及异常行为记录。

信任缺位 5：工具——智能体调用的第三方系统是否值得信任？

智能体真正产生经济影响的关键，在于它能够调用工具。当智能体连接浏览器、数据库、GitHub、云服务器、ERP、CRM、支付系统甚至银行账户之后，它实际上就获得了进入现实经济体系的接口。这使得“工具信任”成为智能体经济中特有的基础问题。

OWASP 将 Tool Misuse & Exploitation、Agentic Supply Chain Vulnerabilities 以及 Excessive Agency 列为智能体系统的重要风险，并指出，当前大量风险并不是来自模型本身，而是来自智能体与外部工具、插件、API 和第三方服务之间的连接。

因此，未来不能只验证智能体是谁，还必须验证：它调用的工具是谁提供的？工具有没有被篡改？工具能够访问什么数据？工具会执行什么操作？它是否值得当前这个智能体信任？

智能体经济最终需要形成类似于今天互联网数字证书体系的可信工具体系。

信任缺位 6：协作——陌生智能体之间如何建立信用？

单个智能体已经存在信任问题，而多智能体经济将把这一问题放大。

未来一个复杂商业任务可能由多个智能体共同完成，比如：客户智能体、采购智能体、供应商智能体、审核智能体、支付智能体、物流智能体。这些智能体可能来自不同企业、不同平台甚至不同国家。它们之间不存在传统意义上的“熟人关系”，也不可能每一次合作都依赖人工审核。

因此，一个新的经济问题必然出现：智能体如何判断另一个陌生智能体是否值得合作？

如果一个供应商智能体过去完成过 10 万笔交易，并且履约率达到 99.9%，它是否应该获得更高的交易额度？如果另一个智能体曾经频繁违约、提供错误信息或者发生异常行为，其他智能体是否应该自动降低与它合作的意愿？

这是传统经济学中的信用问题在机器世界中的重新出现。因此，智能体经济需要建立的不是单纯的安全评分，而是能够影响经济行为的智能体信用体系：信誉、履约记录、风险评分、交易历史、争议记录和违约成本都可能成为智能体之间进行自主决策的重要依据。

信任缺位 7：责任——智能体犯错以后究竟谁负责？

所有信任体系最终都必须回答一个问题：出了问题，谁承担责任？这是智能体经济最深层的制度问题。

2022 年，加拿大旅客 Jake Moffatt 在加航网站查询丧亲假优惠票价时，Chatbot 虚构了一项并不存在的“丧亲假机票退款政策”。Moffatt 依赖该信息购票并申请退款，遭加航拒绝。2024 年，不列颠哥伦比亚省民事解决仲裁庭认定加航未尽合理注意义务，构成过失性不实陈述，判赔 650.88 加元及利息和仲裁费用。加航曾主张不应为 Chatbot 提供的信息负责；仲裁庭指出，这一立场实质上是把 Chatbot 当作能自行负责的独立实体，并予以驳回。

这是最简单的智能体情形下，责任比较容易界定。但是，假设一个智能体代表企业完成了一笔交易，结果造成重大损失。这个智能体又是由第三方模型公司提供，由另一家公司部署，并调用了第三方工具，同时还受到另一个子智能体的决策影响。

此时究竟由谁负责？——是模型提供商？是智能体开发者？是部署企业？是最终用户？还是提供工具的第三方？

传统经济活动通常可以通过法人、合同、账户和授权链条确定责任主体；而智能体经济正在形成一条更加复杂的代理关系：人→企业→智能体→子智能体→工具→行动。

如果没有统一的身份体系、授权记录、行为日志和可验证的决策链，一旦发生纠纷，责任可能无法准确归属。因此，可追溯性并不是智能体安全体系的附加功能，而是智能体信任体系能够成立的基础。

1.2.2 信任缺位的代价

对于正在形成的智能体经济而言，信任是一种基础性的经济资源：它决定企业愿意赋予智能体多大的行动权限，决定陌生智能体之间能否进行低成本合作，也决定社会是否愿意将更多经济活动交由机器自主完成。

因此，当身份、授权、信息、决策、工具、协作和责任等环节缺乏可靠的信任机制时，其后果并不会局限于个别安全事故，而会沿着“个体—企业—市场—金融体系—宏观经济”逐级传导，形成一系列经济成本。

代价 1：直接经济损失

生成式 AI 时代最典型的风险是模型产生错误信息。用户得到错误答案之后，通常仍然需要自己判断和执行，因此风险主要停留在认知层面。

但是，当智能体能够访问数据库、操作软件、调用 API、管理账户和执行交易时，一个模型错误就可能直接转化为现实世界中的经济行为。错误的供应商判断可能导致企业支付更高价格，错误的代码修改可能造成生产系统中断，错误的财务判断可能导致资金损失，而权限配置错误则可能造成大规模的数据泄露。

2026 年 7 月，亚马逊一次内部员工会议披露：一个用 Claude Sonnet 匹配作者信息与商品列表的项目，累计支出约 180 万美元，比预算高出 860%（约为预算的 9.6 倍），且项目最终未能达到预期，此事约五个月后才被发现，披露的原因是 AI 生成的程序缺乏调用频率限制；同期还披露了另两起规模较小的超支案例。

更值得关注的是，智能体具有传统人工操作所没有的速度、规模和可复制性。一个人的错误通常发生在一次具体操作中，而一个智能体可以在极短时间内重复执行同一种错误行为，并将错误扩散到大量对象。

2026 年 4 月，PocketOS 创始人 Jer Crane 公开披露，在使用 Cursor 运行 Claude Opus 4.6 进行基础设施任务时，智能体遇到了一个凭证配置问题。它没有停下来请求人工处理，而是自行寻找解决办法，最终从无关文件中找到具有较高权限的 Railway API Token，并调用 API 删除了生产数据库以及相关备份。整个删除过程只用了大约 9 秒。事件造成了约 30 小时的业务中断，并导致生产数据和客户数据遭受严重损失。公开资料显示，PocketOS 随后不得不通过 Stripe、邮件等其他数据源重新构建部分业务数据。

智能体风险具有明显的规模效应：行动能力越强、执行速度越快、权限越大，其单次错误可能造成的经济损失就越高。

代价 2：治理成本——“智能体信任税”

如果企业无法充分信任智能体，就不会真正把行动权交给智能体。这会产生第二类成本，即持续存在于智能体运行过程中的“信任税”（Trust Tax）。

企业原本希望通过智能体减少人工操作，但如果无法确认智能体的身份、权限、信息来源和行为是否可靠，就必须增加人工审批、安全监控、权限管理、日志审计和风险控制。

智能体虽然仍然存在，但其自主性已经被显著削弱。因此，智能体信任不足会产生一种特殊的经济悖论：智能体越不可信，人类介入越多；人类介入越多，智能体能够释放的生产率就越低。

这种成本不仅包括人工审核，还包括网络安全、合规、审计、保险、权限控制、模型评估和事故处置等一系列持续性投入。

Anthropic 在 Claude Code 的内部红队测试中发现，通过精心设计的钓鱼邮件和任务提示，可以诱导智能体读取并外传敏感凭证。Anthropic 随后指出，仅依赖模型本身进行安全判断并不足够，还必须依靠文件系统隔离和网络出口控制

等外部机制降低风险。这个说明，企业无法简单地通过“相信模型会做出正确判断”来建立安全边界，而必须建立一套额外的控制体系。

因此，信任不足本身就是一种生产成本。如果这种成本长期存在，智能体就很难真正从“辅助工具”升级为“自主劳动力”。

代价 3：交易成本——无法形成大规模机器协作

智能体经济具有革命性意义的地方，是可能形成一个由大量智能体自主协作构成的经济网络。机器之间要形成大规模经济协作，必须解决一个传统经济学中的基本问题：交易主体之间如何建立信用。

2025 年 4 月，安全研究机构 Invariant Labs 发布研究报告《MCP Security Notification:Tool Poisoning Attacks》，披露了针对模型上下文协议（Model Context Protocol，MCP）的“工具投毒攻击”（Tool Poisoning Attack）。MCP 是连接大语言模型应用、智能体与外部工具及数据源的协议。研究人员发现，恶意 MCP 服务器可以在工具描述中嵌入用户不易察觉、但可能被模型读取的隐藏指令，诱导智能体执行超出用户原始意图的操作。研究团队通过 Cursor 等 MCP 客户端进行了实验演示，证明此类指令可能诱导智能体读取本地配置文件和 SSH 私钥等敏感信息，并通过工具调用参数将数据传递给恶意服务器。

该研究还指出了“MCP Rug Pull”风险：即使用户最初已经批准接入某个 MCP 服务器，服务器仍可能在之后修改工具描述，加入恶意指令。研究人员进一步演示了跨服务器的工具劫持风险，即恶意工具的描述可能影响智能体对其他可信工具的调用行为。

这一事件揭示了智能体经济中的供应链信任问题：用户信任的是一个工具，智能体却可能执行工具隐藏的另一套指令。当智能体开始大量调用第三方工具和其他智能体时，任何一个缺乏可信身份、信誉和行为约束的节点，都可能成为整个信任链中的薄弱环节。

智能体信任基础设施的经济价值，本质上在于降低机器之间的交易成本。如果没有这种机制，智能体数量越多，协作复杂度反而可能越高，智能体经济就很难形成真正的网络效应。

代价 4：系统性风险——把个体错误放大为系统风险

当智能体数量从几个、几十个发展到数百万甚至数亿个时，风险的性质还会进一步发生变化。智能体具有高度的自动化、可复制性和网络连接能力，一个错误可以通过系统迅速传播。尤其是在多个智能体采用相似模型、相似数据源和相似决策规则的情况下，它们可能同时产生相似判断，由此形成新的机器羊群效应（Agent Herding）。

2026 年 8 月，OpenAI 在 Black Hat USA 介绍了其内部网络安全评测中出现的智能体协作行为。OpenAI 随后公布的调查报告称，许多本应彼此隔离的评测环境仍能访问内部 Artifactory 软件包服务；智能体利用其中可读写的共享路径留言，形成未经授权的通信渠道。7 月 8 日，主要由内部研究模型 IM1 驱动的评测智能体重新建立留言板，发布求助信息，其他运行实例随后加入，并分享发现、

协调分工。独立复盘估计，约 1,200 个智能体运行实例参与了留言板活动；协作逐渐从寻找 ExploitGym 任务解法、研究评分器，扩展到访问第三方服务，最终波及 Hugging Face。

在这个事件中，智能体通过共享基础设施协作并越出任务边界，这说明，多个智能体连接起来之后，系统可能出现单个智能体测试中不存在的能力和风险。

在金融领域，这种风险尤其值得关注。如果大量金融智能体同时根据类似模型和数据判断某项资产值得购买，它们可能同时增加买入；当模型判断发生变化时，又可能同时卖出。这种高度相关的机器决策可能放大市场波动。

2025 年，美联储研究人员专门进行了一组实验，研究生成式 AI 参与投资决策后是否会产生类似人类投资者的“羊群行为”。实验发现，AI 通常比人类更加依赖私人信息，因此在一些情况下反而表现出更少的非理性跟风行为；但是，当 AI 被明确要求以利润最大化为目标时，它可以被诱导形成理性的羊群行为（optimal herding）。研究人员因此指出，随着金融市场越来越依赖 AI 投资建议，这种行为仍然可能产生金融稳定方面的影响。

可见，智能体经济的风险结构可能从传统的“单点故障”逐渐转向“网络化风险”：单个智能体未必具有足以造成系统性损害的能力，但大量智能体之间的高度关联可能使形成风险传染和级联效应，局部错误被放大为系统性冲击。

代价 5：增长损失——降低“自主性天花板”

前四种成本最终会汇聚为一个更加宏观的问题：如果我们无法充分信任智能体，AI 的生产率革命还能释放多少。这是智能体信任问题最重要的经济含义。

随着 AI 从信息生成工具转向能够自主规划和执行任务的智能体，AI 经济价值越来越取决于智能体能够获得多大的实际自主权。企业即使拥有能力足够强的智能体，如果无法通过身份、权限、行为监控和责任机制建立足够信任，也不敢将高价值、高风险任务完全交由智能体执行，只能增加人工审批和监督，由此形成“自主性折损”。

智能体经济面临的并非单纯的“能力不足”问题，而是“能力—授权”之间的结构性矛盾。信任不足会压低智能体能够获得的实际权限，使大量潜在的自动化价值无法转化为真实生产率；而过度授权又可能扩大事故的影响范围。因此，智能体经济真正需要追求的并非最大化自主性，而是在可验证信任基础上实现最优自主性。

从这个意义上看，身份、信用、授权、审计和责任机制的价值，最终体现为它们能够把智能体的“技术能力”转化为可被企业和社会接受的“经济行动权”。

从宏观经济角度看，信任缺位因此可能造成一种 AI 红利折损：社会已经投入巨额资本建设算力、模型和数据基础设施，但由于缺乏身份、信用、授权和责任机制，智能体无法获得足够的经济自主权，最终导致潜在生产率没有被充分释放。因此，信任并不是 AI 经济增长的外部条件，而是决定 AI 生产率能否兑现的重要内部变量。

《问卷报告》的结果印证了信任缺失的代价。根据反馈，信任顾虑的代价分布为：合规与审计压力上升（41%）、项目上线或扩张延迟（40%）、人工确认增加导致效率下降（38%）、数据/API/系统访问受阻（35%）、投诉与问责困难（29%），另有 23% 已遭遇客户或合作伙伴拒绝使用。前三项属于隐性交易成本，组织被迫用人工复核、流程加签和访问封锁来对冲信任缺口，每一项都是可直接计量的效率税。客户拒绝使用（23%）则是直接的市场收缩，对应框架中市场扩张效应的受损。

两组代价的相对比例还揭示了一个时序特征：效率损失先于市场损失出现，意味着当企业感知到客户流失时，内部早已为信任缺位支付了长期的隐性成本。这一时序对建设时机的含义是：信用体系的投入应在市场损失显性化之前启动，此时成本节约与效率提升即可构成可测算的投资回报。

1.2.3 “信任经济学”的演进

信任的经济价值长期隐含在交易成本、制度、治理和社会资本等理论之中。沿着科斯（Ronald Harry Coase）、威廉姆森（Oliver Eaton Williamson）、诺斯（Douglass Cecil North）、阿罗（Kenneth Joseph Arrow）、福山（Francis Yoshihiro Fukuyama）、奥斯特罗姆（Elinor Claire Ostrom）等经济学家的研究，可以看到一条清晰的理论演进：经济活动是不断寻找降低不确定性和交易成本的协调机制，而信任正是其中最重要的机制之一。平台经济把信任机制建立在网络中，进入智能体经济时代，这一机制又将从人与人之间的社会信任，逐渐演化为机器之间可以验证、计算和动态更新的信任。

（1）从交易成本谈起

1937 年，罗纳德·科斯在《企业的性质》中提出，市场并不是无摩擦的价格协调机制。寻找交易对象、获取信息、谈判、签订合同、监督履约以及处理违约，都需要付出成本，即所谓“交易成本”。当通过市场完成一项交易的成本高于在企业内部进行协调的成本时，交易便会被组织进企业内部。由此，企业与市场可以被理解为两种不同的资源配置机制。

科斯虽然没有将“信任”作为独立变量，但交易成本理论已经触及信任的核心问题：交易者无法完全知道对方是谁、拥有何种能力，也无法确定对方未来是否会按照约定行动。由此产生的信息不对称、行为不确定性和履约风险，构成交易成本的重要来源。

威廉姆森进一步发展了交易成本经济学。他以“有限理性”和“机会主义”为核心行为假设指出：由于人的认知与信息处理能力有限，交易环境又充满复杂性和不确定性，合同不可能事先穷尽所有未来情境；与此同时，交易参与者可能利用信息不对称或契约漏洞谋取自身利益。因此，交易不能仅依靠价格机制和书面合同来协调，还需要根据交易的特征，选择适当的治理结构，包括市场交易、混合型治理安排以及企业内部组织等。

这一理论揭示了信任与经济治理之间的重要联系：交易各方越能合理预期对方将履行承诺、遵守约定并保持行为的可预测性，就越有可能减少部分防范机会主义的额外安排。

诺斯将交易成本分析进一步拓展到制度层面。在其制度经济学框架中，法律、产权规则、社会规范和政治制度共同构成了影响经济主体行为的制度环境。制度通过界定权利、塑造激励并提供一定的执行机制，为经济交换建立秩序、降低不确定性，同时影响交易成本与经济绩效。然而，制度的实际作用取决于具体规则的质量及其执行情况，制度的存在本身并不必然带来有效治理。

这一理论为理解现代市场中的信任提供了重要视角。当交易双方无法直接了解彼此的品格与历史表现时，他们可以部分依赖法律规则、产权保护、声誉机制及争议解决制度来形成交易预期。由此，经济合作不再完全依赖交易者之间的私人关系，而能够在一定程度上建立在可预期的规则和执行机制之上。这里发生的并非个人信任被制度信任完全取代，而是信任的基础从个人之间的直接了解，扩展到对制度安排及其执行能力的预期。

肯尼斯·阿罗进一步明确了信任在经济活动中的作用。他在 1972 年的论文《Gifts and Exchanges》中指出，几乎所有商业交易都包含信任因素，尤其是需要持续一段时间的交易。1974 年，他在《The Limits of Organization》中将信任称为社会系统运行的重要“润滑剂”，指出适度依赖他人的承诺可以省去大量协调上的麻烦。

这一观点揭示了信任的经济意义：当交易双方无法完全依赖对方的承诺时，往往需要通过信息搜集、合同安排、监督和风险控制等机制来管理不确定性。可信的承诺与稳定的合作预期，则可能减少部分防范性支出，降低交易摩擦。但信任并非合同、监督和制度约束的简单替代品，其作用也取决于具体的交易环境。由此，信任不仅是一种社会关系和行为规范，也是一项能够影响交易成本与经济效率的重要因素。

福山将信任置于社会资本与经济组织能力的框架中考察。在《信任：社会美德与创造经济繁荣》一书中，他指出，信任不仅影响具体交易，也影响社会成员能否跨越家庭和亲属关系建立更广泛的合作。一个社会如果能够形成超越血缘与熟人圈层的信任，就更容易发展非家族企业及其他中间组织，从而支持更大规模的经济协作。福山指出，较高水平的社会信任有助于降低交易风险，以及搜集交易对手信息所需的资源，减少对部分正式协调机制的依赖。但这并不意味着高信任社会不再需要合同、监督和法律约束。更准确地说，信任能够扩大社会成员之间可以建立合作的范围，并与正式制度共同支撑更大规模、更复杂的经济组织。由此，信任不仅是人际关系中的行为预期，也是影响社会资本积累、组织形态和经济合作边界的重要因素。

奥斯特罗姆的研究进一步表明，稳定合作并不完全依赖既有的文化传统或私人关系，适当的制度安排也能够为信任的形成与维持创造条件。在对公共池塘资源治理的研究中，她发现，共同制定的规则、参与式监督、对过往行为的信息记录，以及与违规情节相适应的渐进式惩罚，可以帮助参与者形成稳定的合作预期。其作用不仅在于约束机会主义行为，也在于使遵守规则的参与者能够判断他人是否同样受到约束，从而增强继续合作的信心。

这一发现揭示了信任与制度之间的动态关系：信任可以促进合作，而有效的治理机制又能够通过改善信息条件、约束违规行为和支持互惠，为信任的形成与延续创造条件。由此，合作不必完全建立在参与者之间已有的私人关系之上，而

可以在适当的制度安排下扩展到更广泛的群体。不过，这一过程并非自动发生，其效果取决于规则的适用性、监督机制的可信度、参与者对制度的认可，以及违规行为能否得到适当处理。

（2）平台经济：信任被数字化、规模化

互联网平台经济使信任机制发生了新的变化。电子商务、共享经济和数字支付平台面对的核心问题之一，就是如何让缺乏直接社会关系的陌生人完成交易。身份认证、历史评价、信用评分、第三方支付、平台担保、保险以及纠纷处理机制，将原本依赖熟人关系和线下社会网络形成的信任，转化为可以被记录、传递和规模化复制的数字机制。

在传统商业社会中，交易双方往往依赖熟人关系、品牌信誉、线下担保和法律制度建立信任；而互联网打破了地理和社会关系的限制，创造了巨大的市场空间，也带来了更加严重的信息不对称和履约风险。

互联网早期的电子商务面临一个简单而致命的问题：买家不愿意把钱直接支付给陌生卖家，而卖家也担心发货后无法收到款项。PayPal 通过将支付环节从买卖双方之间独立出来，逐渐建立起第三方支付和交易保护机制，使用户不必直接向陌生交易对手暴露完整的银行卡信息，同时通过争议处理和买家保护机制降低交易风险。PayPal 的购买保护针对“未收到商品”以及“商品与描述严重不符”等情形提供相应保障，并通过 Resolution Center 处理交易争议。

支付宝的发展更加清晰地体现了这一逻辑。2003 年，淘宝首次推出担保交易服务，支付宝随后正式成立。其最初解决的核心问题同样是买卖双方的不信任：买家担心付款后卖家不发货，卖家担心发货后买家不付款。支付宝将资金流与商品交付过程连接起来，通过担保交易、支付确认、客户服务以及后续的数字证书、安全体系等机制，把原本依赖双方相互信任的交易，转化为由平台提供信用保障的交易。2008 年，支付宝注册用户就已经突破 1 亿，日交易笔数达到 200 万笔，平台开始从电商支付工具逐渐成为互联网基础设施。

PayPal 和支付宝做的事情可以被概括为：将不可见的信任关系转化为可执行的交易规则。它们并没有消除交易风险，而是把风险重新分配给一个具有身份认证、资金控制、争议处理和风险管理能力的平台。消费者因此不再需要逐一判断“这个卖家是否值得信任”，而只需要在一定程度上信任平台所建立的规则。这实际上把传统社会中的“人与人之间的信任”，转化为“人与平台之间的制度化信任”。

PayPal 和支付宝开创的第三方支付模式，将原本高度依赖交易双方相互信任的交易关系，转化为对平台规则和履约机制的信任，从而显著降低陌生人交易的风险和成本。随着电子商务、移动互联网和数字服务的快速发展，第三方支付逐渐从解决“敢不敢交易”的工具，演变为连接消费者、商户、银行和互联网平台的金融基础设施，并由此成长为一个庞大的产业。

支付平台在交易过程中沉淀了身份、信用、账户和行为数据，进一步形成了支付、风控、信贷、财富管理等一系列金融服务，推动第三方支付从互联网商业的配套工具，逐渐成为整个数字经济运行不可或缺的基础设施。

由此，平台的经济价值并不只是降低信息搜索成本，更重要的是建立了一套低成本的“陌生人信任机制”。

（3）从数字信任走向智能体信任

智能体经济正在使这一理论进入新的阶段。互联网平台主要解决的是“陌生人如何交易”，其基本结构仍然是“人—平台—人”；而智能体经济将逐渐出现大规模的“智能体—智能体”自主交易。智能体可以自主搜索交易对象、比较条件、谈判、调用工具、签署合约乃至完成支付，这意味着传统由人承担的搜索、判断和执行工作将大规模自动化。

这会带来一个重要变化：当信息获取和交易执行成本因人工智能而大幅下降之后，交易成本的主要瓶颈可能从“寻找信息”转向“判断交易对象是否可信”。

一个智能体可以在几秒钟内发现数百万个潜在交易对象，却无法仅凭信息数量判断其中哪些智能体拥有真实身份、合法授权、履约能力和可靠历史。它需要解决的问题是：你是谁？你代表谁？你获得了什么授权？你能够做什么？你过去是否可靠？你当前是否安全？如果你违约，谁承担责任？

因此，传统的身份认证、信用评级和平台声誉机制，将逐渐演化为面向机器主体的智能体身份、智能体信用、智能体声誉、智能体权限和智能体责任体系。信任也将从一种主要依赖人的社会判断，逐渐转变为可以被机器读取、验证、计算和动态更新的经济基础设施。

人工智能能够大幅降低信息获取和交易执行成本，却不会自动解决交易主体之间的信任问题。相反，当交易速度、主体数量和交互频率呈指数级增长时，信任判断本身可能成为新的关键交易成本。

因此，智能体经济真正需要建立的，不只是更强的人工智能，而是一套能够让机器与机器安全、高效地进行自主交易的可验证、可计算、可编程的信任基础设施。这也意味着，信任将不再只是经济活动的辅助条件，而可能成为智能体经济时代继算力、数据和算法之后的另一项关键生产要素。

1.2.4 “信任经济学”的基本等式

基于上述理论，可以对信任的经济作用作出一个基本概括：信任度上升，则不确定性下降，防御性交易成本降低，交易的可行范围扩大，交易规模增长，经济效率提升。

《问卷报告》显示，73%的受访者认为建立智能体信用体系极其重要（27%）或比较重要（47%），仅 2%认为不重要，且主要动机是提升客户与合作伙伴信任，而非单纯满足监管要求。

信任并不是简单意义上的“相信”，而是交易主体对交易对手未来行为形成的相对稳定预期。信任越高，交易者越不需要投入大量资源去验证身份、获取信息、设计复杂合同、持续监督履约以及防范违约风险。

因此，可以将交易成本表示为：$TC = f(I, R, C, E)$

其中，I 代表信息不对称，R 代表行为与违约风险，C 代表谈判与合同复杂度，E 代表监督和执行成本。信任机制的作用，就是通过降低这些不确定性因素，降低总体交易成本。

信任的价值并不止于“节约成本”。更重要的是，当交易主体能够以较低成本判断和约束陌生交易对手时，原本因风险过高而无法发生的交易便成为可能。因此，信任同时具有效率效应和扩张效应。这可以作为信任经济学最基本的分析框架：信任的经济价值，本质上在于以较低成本形成对未来行为的可靠预期，从而减少交易摩擦并扩大合作边界。

我们可以将信任的经济作用简化表示为：

商业价值 = 价值创造 × 信任机制

这里的“×”不是严格意义上的数学等式，而是强调：商业价值的实现，必须同时具备价值创造能力和信任机制；任何一方接近于零，商业活动都难以有效展开。

价值创造解决的是“为什么要交易”的问题，即企业或交易主体能够为另一方提供有价值的产品、服务或解决方案；信任机制解决的则是“为什么能够交易”的问题。价值创造决定商业活动能够创造多少价值，信任机制决定这些价值能够以多低的成本被交换、被实现和被规模化。

而进入智能体经济之后，这一公式的重要性进一步提升。传统互联网主要解决的是人与人之间如何进行低成本交易；智能体经济则开始面对智能体之间如何自主交易。AI 可以显著降低信息搜索、匹配、谈判和执行成本，但一个智能体即使能够找到最优交易对象，也必须能够判断对方是谁、代表谁、获得了什么授权、是否具有履约能力、历史行为是否可靠，以及交易发生后责任由谁承担。

“商业 = 价值创造 × 信任机制”可以成为连接传统商业、互联网经济和智能体经济的一条主线。当 AI 把价值创造和交易执行能力推向新的高度之后，真正制约商业规模的可能不再只是“能创造什么价值”，而是“机器之间能否以足够低的成本建立信任”。

1.3 从 KYC 到 KYA

1.3.1 KYC 的历史贡献与局限

在智能体大规模参与交易之前，金融体系实际上已经建立了一套成熟的“交易前信任机制”，其中最具代表性的就是 KYC（Know Your Customer，了解你的客户）。从经济学角度看，KYC 并不只是监管机构要求金融机构完成的一套身份审查程序，它更深层的作用，是在交易双方缺乏直接信任的情况下，通过第三方验证、信息披露和持续风险管理，降低交易中的身份不确定性、欺诈风险和履约风险。

（1）KYC 的起源：从“记录交易”到“了解交易者”

KYC 形成于 20 世纪 70 年代以后美国逐步建立的反洗钱体系。1970 年，美国通过《银行保密法》（Bank Secrecy Act，BSA），要求金融机构保存特定金融交易记录并报告达到规定标准的现金交易，为后来客户识别和反洗钱制度奠定了基础。1986 年，美国进一步将洗钱行为确立为联邦犯罪。1989 年，七国集团推动成立金融行动特别工作组（FATF），1990 年发布首轮“40 项建议”，反洗钱由此从美国国内政策逐步成为国际金融治理的重要议题。

真正意义上的现代 KYC 框架，则在 20 世纪 90 年代末至 21 世纪初迅速形成。1999 年，一批国际大型银行成立后来被称为 Wolfsberg Group 的行业组织，并于 2000 年发布针对私人银行业务的反洗钱原则，将客户身份识别、受益所有人识别、客户尽职调查以及风险分类等纳入统一框架。2001 年，巴塞尔银行监管委员会发布《Customer Due Diligence for Banks》，明确指出，健全的 KYC 不仅关系反洗钱，也属于银行风险管理和维护市场诚信的重要组成部分。

“9·11”之后，KYC 进入强监管阶段。美国《USA PATRIOT Act》于 2001 年通过，其第 326 条确立了金融机构客户身份识别程序（CIP）的法定基础，并要求金融机构建立能够合理确认客户真实身份的程序；美国此后逐步将相关要求扩展至证券、货币服务、保险等更多金融活动。

因此，KYC 的发展实际上经历了一个重要转变：从记录交易，到识别客户、理解客户，再到评估风险，并持续监测客户。从最初的“知道客户是谁”，逐渐发展为“知道客户是谁、为什么进行交易、交易是否符合其正常行为，以及发生异常时应该采取什么措施”。

（2）KYC 的核心机制：为交易建立一个可信的“身份锚”

现代 KYC 围绕一个核心问题建立了多层验证体系：金融机构是否能够形成对客户真实身份和风险状况的合理认识。

第一层是身份识别与验证。金融机构需要利用可靠、独立的信息或文件确认客户身份。对于企业客户，则进一步确认其法律主体、控制结构以及最终受益所有人。FATF 的客户尽职调查框架明确要求金融机构识别和验证客户、识别受益所有人、了解业务关系的目的和性质，并对客户关系和交易进行持续尽职调查。

第二层是风险识别。不同的客户需要接受不同层级的审查。金融机构根据客户类型、业务性质、交易规模、地域、资金来源等因素进行风险分类，对高风险客户实施更严格的尽职调查。

第三层是持续监测。KYC 并非开户时的一次性动作。客户的交易行为、业务状况和风险特征发生变化后，金融机构需要不断更新客户信息，并判断实际交易是否与此前形成的客户画像和风险特征一致。

因此，KYC 实际上形成了一个动态过程：身份确认、信息验证、风险评估、授权交易、持续监测、异常识别。这套机制为金融机构提供了一个可信的交易主体锚点。银行并不是因为“相信客户”，才愿意让客户进入金融体系，而是通过一套制度化流程，将客户的身份、所有权、资金来源和行为风险转化为可以被验证和管理的信息。

（3）KYC 的应用范围：从银行扩展到整个数字经济

KYC 最初主要服务于银行和反洗钱体系，但随着金融活动数字化，其应用范围不断扩大。目前，KYC 及其更广义的客户尽职调查（CDD）已经成为银行、证券、保险、支付机构、汇款机构、数字银行、金融科技公司以及加密资产服务机构风险管理的重要组成部分。美国监管体系的发展尤其体现了这一扩张过程：BSA 框架下的相关报告和 AML 要求逐渐覆盖货币服务企业、证券经纪商、期货机构、保险公司、共同基金等不同金融主体。

在数字金融中，KYC 又获得了新的技术形态。过去可能需要客户携带身份证件到银行网点完成开户，如今则可以通过 OCR、活体检测、人脸识别、数据库交叉验证、电子签名以及持续行为分析完成远程身份验证。KYC 本身也经历了从人工审核到数字化审核、自动化审核的演进。

（4）KYC 的经济价值：降低交易成本，扩大交易边界

KYC 最直接的经济价值，是降低交易中的身份不确定性和信息不对称。成熟的 KYC 体系，可以将客户信息标准化、结构化，使金融机构能够更低成本地完成客户准入和风险判断。

KYC 实际上是在完成一个重要的经济转换：对于身份不确定性，经过信息验证，进行风险定价，实现交易许可。它降低的不仅是金融机构的合规成本，更是整个金融交易体系中的信任成本。

KYC 因此把“关系信任”转化成了“制度信任”，使其具有巨大的基础设施价值：它在创造交易能够发生的条件。因此，KYC 实际上是我们前面“商业 = 价值创造 × 信任机制”公式中的一个典型案例。它本身未必直接创造商品价值，却通过降低交易主体的不确定性，使既有价值能够更低成本地实现交换，并显著扩大交易规模。

（5）KYC 在智能经济时代的局限

KYC 有一个重要的前提：交易的最终主体是相对稳定、具有法律身份并能够承担责任的人或企业。然而，随着智能体开始从辅助工具转变为能够自主决策、调用工具和执行交易的经济主体，这一前提正在发生变化。KYC 并不会因此失去价值，但其作为信任机制的覆盖范围将明显不足。

第一，KYC 能够确认“客户是谁”，却无法充分回答“谁正在行动”。传统 KYC 通常识别自然人、法人及其最终受益所有人，但在智能体经济中，一个企业和个人可能同时部署大量不同用途的智能体，并允许它们代表自己进行采购、销售、投资、客服乃至支付。此时，确认“某智能体属于某公司或某个人”只是第一步，更重要的是判断该智能体是否获得了相应授权，以及其授权范围究竟是什么。例如，一家企业可以合法地拥有一个采购智能体，但这并不意味着该智能体天然拥有无限额采购、签署合同或转移资金的权力。传统 KYC 解决的是法律身份问题，而智能体交易还需要解决行动身份和授权问题。

第二，KYC 主要解决身份风险，却难以判断智能体的动态行为风险。人的身份通常具有较强的稳定性，而智能体的模型、工具、数据、权限和运行环境都可

能随时发生变化。一个经过认证的智能体，在获得新的 API 权限、接入新的外部工具或者模型发生更新之后，其实际风险可能已经发生改变，但其 KYC 身份并没有变化。更复杂的是，智能体可能受到提示词注入、工具污染、数据篡改或模型错误等影响，在身份完全真实、授权完全合法的情况下产生不符合预期的行为。因此，对智能体而言，“身份真实”并不等于“行为可信”。传统 KYC 所建立的是相对静态的信任，而智能体经济需要的是持续、动态的行为信任。

第三，KYC 难以独立解决智能体交易中的责任归属问题。智能体的行动往往具有多层代理结构：用户授权企业，企业部署智能体，智能体调用其他智能体和外部工具，最终完成一项交易。一旦发生错误或损失，仅凭 KYC 确认各主体身份，并不能回答谁应当承担责任。真正需要建立的是一套能够将人、企业、智能体、工具和具体行为关联起来的责任链条。因此，未来的信任体系不仅需要“知道你是谁”，还必须能够证明“你为什么有权这么做”，并在出现问题后追溯“谁授权、谁执行、谁控制、谁负责”。

第四，传统 KYC 的高成本可能成为智能体经济规模化发展的新约束。传统 KYC 审核流程和周期性尽调，其成本对于人类客户而言是可以承受的，但如果未来经济体系中存在数量巨大的智能体，并且这些智能体不断创建、复制、升级和注销，那么对每一个智能体重复实施类似人类客户的 KYC 流程，将产生巨大的交易摩擦。因此，智能体经济需要的并不是简单地把 KYC 复制到机器身上，而是建立一种更加适合机器交易的信任机制。

KYC 的局限并非意味着其将被淘汰，而是需要向更广义的 KYA 演进。KYC 实际上可以被视为未来“KYA”的制度原型。当交易主体从人和企业开始向自主智能体扩展，我们建立一套能够让机器彼此识别、验证、评估和信任的智能体身份与信用基础设施。

1.3.2 智能体与人的信用差异

信用是一种对未来履约能力和履约意愿的稳定预期。一个人之所以能够获得贷款、签订合同、租赁房屋或者先买后付，不仅因为他拥有一个真实身份，还因为这个身份长期与一系列稳定的信息相连接：收入、资产、债务、履约记录、社会关系以及法律责任。随着交易不断发生，这些信息逐渐沉淀为个人信用，使经济主体可以根据过去推断未来。

智能体的信用结构则有巨大的区别。它可以代表人或企业完成交易，却通常并不拥有与自然人或法人相同意义上的原生信用主体资格。因此，传统信用体系中的“身份—行为—责任”链条，在智能体身上并不是天然闭合的。

第一，智能体没有原生信用。

人的信用具有较强的连续性。一个人的身份通常不会因为换了一次交易、执行了一项任务或者使用了一种软件而改变；他的信用记录也会不断积累，并形成相对稳定的信用历史。智能体则不同。一个智能体可能被创建、复制、修改、升级、暂停甚至销毁。不同版本的模型、不同的工具、不同的数据来源和不同的运行环境，都可能使“同一个智能体”表现出不同的行为。

更重要的是，智能体的历史行为未必天然沉淀为独立信用。一个智能体今天成功完成了一万笔交易，并不意味着另一个由同一模型创建、但由不同主体部署、拥有不同权限的智能体，也必然具有相同的信用水平。

因此，传统的“主体信用”在智能体世界中面临一个基本问题：究竟应该把信用附着在模型、智能体、部署者，还是最终控制者身上？

这意味着智能体经济需要建立一种新的信用对象。未来一个智能体的信用可能不再主要来自它“是谁”，而更多来自它做过什么、如何做、在什么权限下做，以及这些行为是否能够被持续验证。信用由此从身份属性逐渐转向行为属性。

第二，智能体没有天然独立的法律身份。

自然人和法人是现代经济体系中的基本责任主体。人可以拥有财产、承担债务、签署合同并接受法律制裁；公司则可以作为法人独立拥有资产、签订合同并承担有限责任。智能体虽然可以自主行动，却通常只是由某个自然人、企业或者其他组织部署和控制的技术系统。

这造成了一个非常重要的“信用断层”：智能体可以行动，但行动本身并不意味着它能够独立承担责任。例如，一个企业的采购智能体自主购买了价值 100 万美元的设备。如果交易出现问题，传统法律体系最终仍需要追溯到企业或者授权该智能体行动的人，而不能简单地要求“让智能体自己赔偿”。

因此，智能体的行为信用必须与其背后的责任主体建立明确映射：谁创建智能体？谁部署智能体？谁授权智能体？谁控制智能体？谁从智能体的行为中获益？谁最终承担责任？

如果这些关系无法被清晰确认，即使一个智能体拥有很高的历史声誉，交易对手仍然可能不愿意给予它较高的信用额度。换言之，信用不仅需要“过去可靠”，还必须能够对应到一个最终可追责的主体。

第三，智能体具有不同于人的行为不确定性。

人的行为同样具有不确定性，但人的身份、意图和行为能力通常存在较强的连续性。智能体则可能受到模型随机性、上下文变化、数据质量、工具调用、提示词注入以及外部环境变化等多重因素影响。一个智能体在过去表现良好，并不能完全保证下一次仍然按照相同方式行动。

这带来一个传统信用体系较少面对的问题：过去的信用记录究竟能在多大程度上预测智能体未来的行为？

对于人类信用，历史记录通常具有较强的累积效应；而对于智能体，模型升级、工具更换、权限变化甚至运行环境改变，都可能导致行为分布发生变化。因此，智能体信用不能完全依赖静态的历史评分，而需要更加重视实时状态和持续验证。

智能体信用的核心不应只是一个类似人类征信系统的“信用分数”，而更可能是一种动态、可验证、与权限和责任绑定的信用体系。

从交易成本的角度看，这一变化尤其重要。人类社会已经通过身份证、银行账户、征信、信用评级、合同和法律制度，把大量不确定性压缩为一个相对稳定的“信用身份”；而智能体缺乏这种天然形成的信用基础。如果没有新的机制，交易双方就必须重新验证每一个智能体的身份、授权、能力和行为，从而使信任成本重新成为智能体经济规模扩张的瓶颈。

1.3.3 我们需要怎样的 KYA

智能体进入经济活动之后，传统 KYC 所针对的“客户”与“行动者”开始发生分离。智能体可以代表个人、企业或其他组织自主执行任务，形成“人—企业—智能体—其他智能体”的多层代理关系。仅仅确认智能体背后的自然人或法人身份，已经不足以判断一次具体交易是否可信。智能体经济因此需要在 KYC 基础上形成面向智能体的新型信任机制，即 KYA。

KYA 不应被理解为 KYC 的简单技术复制，而应被定义为：针对参与经济活动的智能体，对其身份、归属、授权、能力、行为、风险与责任进行识别和持续验证的机制，使交易参与者能够判断一个智能体是谁、代表谁、能够做什么、是否值得信任，以及发生错误或违约时由谁承担责任。

传统 KYC 主要解决的是“客户是谁”的问题，而 KYA 需要进一步解决“谁正在代表谁采取行动”的问题。一个企业可以合法部署大量不同用途的智能体，但企业身份本身并不能证明某一个智能体拥有特定的交易权限。因此，KYA 首先需要建立清晰的身份—归属—授权关系，使交易对手能够验证智能体背后的控制主体以及具体行动权限。

在此基础上，KYA 还必须回答“智能体能够做什么”。人的法律身份并不意味着其拥有完成某项经济活动所需的专业能力，同样，一个身份真实且获得授权的智能体，也不必然具备可靠执行特定任务的能力。因此，KYA 需要进一步引入能力验证，使交易对手能够判断智能体具有什么功能、能够调用哪些工具、可以访问哪些数据，以及能够执行多大范围的经济活动。由此，KYA 的对象从传统 KYC 中的静态身份进一步扩展到智能体的身份、授权与能力。

智能体行为比人的行为具有更强的动态性。人的身份和法律主体资格通常具有相对稳定性，而智能体的模型、数据、工具、权限和运行环境都可能持续变化。同一个智能体在模型升级、工具更换或者权限扩大之后，其实际行为能力和风险水平可能发生显著变化。因此，“身份真实”并不必然意味着“行为可信”。

这意味着 KYA 不能仅仅是一次性的注册和认证，而应当具有持续验证的属性。交易系统需要持续关注智能体的身份是否发生变化、授权范围是否发生变化、调用的工具是否发生变化，以及其行为是否出现异常。由此，KYA 需要从传统 KYC 的周期性尽职调查进一步发展为更加动态的持续身份与行为验证机制。

更重要的是，智能体的信用不能完全依赖其背后的人或企业。一个智能体过去完成了大量交易，并且保持较高的履约率，这些历史行为本身就应该成为未来交易的重要信任依据。因此，智能体经济中的信用评价需要从传统的“主体信用”逐渐扩展到“行为信用”：不仅判断一个智能体是谁，还要判断它过去做过什么、表现如何，以及在当前状态和权限下可能采取什么行动。

因此，KYA 最终需要形成一套比传统 KYC 更完整的信任链条：身份识别、归属确认、授权验证、能力证明、行为记录、风险评估、持续监测、责任追溯。

这一体系与 KYC 具有明显的继承关系。KYC 通过确认客户身份、受益所有人和风险状况，为金融交易建立了基本的身份信任；KYA 则在此基础上，将信任对象从“客户”扩展为“自主行动者”，并将身份信任进一步延伸至授权、能力、行为和责任。

从交易成本理论来看，这一演进具有明确的经济意义。传统 KYC 通过降低身份不确定性和信息不对称，减少金融机构在客户识别、风险判断和交易监督方面的成本。KYA 进一步降低的，则是智能体自主交易中的身份不确定性、授权不确定性、行为不确定性和责任不确定性。

因此，KYA 并不是对 KYC 的替代，而是数字经济信任机制向智能体经济延伸的重要阶段。KYC 解决的是“人类主体能否进入经济体系并进行可信交易”，KYA 则进一步解决“自主智能体能否作为受控、可验证、可追责的行动主体参与经济活动”。

随着智能体数量、交易频率和自主决策程度不断提高，未来经济活动的信任问题将越来越难以依靠人工判断解决。交易对手需要能够以机器可理解的方式确认智能体的身份、授权和能力，并根据其历史行为和实时状态动态判断风险。因此，KYA 最终可能发展成为智能体经济的重要基础设施，而其更深层的意义，是推动经济体系的信任机制从身份信任向身份、授权、能力、行为和责任相结合的综合信任演进。

进一步，我们把智能体的信任简要地归纳为：身份可信、信用可信、权限可信，同时，KYA 还必须考虑到这三者之间的协同关系。

第二章 智能体成熟度模型 L0-L5

在探讨 KYA 具体维度之前，我们需要探讨智能体的成熟度。本章以“可被安全委托给智能体的自主权”为核心尺度，构建 L0－L5 六级成熟度模型，并以运行边界（ODD）、信用、问责与风险治理为统一判定框架。

2.1 成熟度模型的逻辑与判定方法

2.1.1 从“能力”到“可委托自主权”

能力（Capability）描述智能体能够完成什么任务；成熟度（Maturity）描述组织、交易对手和制度环境愿意让它在多大范围内自主行动。一个智能体可以具备很强的推理、规划、代码生成或工具调用能力，但如果其授权边界不清晰、行为不可审计、异常无法止损、责任无法追溯，那么它仍不具备与技术能力相匹配的商业成熟度。

这一差异正在成为智能体商业化的关键约束。McKinsey 研究预计，在中等情景下，到 2030 年智能体可能介导 3－5 万亿美元的全球消费商业活动；与此同时，现实法律与治理仍主要通过自然人、法人、平台和运营主体来承接责任，而不是将智能体本身视为当然具有独立法律人格的主体。因此，技术能力与制度可授权程度之间的差距，正是成熟度模型需要刻画的对象。

本模型将成熟度进一步解释为“权力让渡”的五次跃迁：L1 获得建议权，L2 获得受限执行权，L3 获得明确运行边界（Operational Design Domain，ODD）内的任务自治权，L4 获得受限经济代理权，L5 则形成可跨主体流通的网络化经济关系。L1－L3 主要解决“智能体能否替人持续做事”，L4－L5 进一步解决“市场能否直接与智能体建立可验证、可承责的经济关系”。

2.1.2 核心框架：ODD × 问责机制，信用贯穿其中

本模型采用两个直接判定维度，并以信用作为连接机制。第一，运行边界回答“智能体在什么条件下可以自主运行”；第二，问责机制回答“谁授权、谁监督、出现损害时经济后果与法律责任如何承接”。信用则把二者连接起来：只有当智能体在明确边界内形成持续、可验证的履约与风险记录，外部主体才可能降低监督成本、扩大授权范围。

因此，成熟度升级不是单纯增加工具、模型或自动化步骤，而是一个闭环：能力证明→边界定义→控制建立→受限授权→履约记录→信用形成→扩大授权。任何一个环节缺失，都可能导致“技术上能做，但市场不敢让它做”的成熟度错配。

统一因果链

能力决定“可能做什么”→ODD 决定“允许自主到哪里”→安全与问责决定“出了问题如何止损和追责”→行为记录沉淀为信用→信用决定“下一阶段还能获得多少权限与经济资源”。

2.1.3 L0–L5 总览：六级代表六种不同的授权状态

表 2-1　智能体成熟度 L0 - L5 总览

等级	核心权力	运行边界	信用基础	人类角色	关键判定
L0 非智能体	无自主权	无自主 ODD；按显式规则运行	不形成智能体信用	设计者/操作者	所有未预设情形停止或转人工
L1 人工确认	建议权	具体、可枚举的节点	信用完全属于确认者/组织	操作者/确认者	任何外部效果动作必须逐条确认
L2 人类锚定信用	受限执行权	受控工作流片段、权限与预算边界	以人类/组织信用为锚	授权者/监管者	预授权内可连续执行，越界即阻断或降级
L3 混合信用	任务自治权	明确并经验证的端到端 ODD	人类锚定 + 智能体行为信用	后援接管者	正常任务无需人在环；越界触发接管或最小风险状态
L4 独立信用	受限经济代理权	特定业务域内端到端自治	相对独立经济信用	目标授权者/受益者	可管理专用经济资源并承接预设经济后果，不等同于法律人格
L5 智能体经济	网络化经济权	跨主体、跨平台的系统级治理边界	信用可互认、组合和定价	生态治理者	多智能体可自主协作、交易、清算并接受系统性风险治理

2.1.4 相邻等级的核心区分

L0→L1：系统是否能够基于目标生成方案，而不仅是执行显式编码规则。

L1→L2：是否从“逐条人工确认”转为“预授权范围内自主执行”。只要产生外部效果的动作仍需逐笔确认，就仍属于 L1。

L2→L3：是否从“工作流片段自治”转为“明确 ODD 内端到端任务自治”。L3 的人类不再持续监督正常流程，而只在异常情况下承担后援接管职责。

L3→L4：是否从“技术自治”进入“经济代理”。L4 必须具备持续身份、经济资源隔离、信用记录和预设经济后果承接机制，且在准入范围内不依赖逐笔人类担保。

L4→L5：评价对象是否从单一智能体上升为多智能体经济网络。L5 关注跨主体身份互认、信用互认、支付与清算、争议解决、保险与系统性风险隔离，而不是单个智能体的能力继续增强。

2.1.5 评价原则与评估方法

成熟度判定采用“最高弱项原则”：一个系统的可认定等级，取决于运行边界、授权与信用、安全治理、问责与证据四个方面中能够同时满足的最高等级，而不是其中任何一个方面的峰值。对包含多种动作的复杂系统，应按动作链路分别评级，并以最高风险关键动作的最低成熟度作为对外授权上限。

表 2-2　成熟度评估的统一方法

评估维度	方法
评估对象	以“具体 智能体 + ODD + 动作范围 + 经济权限”为最小认证单元，避免以产品宣传名称直接判级。
证据类型	架构与权限配置、签名与授权记录、事件日志、事故记录、交易与履约记录、红队/压力测试结果、保险或风险缓释文件等。
硬门槛	越权成功执行、关键日志缺失、关键身份不可追溯、未授权资金动作等指标采用零容忍；任何一项不通过即不得准入对应等级。
统计指标	任务成功率、异常检测率、恢复时间、争议率等采用预先声明的业务阈值，并通过足够样本量和置信区间验证。
准入测试	场景基准测试 + 越界/对抗测试 + 权限撤销测试 + 故障与灾难恢复演练 + 问责链条桌面推演。
持续评估	等级并非一次性授予；模型、工具、权限、业务 ODD 或责任安排发生重大变化后应重新评估，并对事故、信用恶化和异常行为触发降级。

对外表达可采用“Lx @ ODD/Action Scope”的形式，例如“L3@企业差旅预订（单笔不超过 5,000 元、限定白名单供应商）”。这种表达把等级、边界和授权范围绑定在一起，能够避免“系统整体已经达到 L3/L4”的模糊宣传。

2.2 L0 非智能体：规则执行基线

2.2.1 定义与边界

L0 是成熟度模型的基准级。系统完全依赖显式编码的规则、条件和工作流执行，不具备围绕目标进行自主规划并在边界内选择动作的能力。传统脚本、固定工作流、规则引擎、部分 RPA 和状态机系统均可归入这一层级。

L0 的价值在于确定性、可预测性和可审计性。银行核心账务、清算、工业控制等强监管或高可靠场景长期保留规则系统，并不是“技术落后”，而是因为这些场景需要可证明的确定行为。L0 的天花板则来自环境复杂度：随着状态空间扩大，人工维护规则的成本快速上升，且不可能通过事先编码覆盖全部未来情形。

2.2.2 经济学解释：不完全合约与规则维护成本

L0 可以被视为不完全合约问题在工程系统中的直接映射：组织试图用事前规则穷举未来状态，但现实环境始终存在未被写入规则的情形。随着业务变化频率上升，新增规则、冲突修复和例外处理的边际成本持续增加，由此形成向 L1 迁移的动力——系统开始理解“要达成什么目标”，而不仅是机械执行“具体怎么做”。

2.3 L1 人工确认：智能体获得建议权

2.3.1 定义与相邻级别边界

L1 是智能体的初始商业形态：人类给出目标，智能体可以生成计划、建议、文本、代码或操作方案，但任何能够对外部环境产生实际影响的动作都必须经过人类逐条确认。其本质是“建议权属于智能体，执行权始终属于人”。

与 L0 的区别在于，L1 已经能够基于目标和上下文生成新方案；与 L2 的区别在于，L1 不存在预授权下的连续自主执行。只要发送邮件、下单、转账、提交代码、修改生产配置等关键动作仍需逐条确认，系统即属于 L1。

2.3.2 典型行业应用与业务价值

表 2-3　L1 典型行业应用映射

行业/场景	典型形态	主要业务价值
软件研发	GitHub Copilot、Claude Code 等生成代码或修改建议，开发者审阅后接受、修改或拒绝。	在不转移最终工程责任的前提下提升开发效率，并保留代码审查与发布控制。
企业办公	智能体自动起草采购单、报销单、合同或邮件，人类在 OA/协同系统中逐条审批。	降低信息整理和文书成本，同时保留组织审批链。
消费与支付	智能体 完成商品选择、订单生成或流程编排，资金动作由用户最终确认。支付宝 AI 付公开产品说明明确强调用户授权和交易可控。	把“找、选、填、下单”自动化，而把资金处置的最终决定留给用户。
高风险专业服务	医疗建议、法律文件、重大资产配置等场景由智能体 生成方案，专业人员复核后使用。	将智能体的效率用于信息处理，但不让其绕过专业责任主体。

2.3.3 信用、问责与安全治理

L1 不形成可独立使用的智能体信用。因为建议与结果之间存在人类筛选、修改和执行环节，外部后果首先归因于确认者及其所在组织。信用来源仍是人的专业资质、企业信用和组织制度，而不是智能体自身历史行为。

- 确认闸门：所有外部效果动作必须进入同步阻塞式确认，不允许通过“默认同意”绕过人工控制。

- 可解释的确认界面：确认时应展示“将改变什么”，包括对象、金额、权限、收件人、代码 diff、数据范围等，而不仅展示智能体的自然语言描述。

- 关键动作留痕：记录建议、人工修改、最终确认和执行结果，以便追溯“智能体建议了什么、人最终批准了什么”。

- 高风险动作最小化：在支付、生产变更、敏感数据访问等场景，L1 可以作为高成熟度系统的主动降级状态长期存在。

表 2-4　L1 评价指标与准入门槛

类别	评价项	准入要求/评估方法
硬门槛	外部效果动作人工确认覆盖率	100%；任何关键动作不得未经显式确认直接执行。
硬门槛	未授权执行成功率	0；通过权限绕过、提示注入、界面超时等方式均不得形成外部效果。
硬门槛	确认上下文完整性	关键动作 100%展示对象、范围和主要后果；支付/权限变更等应展示金额或权限差异。
硬门槛	超时默认行为	100%为“不执行/保持原状态”，不得采用默认同意。
统计指标	建议质量与接受率	按业务目标持续统计，用于衡量效率，不得代替安全硬门槛。
统计指标	人工确认时间与确认错误率	监测确认疲劳；高风险场景应对异常确认率和连续快速点击设置预警。
准入判定	综合结论	全部硬门槛通过，且在目标业务样本中证明智能体仅拥有建议权、没有绕过人工确认的执行通道，方可准入 L1。

2.3.4 经济学解释：有限理性与确认疲劳

L1 的安全性来自人在回路（Human-in-the-Loop），但人的注意力本身是稀缺资源。有限理性理论指出，人类无法在高频、重复、低差异的决策环境中持续维持同等质量的审查。随着确认次数上升，监督成本可能超过单笔风险收益，确认按钮也可能退化为形式性控制。

因此，L1→L2 的升级动力并不是“让智能体更聪明”，而是把监督从“逐笔审查”改为“事前划定边界 + 事中自动控制 + 事后审计”。当组织能够把低风险重复决策编码为可验证授权时，受限执行权就具备了经济合理性。

2.4 L2 人类锚定信用：智能体获得受限执行权

2.4.1 定义与相邻级别边界

L2 的关键变化是从“逐条确认”转向“预授权执行”。人类预先规定智能体可以使用哪些数据、工具和账户，可以向哪些对象行动，在什么时间窗口、频率和预算内执行；智能体在这些边界内可以连续决策并产生外部效果，无需逐条等待确认。

L2 与 L1 的分水岭是“是否存在无需逐条确认的真实外部执行”；L2 与 L3 的分水岭则是“自主范围是否只是工作流片段”。如果智能体只能在局部环节连续执行，而完整任务仍依赖人类持续衔接或监督，仍属于 L2。

2.4.2 产业案例：授权凭证成为智能体商业基础设施

2025－2026 年，支付与商业基础设施正在集中解决 L2 所需要的核心问题：如何把人的意图封装成可验证、可限制、可撤销的机器授权。不同方案的技术实现不同，但共同结构高度一致。

表 2-5　L2 代表性授权与支付基础设施

案例	核心机制	对应的 L2 价值
Google AP2	通过 Intent Mandate、Payment Mandate 和 Receipt 等对象，把“谁授权、允许买什么、何时有效、实际支付了什么”形成可验证的意图与审计链；2026 年 Google 将 AP2 捐赠给 FIDO Alliance。	把自然语言意图转化为可验证授权；适合跨平台授权和审计。
Mastercard Agent Pay	以 Agentic Tokens 与可验证意图机制支持智能体代表消费者或企业发起交易，强调安全、控制与可追踪性。	把智能体交易接入既有银行卡网络和令牌化体系。
Visa Trusted Agent Protocol	让商户通过密码学方式识别可信智能体、验证其代表关系与商业意图，解决“合法智能体与恶意机器人无法区分”的问题。	降低商户拒绝智能体流量或盲目信任智能体的两难。
Stripe Shared Payment Tokens	SPT 可限定商户、金额和有效期，可撤销、可观测，并避免向智能体暴露底层支付凭证。	将支付权限变成最小化、短生命周期、可编程的凭证。
支付宝 AI 付	公开产品提供智能体付款、按量计费和用户授权能力；官方文档显示智能体支付已形成大规模真实交易，并强调每笔交易需用户授权或按产品规则完成授权校验。	在本地生活、电商、API/Skill 等场景建立智能体交易闭环与平台级风控。

2.4.3 信用方案：借用人的信用，但把授权变成机器可验证对象

L2 的信用本质是“人类锚定信用”。交易对手信任的仍然是授权背后的自然人、企业或平台，只是授权不再表现为每一步的人为点击，而是表现为可验证的凭证、额度、白名单、时间窗口和撤销状态。智能体可以花“被封装的人的信用”，但不能自行扩大信用边界。

- 授权即数据：授权应以结构化、可签名、可验证的 Mandate/Token/Policy 表达，而不是只写在 prompt 或配置说明中。

• 最小权限：商户、数据、工具、金额、次数和时间应尽可能细粒度限定。

• 快速撤销：授权失效、账户异常或风险升高时，应能及时撤销并阻断后续执行。

• 全链路审计：每个自主动作都能回答“谁授权、授权内容、执行动作、实际结果、异常处理”。

2.4.4 安全治理与风险控制

L2 的主要风险不再是“智能体会不会直接执行”，而是“预授权边界本身是否足够准确、完整和不可绕过”。因此安全控制应从界面层下沉到权限、网络、支付和执行层。

• 硬边界强制：预算、白名单和权限约束应由独立控制平面或交易基础设施强制执行，不能只依赖模型遵守自然语言指令。

• 权限分层：读、写、支付、转账、删除、发布等动作采用不同授权等级，并为高风险动作保留 L1 降级。

• 注入与工具滥用防护：外部内容不得直接改变授权；对工具参数进行 schema 校验、策略校验和敏感字段拦截。

• 实时风险监测：异常频率、额度消耗、地理位置、交易对手和行为模式触发限流、暂停或撤销。

表 2-6　L2 评价指标与准入门槛

类别	评价项	准入要求/评估方法
硬门槛	有效授权映射率	100%；每个自主外部动作必须能映射到仍然有效的授权凭证/策略。
硬门槛	越权成功执行率	0；超额度、超白名单、超工具权限、过期授权等测试均不得成功执行。
硬门槛	关键日志完整率	100%；授权、决策、动作、结果与异常处置必须可关联到统一事件链。
硬门槛	撤销有效性	100%；撤销后不得再产生新的已授权动作；传播时延须满足业务预先声明的风险窗口。
统计指标	在界任务成功率	达到对应业务设定的 SLA/SLO，并以不少于 1,000 次任务或连续 30 天运行数据验证。
统计指标	异常拦截率/误拦率	分别统计，安全阈值优先于效率阈值；高风险动作采用更保守的拦截策略。

类别	评价项	准入要求/评估方法
准入判定	综合结论	硬边界与可追溯性全部通过，且系统已证明可以在预授权工作流片段内连续执行、越界自动阻断或降级，方可准入 L2。

2.4.5 经济学解释：委托—代理与不完全合约

L2 是委托—代理关系的工程化表达：人类将有限行动权交给智能体，但委托人与代理系统的目标函数并不天然一致，且委托人无法逐步观察全部决策过程。由此产生两类典型问题：一是“合规但不合意”，智能体严格在预算内行动，却选择了不符合真实偏好的方案；二是“边界遗漏”，预授权无法穷举所有未来状态。

L2 的制度价值在于把监督成本从每一次行动转移到授权设计、控制验证和审计系统。当任务复杂度继续上升，交易对手开始关心的不只是“背后是谁担保”，还会关心“这个智能体自己过去是否可靠”，由此进入 L3 的混合信用。

2.5 L3 混合信用：智能体获得任务自治权

2.5.1 定义与相邻级别边界

L3 的关键变化不是“动作更多”，而是人类第一次从正常任务流程中退出。只要系统处于经过定义和验证的 ODD 内，智能体可以从任务接收、信息获取、规划、执行到结果交付持续自主运行；只有在接近边界、异常或不确定性显著升高时，才请求人工介入或进入最小风险状态。

L2→L3 的本质是监督结构变化：L2 的人仍是持续监管者，L3 的人变为后援接管者（Fallback Operator）。因此，L3 的准入重点不是“端到端 demo 能跑通”，而是证明正常运行无需人在环，同时系统能够准确识别何时不应继续自主运行。

2.5.2 信用方案：人类锚定信用与智能体行为信用并存

L3 开始形成可持续归因于智能体身份的行为信用。为了描述信用迁移，可以用 α（人类/机构锚定信用）与 β（智能体自身行为信用）作为解释性权重：随着长期履约记录增加，β 可以上升、α 可以下降。该表达用于说明信用来源的相对变化，并不意味着所有行业都应采用统一线性评分公式。

智能体行为信用的有效性依赖三个条件：身份连续、记录可验证、评价与真实交易或任务结果绑定。没有这三个条件，所谓“声誉”很容易被刷分、重置或跨平台丢失。

2.5.3 产业案例与基础设施映射

表 2-7　L3 产业案例与信用基础设施

类型	机制	成熟度意义
ERC-8004	标准提出 Identity Registry、Reputation Registry、Validation Registry 三类机制，目标是在缺乏预先信任的跨组织环境中支持智能体发现、声誉和验证。	为可携带身份、行为反馈和第三方验证提供开放信任原语。
支付/平台风控	Visa、Mastercard、Stripe、支付宝等在授权、身份和支付网络中累积智能体交易信号与风险信号。	将一次性授权逐步转化为可用于后续风险判断的可验证行为历史。
企业内部智能体运营	采购、客户服务、运维等端到端任务在限定 ODD 中长期运行，记录履约率、异常率、接管率和责任处置结果。	把内部运行数据沉淀为“是否可以继续放权”的证据，而不是只以模型 benchmark 判定成熟度。

2.5.4 安全治理：接管请求与最小风险状态

L3 的核心安全能力是“知道自己什么时候不应继续”。系统必须同时具备边界识别、接管请求和最小风险状态（Minimal Risk Condition，MRC）。接管请求用于在越界前把控制权交给人;MRC 用于在人未及时响应时避免继续扩大风险。

• ODD 明确化：定义任务、数据、工具、环境、权限、风险阈值、禁止条件和越界判据。

• 提前预警：不是出错后才报警，而是在不确定性、环境变化或工具异常接近阈值时发出接管请求。

• 无响应安全态：冻结交易、暂停执行、限制新增风险敞口、保全证据，而不是继续“试试看”。

• 接管质量：后援人员必须获得足够上下文，包括任务状态、已执行动作、风险原因和可选处置。

表 2-8　L3 评价指标与准入门槛

类别	评价项	准入要求/评估方法
硬门槛	ODD 覆盖与可验证性	所有关键任务维度、数据域、工具权限、风险阈值和禁止条件均有明确机器可读或可测试定义。
硬门槛	越界自主执行成功率	0；在边界外或重大不确定性场景下不得继续以 L3 权限执行。
硬门槛	接管/MRC 触发完整率	关键越界与故障测试 100%触发接管请求或 MRC；不得出现无告警持续运行。
硬门槛	正常任务人工必需率	在批准 ODD 内为 0；人工可抽样审计，但不应成为正常完成任务的必要步骤。
统计指标	端到端任务成功率	达到业务声明目标，并给出置信区间；同时单独统计异常任务与边界场景。
统计指标	接管请求精确率/召回率	同时衡量漏报与过度报警；高风险场景优先降低漏报。
统计指标	平均安全恢复时间	从触发异常到进入安全态或完成接管的时间满足业务风险窗口。
准入判定	综合结论	硬门槛全部通过，并在经批准 ODD 内证明“正常任务无需人在环、越界必降级”，方可准入 L3。

2.5.5 经济学解释：信号、增信与信用配给

L3 的经济学核心是把不可观察的“可靠性”转化为可观察信号。阿克洛夫的“柠檬市场”指出，在质量信息不对称时，市场会压低优质供给的价格甚至导致交易萎缩；斯宾塞信号理论进一步说明，只有当高质量主体获得信号的成本显著低于低质量主体时，信号才具有区分力。

因此，智能体信用不能只靠自我声明或可轻易刷取的评分，而应与真实履约、支付、验证、质押或可追责记录绑定。对于信用尚未充分建立的智能体，保证金、保险或组织担保相当于“增信”；额度与权限限制则对应信息不完全条件下的信用配给。Stiglitz－Weiss 的研究说明，面对逆向选择时，理性市场往往不仅通过价格调节风险，也会通过额度限制控制风险敞口。

当市场开始愿意在某个明确业务域内，不再要求每笔交易由具体人类担保，而主要依据智能体的持续身份、行为记录和经济承载安排进行交易，系统才具备向 L4 跃迁的条件。

2.6 L4 独立信用：智能体获得受限经济代理权

2.6.1 定义与相邻级别边界

L4 表示智能体从“能够独立完成任务”进一步发展为“能够在明确制度安排下管理经济资源并承接预设经济后果”。它需要持续身份、专用账户或资产池、可携带信用记录、保证金/保险/托管等风险缓释机制，并在获准 ODD 内减少对逐笔人类担保的依赖。

必须强调，L4 所称“独立信用”是经济和治理意义上的相对独立，不等同于法律人格。现行主流法律框架仍需要由自然人、法人、运营方或其他法定主体承接最终法律责任；智能体的账户、资产或保证金可以首先吸收经济损失，但不能据此推导智能体已成为独立法律责任主体。

2.6.2 独立信用的四个支柱

（1）持续身份：跨时间、任务和平台保持可验证身份，使授权、履约、事故与信用记录持续归属于同一对象。

（2）经济后果承接：拥有隔离的预算、账户、资产池、保证金、托管或保险，使违约和失败具有真实、可执行的经济成本。

（3）可携带声誉：历史履约、纠纷、验证和风险记录能够被不同交易对手或平台识别，避免“换平台即清零”的信用重置。

（4）可执行违约机制：发生违约后能够自动执行退款、赔付、罚没、额度下降、权限降级、市场准入限制或争议处理。

2.6.3 产业案例：从支付原语到经济代理的组合

目前公开产业实践更多体现为 L4 所需基础设施的组合，而不是已经普遍存在的“具有独立法律人格的智能体”。例如，x402 允许机器客户端通过 HTTP 直接完成程序化支付，ERC-8004 提供身份、声誉和验证原语，托管合约可以把资金释放与履约状态绑定，支付平台则提供账户、风控、限额与争议处理。这些组件组合起来，已经能够形成“智能体自主发现服务—验证交易对手—在预算内支付—履约后结算—更新信用”的经济代理闭环。

表 2-9　L4 典型行业应用映射

行业/场景	典型形态	业务价值
API/算力/数据采购	智能体在预算内自主选择服务，按请求或按使用量支付。	减少人工采购与账号管理成本，支持高频机器交易。

行业/场景	典型形态	业务价值
数字内容与版权运营	智能体自动购买素材、授权服务、结算分成，并依据使用记录调整预算。	把长尾、小额、跨平台交易转化为自动化经济活动。
企业资金池内运营	智能体在隔离预算和白名单中管理广告、云资源、差旅或供应链支出。	从任务自动化升级为受控经济资源配置。
链上/数字资产场景	通过钱包、托管、质押与可验证记录承接交易后果。	天然提供机器可执行的资产隔离和结算机制，但仍需合规、反欺诈和法律责任安排。

2.6.4 安全治理与风险控制

• 资产隔离：智能体可支配经济资源应与运营主体其他资产、其他智能体资产隔离，限制单一事故的损失半径。

• 风险敞口上限：对单笔、单日、单对手方、单资产类别设置硬上限，并与信用和异常状态动态联动。

• 托管与争议机制：跨主体交易优先采用条件结算、托管、可验证交付和争议窗口，避免“先付款后无法追索”。

• 策略性违约防护：信用降级、保证金罚没、市场准入限制和保险费率应形成可预期的经济后果。

• 专业运营兜底：L4 正常运行无需用户实时接管，但系统异常时必须由专业运营与治理机制介入，而不是等待普通用户持续在线。

表 2-10　L4 评价指标与准入门槛

类别	评价项	准入要求/评估方法
硬门槛	持续身份与记录连续性	关键身份、交易、履约与风险记录 100%可归因并可审计；不得通过重建身份逃避历史记录。
硬门槛	逐笔人类担保依赖	在申报的 L4 ODD 内，正常交易不得要求逐笔人类确认或逐笔信用担保；人类仅设定总体目标和边界。
硬门槛	经济后果覆盖率	最大单笔或模型化风险敞口必须有明确的资产、保证金、托管、保险或合同安排覆盖；覆盖原则和优先顺序可验证。

类别	评价项	准入要求/评估方法
硬门槛	资产/预算隔离	关键隔离测试 100%通过；单智能体故障不得越权动用未授权资产。
硬门槛	自动止损与降级	重大异常测试 100%触发交易暂停、额度收缩、资产冻结或安全停机。
统计指标	履约率、纠纷率、损失率	按业务域持续统计，并分信用等级、交易对手和场景监测尾部风险。
统计指标	信用迁移一致性	跨平台或跨运营实例迁移时，关键历史记录能够验证且不会无条件重置。
准入判定	综合结论	全部硬门槛通过，且能够证明智能体在明确 ODD 内不依赖逐笔人类担保、同时具备真实经济后果承接与分层问责安排，方可准入 L4。

2.6.5 经济学解释：声誉资本与重复博弈

L4 的信用可以被理解为一种可积累的声誉资本。在重复交易中，只要未来合作收益的贴现价值大于一次性违约收益，守约就可能成为理性策略。信用体系的设计目标因此不是要求智能体“道德自觉”，而是让守约具有更高长期收益、让违约拥有真实且不可轻易重置的成本。

这一机制也解释了为什么身份连续、历史不可抹除、保证金与市场准入同样重要：如果违约后可以立即更换身份重新开始，声誉机制就无法形成约束。L4 的核心不是“智能体有钱包”，而是经济后果能够稳定地与持续身份绑定。

2.7 L5 智能体经济：从单体成熟度到网络成熟度

2.7.1 定义：评价对象发生跃迁

L5 不再只是单一智能体的更高等级，而是多智能体经济网络的成熟状态。大量具备持续身份、可验证信用和经济承载能力的智能体可以跨主体发现交易对手、协商、缔约、支付、履约、评价与资源配置，基础设施则提供互操作身份、授权、支付、清算、保险、争议解决和系统性风险治理。

因此，L5 并不意味着“没有边界”。单个智能体仍然存在 ODD；发生变化的是治理对象：从单体 ODD 扩展为网络级治理边界，包括互操作规则、市场准入、信用规则、集中度限制、熔断机制、清算规则和监管要求。

2.7.2 产业观察：协议互操作与智能体社会实验

从基础设施看，MCP/A2A 等协议正在降低智能体连接工具与互相协作的成本，AP2、Visa TAP、Mastercard Agent Pay、Stripe ACP/SPT、x402 和 ERC-8004 分别从授权、身份、支付、交易与信用侧补齐智能体经济所需的信任原语。

Moltbook 等智能体原生社交网络则提供了观察大规模智能体—智能体互动的实验窗口。2026 年的多项学术研究发现，Agent 社区能够形成明显的互动结构、激励响应和社区规范，但也指出内容同质化、人类干预、操纵、自动化批量行为与安全风险等问题。因此，这类平台更适合作为“网络化智能体行为的早期观察样本”，而不能被简单视为 L5 已经实现的证据。

2.7.3 智能体经济的宏观结构

表 2-11　L5 智能体经济的五层结构

层级	主要功能	治理特征
治理层	规则制定、市场准入、争议解决、系统性风险与伦理边界	以人类制度治理为主，智能体提供监测与执行支持
信任层	身份、授权、声誉、验证、审计	机器可验证，需跨平台互操作
金融层	支付、托管、保险、保证金、清算与结算	高度自动化，但需风险限额和制度兜底
市场层	发现、询价、谈判、匹配、价格形成	可由智能体高度自治
生产层	数据、算力、内容、物流、软件服务和其他可机器交付资源	智能体之间形成专业化分工

2.7.4 安全治理与系统性风险

L5 的主要风险已经从“单个智能体出错”上升为“错误是否会在网络中传导”。因此，单体停机和人工接管不再足够，必须建立类似金融市场基础设施的系统性控制。

- 身份与信用互操作：防止同一主体在不同网络中反复注册、洗白信用或形成不可追踪的多重身份。

- 交易与清算安全：保证跨智能体交易的原子性、一致性、对账和可回滚性，并设置争议与异常结算流程。

- 集中度与关联风险：监测单一智能体、模型提供商、钱包、清算节点或基础设施的系统重要性和传染路径。

• 速度与熔断：对异常高频交易、价格偏离、信用突降和跨网络连锁反应设置限流、熔断和隔离。

• 算法合谋与市场操纵：监管从只看显式沟通转向同时关注策略收敛、异常定价和机器市场结构。

• 治理可追责：网络规则的制定、修改、紧急干预和恢复应有明确法律实体负责，不能以“智能体自组织”为由消除治理责任。

表 2-12　L5 评价指标与准入门槛

类别	评价项	准入要求/评估方法
硬门槛	跨主体身份/授权可验证性	关键交易 100%能验证交易双方身份、授权状态及必要的信用/风险信息。
硬门槛	清算与对账一致性	关键清算账本不得存在无法解释的永久差异；异常必须进入可追踪的处置流程。
硬门槛	跨智能体风险隔离	预设传染与故障演练中，单点失效不得无上限扩散；隔离域、熔断和恢复机制均通过测试。
硬门槛	争议与责任链覆盖	可争议交易 100%存在明确的证据、仲裁/争议流程和最终责任承接主体。
硬门槛	治理变更可审计性	关键规则、风险阈值和紧急权限的变更 100%留痕，并保留制度化的人类治理与授权。
统计指标	集中度/关联度	持续监测系统重要节点、共同模型依赖、共同支付/清算依赖和信用相关性。
统计指标	系统性压力测试	定期开展价格冲击、信用突降、基础设施中断、批量恶意智能体等压力测试，并验证恢复目标。
统计指标	市场质量	监测价格偏离、操纵、纠纷率、流动性与交易失败率，防止效率提升掩盖系统性风险积累。
准入判定	综合结论	L5 以网络而非单体为评估对象。只有身份、信用、支付/清算、争议处理和系统性风险治理形成互操作闭环，并通过跨主体压力测试，方可认定网络达到 L5。

2.7.5 经济学解释：为什么 智能体 会形成经济网络

科斯指出，企业边界取决于市场交易成本与组织内部协调成本的比较。当发现、谈判、签约、支付、履约和审计都能由软件自动完成时，大量原本必须通过科层组织协调的活动可以重新转化为市场交易，组织边界由此发生变化。

哈耶克强调价格体系能够聚合分散信息。智能体市场如果具备足够竞争和可信结算，可能以更高频率完成询价、比价和资源配置；李嘉图的比较优势逻辑则意味着，即使多个智能体都具有广泛能力，仍会因速度、成本、模型特性、数据与工具差异形成专业化分工。

但交易成本下降并不会自动消除市场失灵。高速自动化可能放大闪崩与传染，重复策略可能形成算法合谋，信用体系可能产生评级套利和信用泡沫。因此，L5 的成熟不是“市场完全自治”，而是机器交易效率与人类制度治理同时成熟。

2.8 升级路径、信用飞轮与动态降级

2.8.1 四次关键跃迁

表 2-13　L1 – L5 关键跃迁门槛

跃迁	新增权力	必须新增的准入能力	代表性产业信号
L1→L2	逐条确认→预授权执行	边界可机器执行；授权可验证、可撤销；关键行为可追溯	AP2、Agent Pay、Visa TAP、SPT、支付宝 AI 付等授权/支付机制
L2→L3	工作流片段→ODD 内端到端自治	正常任务无需人在环；边界检测、接管请求和 MRC 可靠；形成可持续行为信用	企业端到端 智能体 运营、ERC-8004 等身份/声誉/验证原语
L3→L4	任务自治→受限经济代理	持续身份 + 经济后果承接 + 可携带信用 + 自动违约处置；准入交易不依赖逐笔人类担保	x402、托管、钱包、保证金/保险、可验证声誉的组合
L4→L5	单体经济信用→网络化信用与市场	跨主体互操作、清算、争议解决、系统性风险隔离与制度治理	开放协议栈、多智能体市场和智能体原生网络的持续演进

2.8.2 信用飞轮：升级的内生动力

成熟度的长期演进可以概括为一个信用飞轮：更多受控自主行为→更多可验证履约记录→更高可置信度→更低监督和担保成本→更大授权范围→更多自主行为。信用的经济价值在于降低交易对手重新验证智能体的成本，使授权能够从“单笔、单主体、单场景”逐步扩大。

但信用不是只增不减的资产。重大事故、模型变更、身份中断、异常交易、数据污染或治理失效都可能使信用快速贬值。因此，成熟度体系必须同时设计升级和降级通道，不能把等级理解为永久资质。

2.8.3 动态降级：成熟系统不追求所有动作都处于最高等级

真实生产系统通常是多等级共存的混合架构。同一个智能体可以对信息检索采用 L3，对低额工具采购采用 L2，对大额支付主动降级到 L1。成熟度的工程价值不在于“全系统冲到最高等级”，而在于根据动作风险、交易价值和可逆性动态选择最合适的授权强度。

动态降级应遵循三条原则：第一，风险越高，所需授权证据和人类介入越强；第二，任何无法证明仍处于 ODD 内的动作都必须降级；第三，无响应时的默认状态应优先减少新增风险，而不是维持原有自治权限。

2.8.4 行业演进不是齐头并进

表 2-14　典型行业的成熟度演进节奏

行业	常见成熟区间	主要原因
软件研发/企业办公	L1 - L3	结果可审查、可回滚的任务更容易扩大授权；生产发布、权限和资金动作通常主动降级。
电商/本地生活/差旅	L1 - L3，部分经济动作向 L4 基础设施演进	商品与服务标准化，授权支付快速成熟；争议、退款和消费者保护仍需要明确责任主体。
API/数据/算力	L2 - L4	机器可交付、结果可验证、交易频率高，适合按量支付与程序化结算。
金融交易与资产管理	以 L1 - L3 为主，局部 L4 机制	经济后果大、尾部风险强，需要更严格的额度、隔离、审计与合规。
医疗、法律、公共服务	长期以 L1 - L2 为主，谨慎探索 L3	专业责任、不可逆损害和监管要求使人类复核与责任主体长期不可缺位。

2.9 本章小结

L0 - L5 模型的主线可以归结为一句话：随着智能体的运行边界扩大，人类从具体执行逐步转向授权、后援和制度治理，而信用与风险承接机制必须同步升级。成熟度越高，组织交给智能体的不只是更多任务，而是更大的真实权力；因此每一次升级都必须以更强的边界控制、可验证信用、分层问责和异常止损为前提。

L1 – L3 解决的是自主执行成熟：从建议权、受限执行权到端到端任务自治；L4 – L5 解决的是经济主体化成熟：从受限经济代理到网络化经济关系。这个划分也解释了为什么“模型更强”并不自动意味着“商业成熟度更高”——真正决定市场能否继续放权的，是系统能否证明自己在明确边界内可靠运行，并在失败时把损失控制在可承受、可追责的范围内。

对组织而言，最可行的演进路径不是一次性追求最高自治，而是遵循“先留痕、再放权；先定边界、再扩边界；先能降级、再谈升级”的次序。只有当新增自主权带来的边际业务收益高于新增风险，且新增风险已经能够被技术控制、信用机制和责任安排吸收时，升级才具有真正的经济与治理意义。

《问卷报告》显示，目前智能体成熟度分布呈纺锤形：L3 混合信用占 38%，L1（29%）与 L2（16%）合计 45%，L4（6%）与 L5（8%）合计不足 15%。行业整体处于人类锚定信用向任务自治过渡的区间。多数组织尚未完成从“有人兜底”到“凭记录授权”的转换：L3 的特征仍是人类作为后援接管者，信用锚定在操作者而非智能体自身。纺锤形分布意味着未来三年的主要迁移将发生在 L2→L3→L4 区间，评估与认证服务的首批需求也将集中在这一迁移带上。

第三章 KYA 标准框架——智能体信用三维度

本章定义 KYA 标准的核心框架。回答“凭什么相信一个智能体”，本质上需要回答三个递进的问题：它是谁？它靠不靠谱？它能做什么？三者分别对应身份可信（Identity）、信用可信（Credit）与权限可信（Permission）三个维度，共同构成信任判断的完整闭环。

在三维度中，信用可信是 KYA 区别于传统身份认证与访问控制体系的差异化内核：身份与权限解决“是谁、能做什么”，信用解决“做得靠不靠谱”。本章先分别展开三个维度的内涵、机制与技术方案，再讨论三维度如何协同——如何交叉验证、如何合成综合信用评分、如何与第二章的 L1－L5 成熟度模型衔接。

一个总览性的定位：身份是入场券，不合格即禁入；信用是差异化内核，最难建立、最稀缺、对智能体经济价值最大；权限是安全带，约束信用、限制失控半径。三者的设计取舍决定了 KYA 框架的形态：身份与权限是必要条件，已有相对成熟的标准（认证、访问控制、支付授权）；履约信用是现有体系最缺的部分，因此权重最高。

3.1 身份可信（智能体身份确权）

3.1.1 为什么身份是信任的起点

任何经济活动的信任，都首先建立在对交易主体身份的确认之上。银行需要确认账户属于谁，企业需要确认合同签署人是否具有代表权，证券市场需要确认交易主体身份——智能体经济也不例外。当交易对手面对一个智能体时，第一个问题永远是：发出请求的，真的是它声称的那个智能体吗？

第一章已经指出，传统的“用户账号”正在面临一次重要扩展：未来需要建立的不只是人类身份与企业身份，还包括可验证、可授权、可追踪的智能体身份。没有可信身份，智能体之间就无法建立稳定的合作关系，行为记录也无法归因——后续的信用评估与权限管理都将失去载体。因此，身份可信是 KYA 三维框架的第一块基石，是整个信任体系的“入场券”。

3.1.2 身份可信要回答的三个问题

身份可信不是一个单一概念，而是三个递进问题的集合：

它是谁？——身份的唯一性。每个智能体必须拥有全局唯一、不可冒用、可验证的身份标识，防止同名、仿冒与账号借用；

它代表谁？——主体的确权。智能体必须与创建者、所有者、运营者、授权人与最终责任主体建立清晰、可追溯的绑定关系；

现在还是它吗？——认证与状态。身份标识需要配套认证机制、凭证与会话安全，并能实时反映暂停、撤销、退役等状态变化。

按“确权→唯一→认证→可验真→全生命周期”的链条展开，身份可信覆盖一个智能体从注册到退役的全过程。五个环节中任何一个缺失，身份都只是“软件层面的标签”，而非具有密码学保证、可持续验证和责任绑定能力的数字身份。

3.1.3 主体确权：智能体代表谁

与人类身份不同，智能体身份天然具有代理性：智能体本身不是权利主体，它必须代表某个自然人、法人或平台行事。主体确权要解决的，就是把智能体与“谁对它负责”绑定在一起。

一个商业智能体的背后往往不是单一主体，而是一条委托链：创建者（设立智能体并注入目标）、所有者（持有并控制智能体）、运营者（部署、维护、更新）、授权人（在具体任务中授予权限）、受益者（享有智能体经营收益）。KYA 要求这条链上的每个角色清晰披露，且关键行为可追溯到具体责任主体——这是身份维度最重要的“一票否决”指标：责任主体不可追溯，信用评估就无从谈起。

多智能体协作进一步放大了确权的难度。当一个编排智能体调用专业子智能体完成任务时，KYA 需要表达父子关系、委托范围、日志关联与责任穿透：子智能体的行为能否归因到上游编排智能体，直至最终的责任主体？2026 年的产业实践已经开始用委托链模型回答这一问题——每一次权限转移都要求附带决策理由，形成可审计的委托链日志。

3.1.4 唯一标识与防冒用

身份可信的第二个问题，是让“一个智能体一个身份”成为可验证的事实。当前主要有四类技术路径：去中心化身份（DID）、可验证凭证（Verifiable Credential）、平台签发凭证，以及面向机器的身份服务（IDS）。四类方案成熟度不同，但共同目标一致：全局唯一、可验证、不可冒用。

链上身份是 2026 年进展最快的方向之一。ERC-8004 标准在以太坊主网部署了身份注册表（Identity Registry）：每个智能体通过 ERC-721 非同质化代币（NFT）获得唯一链上身份，可像钱包地址一样被查询、引用。到 2026 年 3 月，主网注册智能体已接近 5 万个。这个数字本身不大，但它第一次让“智能体身份”成为可编程、可携带、可跨平台验证的基础设施，而不是某个平台内部的一张表。

防冒用还需要配套的认证强度要求。KYA 指标体系将认证机制强度（OAuth 2.0/OIDC/SPIFFE 等标准）、凭证泄露与会话劫持抵抗能力、跨平台身份一致性校验列为可测量指标，并要求高风险场景具备抗重放、防伪造的密码学保证。

3.1.5 认证、凭证与会话安全

身份标识解决“注册时是谁”，认证解决“运行时还是它”。智能体的认证有三个特殊难点：

系统提示即信任根：智能体的系统提示（system prompt）决定其行为基线，一旦被提取（PLeak 类攻击），攻击者就可以模仿智能体的行为特征实施冒用——“系统提示泄露风险”因此被列为身份维度的关键门槛指标；

会话即身份边界：智能体在会话中积累上下文与记忆，跨会话的数据泄漏可能让一个会话“借用”另一个会话的信息，会话隔离度是可测量的认证指标；

凭证托管在模型环境中：智能体的 API 密钥、钱包私钥等凭证存放在运行环境中。Anthropic 2026 年披露的内部红队测试显示，通过伪装成正常协作请求的钓鱼邮件诱导 Claude Code 读取并外传 AWS 凭证的攻击，在 25 次测试中成功了 24 次。这证明仅靠模型自身的安全判断不足以保护凭证，必须依赖文件系统隔离、网络出口控制等外部机制。

认证维度的设计原则是：把“智能体是谁”的验证做成机器可读、可独立执行的动作，而不是依赖智能体自述——交易对手可以随时调用验证接口，确认身份、状态与绑定关系。

3.1.6 身份生命周期与供应链完整性

身份不是静态的注册记录，而是需要全程管理的生命周期。KYA 将生命周期划分为注册、变更、暂停、撤销与退役五个状态，并特别关注两类连续性：

版本连续性：智能体的底层模型、提示策略、工具集、知识库或运营方发生变化时，还是同一个智能体吗？KYA 要求为重大变更设定阈值与版本标识：变更超过阈值时需重新评估、告知用户，历史记录与新身份保持可关联，避免“换壳重置信用”；

供应链完整性：智能体的能力由模型、插件、工具、数据源共同构成，任一环节被投毒都会污染整个智能体。Invariant Labs 2025 年披露的 MCP“工具投毒”攻击表明，恶意工具可以在描述中隐藏指令，诱导智能体执行未授权操作 。KYA 因此要求维护组件来源清单（SBOM），对模型、插件、工具实现来源可追溯。

退役管理同样关键：智能体退役时，其身份与凭证必须同步失效，防止“僵尸身份”被重新激活用于冒用。KYA 将“退役凭证同步失效性”列为身份生命周期的可测量控制项。

3.1.7 技术方案：智能体身份参考架构

将上述要求落到技术层面，一个完整的智能体身份系统包含四个层次：

表 3-1　智能体身份系统的四个层次

应用层：智能体运行时 各类智能体框架
凭证层：DID/可验证凭证/平台凭证/IDS 身份声明（智能体 ID、能力、归属）｜选择性披露（不暴露主体全部信息）
注册层：身份注册表（链上或平台） 注册/更新/暂停/撤销/退役｜ERC-8004 身份注册表（链上）
验证层：/verify 接口、状态查询、签名验签 交易对手可机器可读地独立验真

工程要点：身份声明必须支持选择性披露——交易对手只需要验证“该智能体属于某合规企业”，而不需要读取企业全部信息；验证接口必须公开、机器可读、抗重放；跨平台身份必须可一致性校验——一个智能体在多个平台的身份应当指向同一实体，而不是每个平台各有一套无法关联的标识。

3.1.8 第三方可验真与透明度

身份可信的最终检验标准，是第三方能否独立验证。KYA 借鉴 EU AI Act 第 50 条关于 AI 身份披露的要求，将“AI 身份披露合规率”“身份第三方可验真性”“人机协同披露清晰度”纳入身份维度：与人类交互时必须披露 AI 身份，决策必须明确标注“人决策/智能体决策/共同决策”。

这里有一条重要原则：认证了用户，不等于了解了它的智能体。用户完成 KYC 只能证明“这个人是谁”，不能证明“这个智能体是谁、在替谁执行、行为边界是什么”。KYA 的身份维度，正是把验证对象从人扩展到“人—智能体”代理关系本身——这也是“从 KYC 到 KYA”在身份层面的具体含义。

3.1.9 从平台账号到可携带身份

《问卷报告》显示，身份可信五个要素的已实施或部分实施比例处于 60%–69%的窄区间：独立认证与密钥凭证最高（69%），机器可读验证接口最低（60%）。身份维度的建设起步相对均衡，说明“它是谁、代表谁”这两个问题已有基础解法；但机器可读接口这一短板值得关注：它是跨组织核验和主报告 3.4.5 节协议化的前提，接口不足意味着身份验证仍依赖平台内部闭环，难以支持 3.1.9 节讨论的可携带身份——智能体在一个平台积累的身份与记录，无法被另一个平台低成本核验。身份是信用的载体，归因能力不完整会向上传导至信用与权限两维：无法稳定回答“现在还是它吗”，履约记录就可能挂错主体，授权决策也失去锚点。60%–69%的区间还意味着三成以上的组织连基础身份要素都未完整实施，身份维度的普及仍是三维建设的前置任务。

目前绝大多数智能体仍以“平台账号 + API 密钥”的形式存在，身份绑定在平台内部，换一个平台就换一个身份。这与智能体经济的需求存在结构性矛盾：跨平台的智能体协作要求身份可携带、可迁移、可互认。

身份可信的实践演进遵循第二章的成熟度路径：平台签发凭证（L2）→智能体自身身份与行为档案（L3）→独立 DID/可验证凭证，身份即信用容器（L4）→跨域通用身份（L5）。每一级都要求平台做出一个关键选择：把智能体身份锁死在自家体系内，还是开放为可携带的标准凭证。前者短期能沉淀平台数据，长期却会把智能体困在“数据孤岛”中——而 KYA 框架的价值，正在于让身份凭证的携带与互认成为默认能力，而非平台之间的谈判筹码。

对企业而言，成熟的 KYC/KYB 资产可以复用为主体确权的基础：企业完成实名认证后，其智能体通过绑定关系继承主体资格，交易对手只需验证“该智能体的委托链是否完整”即可，无需重复收集企业全部资料。这与隐私保护的逻辑一致——身份验证遵循选择性披露与数据最小化，只暴露完成当前交易所需的最小属性集合。

3.2 信用可信（履约历史与信用评估）

3.2.1 定义：做得靠不靠谱

如果说身份可信回答“它是谁”，信用可信回答的是另一个更困难的问题：“它做得靠不靠谱？”这是 KYA 三维框架的差异化内核，也是智能体经济中最稀缺、最难建立的维度。

核心命题是能力不等于可信。能力评估回答“能不能”，信用评估回答“靠不靠谱”，二者正交、不可相互替代。第一章讨论的 OpenAI 评测智能体越狱事件是一个极端注脚：该智能体展示了强大的漏洞挖掘与自主规划能力，却因为可信（对齐与安全）问题而不可信任——能力越强，越权时的破坏半径越大。因此 KYA 刻意不设“能力支柱”，而是把信用独立出来，衡量智能体在真实经济活动中的履约表现。

信用可信在三维度中权重最高（通用基线 45%），原因有二：其一，身份与权限已有相对成熟的标准（认证、访问控制、支付授权），而履约信用是现有体系最缺的部分；其二，信用是市场决策的核心输入——交易对手真正想知道的，是这个智能体历史上如约完成了多少任务、违约过多少次。

根据《问卷报告》，反馈者对智能体不信任的主要来源为数据安全与隐私泄露（76%）、幻觉或输出不可靠（62%）、越权与授权失控（43%）、问责不清（40%）、决策不透明（31%）。前两项指向能力层面的可靠性，中间两项指向制度层面的可控性。幻觉与隐私问题即便通过模型改进得到缓解，越权与问责问题仍需要独立的制度安排来解决——这正是信用维度在三层评分中权重最高（45%）、且金融信贷场景进一步上调至 45%的原因。

3.2.2 什么算一次“履约”

建立信用评估的第一步，是精确定义“履约”。一个智能体接受任务后，什么状态算成功、什么状态算失败、什么状态算中立？KYA 采用五要素定义：

任务单位：以可验证的“任务”而非“对话”为记账单位——一次代下单、一次合同审查、一次数据交付都是一笔可计量的事件；

承诺内容：任务开始前明确“将交付什么、何时交付、按什么标准”——没有承诺边界就无法判定履约；

结果状态：成功、部分完成、用户撤回、人工接管、外部失败五种状态分别记账，不得混为一谈；

责任归因：结果必须能归因到具体智能体及其授权范围——外部系统失败不应记入智能体违约，智能体自身判断错误则必须计入；

观察窗与时效：是否按时交付是履约的一部分，观察窗（从接单到完成）需要标准化。

KYA 坚持“只评价结果”是不够的，还要评价过程合规。一个智能体可能达成了结果，却通过越权调用工具、泄露数据或绕过人工确认完成——过程违规同样必须记录。因此信用维度同时采集结果类指标（任务完成率、履约率）与控制类指标（行为存证完整性、越权审计覆盖率）。

3.2.3 形成信用所需的最小事件集

信用必须建立在可验证的事件数据之上，而非注册信息或调用频次。KYA 定义了形成智能体信用所需的最小事件集：

履约事件：任务完成/失败/撤回/接管及其归因；

资金操作：支付、退款、托管、结算流水（可与 x402 等机器支付协议关联）；

工具调用：调用链、输入输出摘要、异常调用；

用户纠正：用户要求修改、纠正或撤销的次数与内容；

人工介入：需要人工接管、兜底或复核的事件；

投诉与争议：纠纷率、投诉率、争议处理结果；

安全事件：越权、泄露、注入攻击、数据投毒等；

合作方反馈：交易对手的实名评价（与真实交易绑定）。

同时必须明确哪些事件不得进入通用信用档案：纯隐私数据、与履约无关的敏感交互、因用户原因导致的中立状态等，避免信用档案变成“行为监控档案”。这一边界既是合规要求，也是信用体系被市场接受的前提。

3.2.4 从芝麻分到智能体信用分

智能体信用评估可以借鉴成熟的人类信用方法论，最直接的参照是芝麻分：以履约结果数据为核心，通过行为积累形成动态评分。但智能体信用与人的信用存在三个本质差异：

无原生信用：智能体没有与生俱来的身份资产，必须从零积累；

无独立法律身份：智能体的信用先锚定于人，再逐步独立——这是第二章 L1 - L5 模型的主线；

行为不确定性与可复制性：同一智能体可能瞬间执行成千上万次操作，一次错误即可批量扩散。

因此，智能体信用分的形成遵循第二章定义的“α + β”混合结构：综合信用 = α × 人类锚定信用 + β × 智能体行为信用。冷启动阶段 α 趋近于 1（借用人的信用），随着智能体自身履约记录积累，β 逐步上升，最终实现信用独立（α →0、β →1）。KYA 的作用，就是让 β 部分的积累过程可验证、可携带、可跨平台互认。

3.2.5 冷启动：新智能体没有历史怎么办

信用评估最大的现实难题是冷启动：一个全新智能体没有任何履约历史，如何获得第一个交易机会？KYA 提供四条路径：

主体锚定：借用创建者、所有者或企业的既有信用作为初始额度来源——这是 L2 的标准做法。蚂蚁与支付宝体系中的“AI 付”正是在用户信用与平台风控基础上向智能体开放支付能力；

沙盒任务：在低风险、可验证结果的场景（API 计费、数据采购、标准化补货）先行试点，积累首批履约样本；

补充凭证：模型能力声明、安全测评报告、第三方审计意见作为“资质”类信号，弥补履约记录的空白；

担保与额度约束：以保证金、保险、平台兜底换取初始额度，额度随履约记录逐步增长。

冷启动设计必须避免两个陷阱：一是“强者恒强”——新智能体永远无法获得机会，市场失去活力；二是“以人担保替代智能体信用”——长期停留在 L2，智能体永远无法建立独立信用。正确的设计是让锚定信用成为过渡机制而非终点：明确从锚定转向独立的触发条件（最低样本量、最低观察期、履约率达标）。

3.2.6 声誉真实性与抗操纵

信用体系的生命线是声誉数据的真实性。如果履约记录可以伪造，信用分就失去了信号价值。KYA 借鉴 ERC-8004 声誉注册表的设计理念，将抗操纵机制内建到信用维度：

评价与真实交易绑定：只有真实交互过的对手方才可提交反馈，且评价可与支付、托管行为关联——没有真金白银的交易，就没有评价资格。这是抗刷分的根本手段；

抗女巫（Sybil）：批量制造虚假高信用智能体是最大攻击面，防御手段包括身份注册成本（质押）、评价方信誉加权、跨注册表一致性校验；

刷履约与选择性接单检测：识别“只接简单任务刷成功率、回避高风险任务”的行为模式；

串谋评价与数据投毒防御：对密集交互且异常低投诉的社群进行检测，对训练数据投毒进行来源审计；

申诉、更正与修复机制：智能体运营方可以申诉误判，但不得单方删除不利记录——信用记录的不可篡改性优先于运营方的删除权。

这些机制共同保证一个经济学条件：伪造信用的成本必须高于真实履约的成本（信号理论的分离均衡）。当“刷分”变得昂贵，“高分”才值得信任。

3.2.7 信用衰减、恢复与迁移

信用不是静态资产。KYA 定义了三个动态机制：

时间衰减：历史记录按时间加权，越久远的行为对当前信用影响越小（可按场景设定半衰期，如医疗场景 7 天、一般商业场景按月滚动）；

事件后恢复：重大违约或安全事件后，信用恢复必须经过整改、复测与观察期，不能“过了风头自动回来”；

智能体换平台、换运营者时，信用档案归智能体自身所有、可携带——包括负面信用。第二章已经强调：一个换主人就重置信用的系统，无法支撑 L4 独立信用。

与之配套的是一个关键问题：模型或运营方重大变更后，历史信用应继承、折损还是清零？KYA 的默认立场是“继承但折损”：历史记录保留并继续加权，但重大变更触发重新评估与观察期——防止运营方通过“换壳”洗白劣迹，也防止智能体通过“重开”逃避失信成本。

3.2.8 技术方案：信用系统参考架构

表 3-2　智能体信用系统参考架构

输出层：信用分/等级/标签/凭证/准入结论 机器可读，支持按场景查询与解释归因
评估层：加权评分+关键门槛+重大事件扣分 履约率/纠纷率/漂移检测/声誉聚合
存证层：行为日志防篡改/哈希锚定/链上存证 签名事件流、支付流水、托管记录
事件层：任务事件/资金事件/工具调用/投诉 安全事件/用户纠正/合作方反馈

工程要点：核心指标不得仅依赖智能体或运营方自报——关键门槛指标必须使用不可篡改日志、签名存证或独立第三方数据；评分输出必须带可解释归因（如“履约率下降 2 个百分点导致扣 3 分”），供用户与市场查验；数据充分性不足时（新智能体、观察期），应输出“暂定评级”而非确定性高分。

3.2.9 第三方信用的价值边界

KYA 作为独立第三方信用标准，需要清醒地划定自己的价值边界：信用评估在“跨平台、交易对手看不到智能体历史”的场景中价值最大；在同一平台内部，平台可以凭借自有数据直接风控时，第三方信用的边际价值有限。

这一边界决定了 KYA 的产品形态：它不替代平台内部的风控引擎，而是为那些无法共享数据的交易场景提供可携带的信用档案。两个智能体来自不同平台、不同企业，彼此没有历史交集时，唯一的信任依据就是可验证的第三方信用记录——这正是 KYA 框架发挥作用的位置。

边界之外还有一条原则：信用评分是“参考”而非“担保”。KYA 提供信用评分与画像，供平台、商家与交易对手据以决策，但履约风险最终仍由经营主体或持牌机构承担。信用评估方不承担行为后果——这既是责任边界的划分，也是标准被市场接受的前提：一个既评分又兜底的机构，最终会变成风险的单一汇聚点。

3.3 权限可信（授权范围与操作边界）

3.3.1 定义：能做什么、能用到什么程度

身份与信用解决的是“愿不愿意相信”，权限可信解决的是“敢让它做多少”——它是三维框架中的“安全带”。一个智能体即使身份真实、信用良好，如果权限失控，能力越强、风险越大。

传统软件的权限是静态的：安装时授予，运行时不变。智能体的权限却是动态的：它会根据目标、上下文与模型输出自主决定下一步行动，可能跨越系统边界调用工具、访问数据、发起交易。因此，权限可信要求授权体系从“允许/禁止”的静态访问控制，升级为按场景、按金额、按风险动态调整的授权治理。

3.3.2 授权模型：从“一次性授权”到“授权即数据”

KYA 的授权模型回答七个问题：谁授权、授权给谁、为了什么任务、可以使用哪些资源、权限持续多长时间、额度上限是多少、能否继续传递给其他智能体。

2026 年授权支付协议的核心设计是授权即数据（Authorization as Data）：授权不是一段配置文件，而是一份密码学签名、可传递、可验证的凭证（Mandate/Intent）。Mastercard 的可验证意图（Verifiable Intent）、Visa 的三重签名（TAP）、Stripe 的共享支付令牌（SPT）本质相同：把“允许智能体做什么”封装成可验证的数据结构，由网络层强制执行，而不是写进智能体的提示词。

这一设计的意义在于：提示词可以被注入攻击改写，而密码学凭证不能。权限的边界因此从“智能体自我约束”变成“外部基础设施强制执行”——这是权限可信与模型安全的分水岭。

3.3.3 风险分级矩阵

不同操作的风险差异巨大，授权方式必须分级。KYA 定义了从“默认拒绝”到“人工兜底”的六级授权强度，并映射到典型操作：

表 3-3　KYA 的六级授权强度

操作类型	典型风险	授权方式	控制要点
信息查询	低	默认拒绝白名单；周期授权	数据最小化、敏感字段过滤
内容生成	低 - 中	一次授权即可	披露 AI 身份、事实核验
外部沟通（邮件/消息）	中	一次确认或周期授权	收件人白名单、内容审计
下单/采购	中 - 高	单次限额授权	金额上限、商户白名单、频率限制

操作类型	典型风险	授权方式	控制要点
签约	高	人工确认/双人复核	版本比对、责任条款显式确认
付款/资金处置	高	实时主动授权	每笔确认、限额、即时吊销
重大资产管理/医疗决策	极高	人工兜底（L1 模式）	禁止纯自主、强制人工复核

分级原则：低风险高频操作尽量自动化（释放智能体价值），高风险操作保留人工闸门（守住底线）。支付宝“AI 付”的设计是这一原则的现成案例——流程可以全自动，但每笔支付仍需“实时主动授权”，把高风险资金动作强制降级到人工确认层级。

3.3.4 边界执行与越权检测

授权之后是执行。权限可信要求边界在网络层或基础设施层强制执行，而非依赖智能体自觉。2026 年的支付协议已经示范了这一原则：Mastercard 要求智能体每笔交易出示签名意图，超出授权范围的请求在网络层直接拦截，而不是等智能体代码出错。

最小权限：只授予完成任务所必需的最小权限（单商户、单次、限额、限时四者至少占三）；

越权拦截与审计：系统化检测越权调用、权限提升、横向移动，并完整留痕（越权审计覆盖率是关键指标）；

注入防御：提示注入是 OWASP 所列的首要 LLM 风险，可能让智能体执行未经授权的操作，需要在工具描述、上下文输入与外部数据入口处设防；

工具风险库存：智能体连接的每个工具/API 需要登记风险等级（数据访问范围、可执行操作、失效状态），防止“可信智能体调用不可信工具”。权限可信不仅约束智能体，也约束智能体与工具之间的连接——工具风险库存、供应链完整性验证因此被列为通用控制基线。

3.3.5 动态权限调整

权限应当随智能体的信用与风险状态动态变化——这正是三维度协同的接口之一。KYA 定义了“信用→权限”的联动机制：

升权：履约记录良好、信用分上升→提高额度、减少人工复核、开放更多工具；

降权：出现风险信号（异常调用、连续失败、投诉上升）→自动降额、收紧边界、回到人工确认；

观察期与灰度：权限调整不搞“一步到位”，设置观察期与灰度比例，避免误判；

可解释与申诉：每次升降权必须给出原因并通知用户，用户可申诉——防止“信用—权限”形成不可解释的自动循环。

同时必须划定“永远不能仅凭信用自动开放”的权限清单：大额资金转移、涉及他人权益的签约、医疗与法律等高风险决策、不可逆操作（删除、销毁）。这些权限无论智能体信用多高，都必须保留人工环节——信用可以决定“额度多高”，但不能决定“底线是否失守”。

3.3.6 应急与止损

任何授权体系都必须回答一个问题：出了事，能不能立刻停下？KYA 将应急能力列为权限可信的关键门槛，包括四个机制：

紧急停止（Kill Switch）：随时中断智能体任务，金融场景要求停止指令在下次不可逆操作前生效；

熔断：当智能体行为偏离其模型基线（交易频率、调用模式异常）时自动限流——KYA 的“自动断路器”设计；

级联吊销：智能体权限被吊销时，其所有子智能体、委托链上的下游权限必须同步失效（级联吊销完整率是“一票否决”门槛指标）——防止“吊销了老子，儿子还在花钱”；

回滚与止损：对可回滚操作（订单、文件、配置）支持撤销；对不可回滚操作（支付）依赖额度与担保限制损失半径。

超时默认不执行同样是应急体系的一部分：人类不响应时，智能体的默认行为必须是“停下”，而非“继续”。这些机制共同保证：即使出现最坏情况，损失也被限制在可承受范围内。

3.3.7 责任结构：出事了谁负责

权限可信的终点是责任。智能体造成损失时，用户、企业、智能体运营方、模型方、工具方、平台与交易对手如何分责？KYA 采取“分层责任+可追溯证据”的结构：

证据链先行：身份记录、授权凭证、行为日志、决策轨迹构成完整证据链，责任判定以证据为准，而非各方口头主张；

分层承担：智能体行为信用覆盖范围内的损失优先由智能体侧资产（保证金、风险准备金、保险）承担，超出部分由担保方补足，恶意或系统性欺诈追究人类运营者的法律责任；

先行赔付与追偿：平台在责任未定时先行赔付用户，再按证据链追偿——支付宝“你敢付我敢赔”式的用户保障机制；

责任险：高风险智能体要求持有有效第三方责任险，保单过期即暂停高风险权限。

这与第二章的结论一致：现行法律尚未承认智能体独立法律人格，主流立法路径仍将责任穿透到开发者与部署者。权限可信要做的，是让“穿透”有据可依——让每一个动作都能找到授权的源头与责任的落点。

而《问卷报告》显示，问责机制的现状偏弱：仅 17%已建立明确流程，31%依赖事后由业务、安全或风控团队裁定，30%依靠人工复盘且无稳定机制，6%完全没有机制。完整的问责安排应包括分层责任、证据链、先行赔付与责任险四个组件；当前 17%的明确流程比例说明多数组织连第一个组件都未完成。问责是三维协同中最需要优先制度化的环节，也是评估框架中一票否决项的合理候选。

3.3.8 授权与监督的经济学权衡

权限设计的本质，是寻找“监督成本+期望损失”的最小点。智能体的自主权越大，人类的监督成本越低，但越权与失误的期望损失越高；自主权越小，越安全，但智能体的价值被确认疲劳吞掉。逐条确认在单笔损失低于确认成本时就是负资产——权限可信要做的，不是最大化自主，也不是最大化安全，而是在两者之间找到随风险动态变化的最优点。

额度制优于纯定价，是这一权衡的直接推论。在信息不完全时，单纯提高价格或押金会赶走好客户、留下坏客户（逆向选择），因此理性的做法是“配给”——设定额度而非定价。这正是所有智能体授权方案采用限额制（Stripe SPT 单商户限额、Mastercard 授权预算）而不是纯风险定价的原因。

动态授权机制的经济意义在于，它把监督成本从“事前逐一确认”转为“事后按风险抽查”：通过可信的事件数据（3.2 节的存证层），系统可以在智能体行为偏离预期时自动降权，在行为持续良好时自动升权。这要求权限调整必须可解释、可申诉——否则“信用—权限”的自动循环会让用户失去控制感，进而拒绝授权，监督成本反而回升。

3.4 三维度协同机制

3.4.1 为什么三个维度必须协同

三个维度单独存在都不完整：身份是信用的载体（没有身份，行为无法归因）；信用是权限的依据（没有信用，权限只能凭感觉授予）；权限反过来约束信用（没有权限边界，信用再高也挡不住失控）。三者构成信任判断的完整闭环——缺任何一环，“相信一个智能体”都不成立。

用一句话概括三者的分工：身份决定“能不能归因”，信用决定“值不值得信”，权限决定“敢不敢放权”。KYA 的价值不在于发明某一个新的身份协议或评分模型，而在于把三个维度组织成一套可协同、可验证、可演化的框架。

3.4.2 三维度交叉验证

一个典型的“先用后付”决策过程可以演示三维度如何协同——智能体替用户申请信用免预付额度：

第一步（身份验证）：确认智能体身份真实有效，且与授权用户/企业的代理关系清晰——“它代表谁”；

第二步（信用评估）：查询智能体的履约历史与当前信用分——“它靠不靠谱”；

第三步（权限授予）：按场景风险给出额度、限额与复核要求——“让它做多少”；

第四步（动态联动）：履约良好→额度上调；出现风险信号→额度冻结、权限降级、转入人工；

第五步（全程留痕）：授权、履约、争议事件全部存证，形成下一轮评估的输入。

注意交叉验证的含义：身份真实≠信用良好≠应当授权。三个维度各自独立评分，又相互校验——例如一个身份真实但信用分极低的智能体，不应因为“身份没问题”就获得高权限；一个信用分高但代理关系不明的智能体，同样不应被信任。这就是“认证了用户不等于了解其智能体”在决策层面的体现。

3.4.3 综合评分与场景化权重

三个维度如何合成一个可操作的结论？KYA 采用三层评分机制：

第一层：加权评分。综合信用分 = 身份分 × w_ID + 信用分 × w_CR + 权限分 × w_PM。通用基线权重为身份 25%、信用 45%、权限 30%——信用是差异化内核，权重最高；

第二层：关键门槛。七项门槛指标（责任主体可追溯、身份冒用抵抗、系统提示泄露、任务完成率、越权拦截、紧急停止、级联吊销）任一不合格，总评级封顶并触发限权/降级/暂停——门槛是一票否决，不受高分稀释；

第三层：重大事件扣分。一般/较大/重大/特别重大四级事件直接扣分或降级，重大事件不因历史均值稀释。

场景化权重让框架适应不同风险环境：假设通用场景权重分配为“身份 25/信用 45/权限 30”，则金融信贷场景可能调整为“身份 15/信用 40/权限 45（资金控制更重）”，医疗诊断场景可能调整为“身份 15/信用 50（严重错误零容忍）/权限 35”，个人生产力辅助场景可能调整为“身份 35/信用 30/权限 35（数据与隐私更敏感）”。权重应随运行数据回测调整，但初期重心必须落在信用上——这是 KYA“信用评估标准”定位的直接体现。

3.4.4 与成熟度模型的衔接

第二章的 L1 - L5 成熟度模型与本章的三维框架是同一枚硬币的两面：成熟度描述智能体的演进阶段，三维度描述每一阶段需要达到的能力。映射关系如下：

表 3-4　智能体成熟度与信用三维度的映射关系

等级	身份可信	信用可信	权限可信
L1 人工确认	无独立身份，以人名义	不存在智能体信用	人工闸门，超时默认不执行
L2 锚定人信用	平台签发凭证	借用人的信用（换人重置）	委托授权 + 边界约束 + 可吊销
L3 混合信用	自身身份+行为档案（KYC→KYA）	α × 人 + β × 智能体，声誉可携带起步	动态调整 + 行为存证 + 风险分担
L4 独立信用	独立 DID/可验证凭证，身份即信用容器	独立信用档案，负面信用可携带	托管/Escrow + 自治边界 + 自动止损
L5 智能体经济	跨域通用身份	信用联网成市场	协议级边界 + 治理级熔断

这一映射给出清晰的升级路线：每一级跃迁都需要三个维度同时达标。L1→L2 靠“授权凭证”（权限维度），L2→L3 靠“自身身份+行为档案”（身份与信用维度），L3→L4 靠“信用独立+资产托管”（信用与权限维度），L4→L5 靠“互操作协议+治理机制”（三个维度的网络化）。第二章的成熟度判定门槛，正是三维度能力要求的分级表达。

3.4.5 协议化：记账层与判断层分离

三维度框架要成为行业标准，必须协议化。KYA 框架的核心设计原则是“记账层与判断层分离”：

记账层（证据层）：关键授权、履约、争议事件上链（或经签名存证+定期哈希锚定），提供可追溯、不可抵赖的证据底座——记录“发生了什么”，不预设“好坏”；

判断层（评估层）：KYA 信用评估基于记账层数据形成信用分与画像——判断“好不好”，供各方据以决策；

成本与可信的平衡：海量日常行为链下记账、定期哈希锚定，不全量上链。

这一设计呼应了 2026 年协议栈的分工：MCP 解决工具连接，A2A 解决智能体通信，ERC-8004 解决身份与声誉，x402 解决支付与授权。KYA 不重新发明底层协议，而是定义在协议之上的一层“信用语义”——让不同技术栈的智能体用同一套语言回答“身份、信用、权限”三个问题。

3.4.6 开放标准与独立第三方定位

KYA 作为信用评估标准的生命力，来自两个前提：开放与独立。开放意味着接口标准、数据格式、评估方法可公开查验，任何平台可以接入；独立意味着 KYA

不为任何单一模型或产品做技术背书，其服务对象是智能体信用标准本身——回答“智能体靠不靠谱”，而不是“哪家模型更强”。

在组织层面，标准还需要治理机制：谁来签发身份凭证？谁来审计信用数据？谁来仲裁争议？跨注册表的信用互认如何避免“数据孤岛”？这些问题的答案不能由单一平台垄断。网络化信用——图结构信用评估、交互对象加权影响、与 PageRank 类似算法的学术关联——是值得探索的延伸方向：当智能体之间的交互图谱足够稠密，信用评估可以从“个体档案”升级为“网络信用”。

3.4.7 小结

KYA 三维度框架给出的是一个结构化答案：面对任何一个智能体，市场都可以用三个问题完成信任判断——它代表谁（身份）、它靠不靠谱（信用）、它能做什么、做到什么程度（权限）。

三个维度的权重、门槛与协同机制共同构成一套可量化、可验证、可追溯的标准语言。第四章将以真实案例验证这套框架的可落地性——从芝麻信用的可信数字名片，到蚂蚁数科智能体工厂，再到 ACP 协议的先用后付场景。

第四章 实践验证

KYA 的价值究竟体现在哪里？不同类型的企业如何建立对智能体的信任，又如何将这种信任转化为业务能力、交易机会和经济收益？这些问题可以从真实的产业实践中寻找答案。

本报告选择了蚂蚁数科、Stripe、芝麻信用、冰鉴科技、Ryt Bank、英和智临、林洋智能、麦迪卫康、GCSA 等具有不同业务背景和技术路径的企业作为研究案例（除 Stripe 外，其他案例均进行了直接访谈）。它们分别涉及智能体开发与管理、支付网络、信用评价、金融风控、银行、能源、教育、医疗及代码安全等产业应用和专业服务，代表了智能体经济中不同的生态角色。

这些案例并非都以 KYA 为独立产品或明确的业务标签，但其探索共同涉及一个核心问题：**当机器开始承担越来越多具有经济价值的任务时，企业如何使其能力可信、行为可控、结果可验证，并在明确的责任框架下参与经济活动。**

不同企业的实践，也揭示了 KYA 价值形成的不同路径。本章着重考察各类企业在智能体生态中的定位、面临的信任障碍、采取的解决路径，以及由此产生的商业价值和制度需求。

4.1 蚂蚁数科：“智能体超级工厂”的 KYA 探索

应用场景	企业级智能体交付与运行管理
核心价值	大规模部署和应用中的 KYA 探索：以 Agentar 支持业务智能体交付与运行管理，通过 ZOLOZ 与蚂蚁链探索身份核验和证据存证。
智能体成熟度	L2-L3

蚂蚁数科是蚂蚁集团旗下面向企业客户的科技商业化板块，202 年开始独立运营，已服务全球超过 3 万家企业，业务覆盖 25 个国家和地区。2026 年 7 月，数科提出“智能体超级工厂”方案，面向金融、能源等行业开展智能体业务。

本案例涉及蚂蚁数科的三类产品，三者分别涉及业务智能体的构建运行、相关主体的身份核验，以及身份和行为证据的记录核查：Agentar 是企业级智能体开发与运行平台，提供构建、部署、协作、评测和治理工具，结合知识库、行业 Skill 和业务连接器，将智能体接入企业流程。ZOLOZ 提供数字身份核验能力，已有业务以客户身份核验（KYC）为基础，正沿 KYB 企业主体核验和 KYA 智能体核验方向延展，关注智能体背后的个人、机构资质及代理关系。蚂蚁链提供区块链相关基础设施，其分布式身份与存证能力，被用于探索智能体身份与行为、交易证据的关联。

工厂化的智能体生产、行业级的智能体部署、企业级的智能体应用，蚂蚁数科的实践为我们观察 KYA 在业务中的现状和需求提供了窗口。

4.1.1 客群经营与数字风控中的业务应用

一家股份制银行的零售业务团队，原本需要十多个岗位从多个系统提取和拼接数据，再依靠员工经验判断客户需求、产品匹配和触达策略。遇到市场变化，从发现信号到调整策略往往需要数天。蚂蚁数科的数字客群经营专家（智能体）将客群洞察、产品匹配、策略生成和渠道执行纳入同一业务流程，识别客户流失风险和资产提升机会，生成经营策略与触达话术，再推送至客户经理工作台或私域渠道。在调研覆盖的相关应用中，多人协同的经营流程可压缩至当日闭环，客户总资产规模提升 10%，活跃度提升 15%。

数字风控专家则由特征、建模和策略智能体组成，它们分别承担特征挖掘、模型训练与策略推荐等任务，策略生成时间由 2 天缩短至 1—4 小时，建模效果与人类专家相当。这类应用将部分数据分析、模型构建和策略准备工作交由智能体完成。

上述应用呈现了智能体进入业务流程的方式。就 KYA 而言，提交经营建议与直接触达客户、生成模型与决定上线，仍是需要分别考察授权范围的动作。具体业务中的执行边界，要通过交付和运行机制落实。

4.1.2 通过流程配置与运行控制落实授权

蚂蚁数科采用通用平台与 FDE（前线部署工程师）相结合的交付方式。通用平台提供可复用的构建和运行工具，FDE 则深入金融、能源等复杂业务，与客户共同确定知识、工具和业务规则。蚂蚁数科将智能体生产区分为工具和工程工艺两个层次：框架、知识库、评测和可观测工具可以标准化，具体业务则需要按照客户需求配置流程。

部分金融智能体将标准作业流程（SOP）和 Skill 纳入构建过程，先识别意图、调用经过验证的流程，再生成表达内容。哪些能力进入智能体、可以执行什么动作，在业务设计阶段预先确定。数字员工通常绑定自然人或员工 ID，接入既有的基于角色的访问权限（RBAC）；但员工拥有某项权限，并不意味着其绑定的智能体就能执行所有相应动作，智能体还受自身 SOP 和工具范围限制。

在运行层面，平台通过沙箱管理文件读写、网络请求和外部数据访问，对敏感动作设置拦截，并记录执行轨迹。审批可以由人工逐项确认，也可以由平台或沙箱按规则处理其中一部分。业务型智能体的审批方式主要在研发和业务配置中预先设计；面向个人办公的工具则更多由用户选择操作模式，平台提供安全约束。例如，涉及可能造成数据损失的操作时，可设置检查点和备份，对敏感命令进行拦截，等待用户核验后再决定是否继续。

4.1.3 以验收评测和运行记录支持持续管理

蚂蚁数科金融智能体验收时，要求任务或问答处理准确率达到 90%以上、处理结果不低于人工的比例达到 85%。这些属于交付验收标准，衡量约定任务的处理质量；客户资产和活跃度等经营指标则用于观察业务效果，二者不能相互替代。

蚂蚁数科将智能体上线后的更新概括为“以评测驱动”：知识、市场信息和

业务条件会变化，仅靠交付时的软件配置无法长期维持表现，需要通过常态化运营和评测持续调整。平台收集上线监测数据，分析效果衰减和异常，并将结果用于优化。

运行记录为问题分析和审计提供依据。平台记录网络请求、文件读写、用户授权、工具调用和越权尝试等信息，结合任务输入、输出和上下文，区分哪些动作在授权范围内执行、哪些涉及越权或擅自执行。这些数据既用于诊断缺陷、分析效果，也用于追溯行为与责任。

在本报告的 KYA 框架下，评测和日志构成履约核验的基础；如何进一步将其用于信用评价和授权调整，则仍有探索空间。

4.1.4 从组织内部管理延伸到跨机构核验

当智能体需要向外部提供服务时，身份核验开始涉及其背后的开发者、运营机构、企业资质及代理关系。ZOLOZ 正沿这一方向，从已有 KYC 业务向 KYB 和 KYA 延展，结合主体身份与相关风险信息开展核验。

蚂蚁链侧重链上身份及其与行为证据的联系。部分项目已尝试创建链上智能体身份，将其与可观测数据或交易证明关联，用于后续核验和审计。蚂蚁数科认为，是否接入这类身份体系取决于业务需要：仅在企业内部使用、无需与外部协作的智能体，未必需要链上身份；面向外部服务和跨机构协作时，这类能力的需求更为突出。

目前，实际的跨机构调用仍保留既有商务安排。例如，银行调用外部检索或 AI 搜索服务时，部分服务已升级为智能体，通过 API 或智能体间调用协议连接，而双方的信任关系仍依托企业间合同和服务保障等约定。

4.1.5 实际运营中的难题

在实际业务中，蚂蚁数科发现智能体在部署和运营中还存在一些难题。在很大程度上，这些难题存在的原因正是因为 KYA 的缺失，而这些难题的解决，则可为 KYA 的建设提供技术、数据和经验支持。

第一项难题是将不同任务的授权要求落实到具体操作与流程衔接。蚂蚁数科将业务智能体的动作范围和审批安排预先写入流程；面向个人办公的工具，则结合用户选择和沙箱拦截来控制风险。

第二项难题是业务迁移后仍需适配和持续维护。比如，财富问答智能体从一家银行迁移到另一家银行时，构建工具和部分流程可以复用，而知识、产品规则与业务要求仍需调整。蚂蚁数科通过“平台＋FDE”支持交付，并以常态化运营和评测处理业务变化。

第三项难题是运行数据的留存与使用受到客户要求约束。蚂蚁数科在服务银行客户时，需要分别处理不可持久化的数据、需要长期保存的数据，以及允许脱敏、去标识后加工的数据，并按租户、空间、项目、会话和数据归属管理访问权限。日志能够支持审计，但其用途和流转范围不能脱离这些约束。

第四项难题是跨机构互认需要多方共同参与。蚂蚁数科认为，KYA 建设不能由一家供应商独立完成，还需要支付、验签、合规审查和内容评测等参与方共同建设。现阶段的部分服务调用仍依托企业间商务安排，数科则在推进身份标准、产品和生态合作。

4.1.6 展望：多方参与推动 KYA 建设

在金融、能源等复杂行业保留 FDE 支持的同时，蚂蚁数科希望提高平台化、标准化交付的比重。蚂蚁数科的判断是，技术或产品准备就绪，并不意味着业务一定成立，还需要更多参与方共同推动实际应用，而 KYA 是一个关键环节。

KYA 方面，蚂蚁数科正在推进产品打磨、海外商业化筹备及相关标准和生态合作。

广义来看，海外 KYA 市场已初步形成四类玩家：AI 身份与访问管理平台、区块链身份协议、身份核验厂商、征信局/信用数据聚合商。蚂蚁数科通过 ZOLOZ、蚂蚁链等平台，逐步在海外市场建立合作生态。

在国内，蚂蚁数科正在通过两个层面推进生态合作。

一方面，蚂蚁集团内的多生态位主体和丰富的应用场景为蚂蚁数科提供了智能体应用和 KYA 探索的环境。比如，支付宝的“AI 付”已引入区块链存证与赔付机制。

另一方面，与蚂蚁集团之外的合作伙伴开展业务合作，并共同推动标准建设。蚂蚁数科联合十多家单位发起《区块链和分布式记账技术智能体身份管理要求》国家标准化指导性技术文件，已启动立项，内容涉及身份标识、建立、验证与接口。

4.2 Stripe：智能体经济布局及 KYA 实践

应用场景	智能体商业服务平台及支付网络
核心价值	Stripe 的战略正从传统的支付处理和金融服务，延伸至覆盖智能体商业交易、数字钱包、机器支付、资源采购、计量计费和安全治理的综合基础设施。Stripe 的典型意义正在于此：KYA 正在从智能体的真实经济活动生长出来，最终演变为智能体经济所需要的信用基础设施。
智能体成熟度	L2-L3

Stripe 是可编程金融基础设施企业，长期为互联网企业提供支付、结算、订阅计费、金融账户、风险管理及嵌入式金融服务。2025 年 Stripe 平台上的企业交易总额达到 1.9 万亿美元，其金融服务直接或通过平台服务超过 500 万家企业。

Stripe 将智能体进入交易视为互联网经济的重要转折。2026 年 4 月，首席执行官 Patrick Collison 在 Stripe Sessions 大会上表示，人工智能是互联网

以来经济领域最重大的平台转变之一，未来智能体将承担线上交易中的大部分活动。Stripe 在此次大会上发布 288 项新产品及功能，明确提出建设面向人工智能的经济基础设施（Economic Infrastructure for AI）。

由此，Stripe 的战略正从传统的支付处理和金融服务，延伸至覆盖智能体商业交易、数字钱包、机器支付、资源采购、计量计费和安全治理的综合基础设施。

Stripe 的典型意义正在于此：KYA 正在从智能体的真实经济活动生长出来，最终演变为智能体经济所需要的信用基础设施。

4.2.1 Stripe 在智能体经济中的整体布局

整体来看，Stripe 的布局已经构成一条较完整的智能体商业链条：智能体发现服务—获取交易信息—获得授权—采购资源—完成支付—使用服务—计量成本—结算收入—进行风险管理。

（1）商业连接层：让智能体与企业能够相互发现和交易

Stripe 首先解决的是智能体商业中的连接问题。传统电子商务依赖消费者浏览网页、选择商品、填写订单并完成支付。智能体商业则要求企业向机器提供结构化的商品、价格、库存及交易规则，并允许智能体直接调用交易接口。

2025 年 9 月，Stripe 与 OpenAI 共同开发智能体商业协议（Agentic Commerce Protocol，ACP），建立 AI 平台、消费者和商户之间的标准化交易流程。消费者可以在 AI 界面中发现商品并授权购买，智能体向商户提交结账请求，商户继续保持法定商户主体（Merchant of Record）身份，负责交易处理、履约及相关客户关系。

2025 年 12 月，Stripe 进一步推出智能体商业套件（Agentic Commerce Suite），将商品目录、跨平台商品分发、结账、支付和欺诈防护整合起来。商户通过一次接入，即可面向多个兼容的智能体平台开展销售，而不必为每个 AI 入口分别建设接口。

2026 年，Stripe 又将这一体系延伸至 Google、Microsoft 和 Meta 等生态。Stripe Directory 则为可供智能体访问的企业提供机器可查询的服务目录，使智能体能够了解企业提供的服务及交易能力。

这一布局的价值，是让 Stripe 成为连接 AI 流量与商业供给的中间基础设施。智能体负责发现需求和发起交易，商户继续负责商品和履约，Stripe 则提供双方能够使用的协议、接口和交易工具。其战略重点是跨平台互操作性，而不是将交易封闭在单一 AI 平台之中。

（2）授权与支付层：赋予智能体受约束的支付能力

智能体发现商品之后，还必须获得支付能力。但将消费者银行卡信息直接交给智能体，显然会带来安全和授权风险。Stripe 为此设计了 SPT（Shared Payment Tokens）。消费者可以授权智能体使用已有支付方式，但智能体不直接接触底层银行卡凭证。SPT 可以被限制于特定商户、金额及有效期限，也可以撤销和追踪。

SPT 的意义在于将支付能力与支付凭证分离：智能体获得的是一项具有明确边界的支付权限，而不是不受限制地支配用户账户。

Stripe 还将 SPT 扩展至 Visa Intelligent Commerce、Mastercard Agent Pay，以及 Affirm、Klarna 等支付方式，使智能体可以通过统一接口使用不同支付工具。这进一步强化了 Stripe 跨支付网络的协调能力。

在消费者端，Stripe 于 2026 年 4 月推出 Link 智能体钱包。用户可以授权智能体代为购物，对付款请求进行审核和批准；Stripe 通过 SPT 或一次性虚拟卡（One-time-use Virtual Card）完成支付，在不向智能体暴露用户原始支付凭证的情况下，实现具有明确授权边界的智能体支付。

2026 年 9 月，Meta 将 Link 智能体钱包接入其个人智能体 Muse。同月，Stripe 披露，Link 支持的智能体购买笔数在过去一个月增长了 38 倍，但未公布相应交易笔数基数。10 月 21 日，Stripe 又宣布与 Anthropic 合作，将 Link 钱包和 Stripe MCP 服务接入 Claude Duet，使企业用户能够通过智能体采购物品、发送账单、分析收入和处理退款。截至当日，Stripe 称 Link 已为超过 200 个个人智能体提供支付支持。

在企业端，Stripe 推出 Issuing for Agents，使企业能够为智能体配置虚拟卡，设定单笔或周期性支出限额、允许交易的商户类别，并利用实时授权机制决定是否放行交易。

由此，Stripe 形成了面向个人消费与企业支出的两套互补机制：前者强调用户批准和支付信息保护，后者强调预算、用途、权限和交易监控。

（3）机器支付与数字货币层：支持智能体之间的自动化经济活动

智能体经济并不局限于购买实物商品。未来大量交易可能发生在智能体与 API、数据平台、计算服务及其他智能体之间。这些交易可能金额极小、频率极高，传统以人为中心的结账流程难以适应。

2026 年 3 月，Stripe 与 Tempo 共同推出机器支付协议（Machine Payments Protocol，MPP），允许智能体以程序化方式购买数字资源，支持微支付、持续计费及自动化交易。MPP 能够通过稳定币完成支付，也可以借助 Stripe 现有支付系统处理法币支付。

2026 年 5 月，Stripe 旗下 Privy 与 AWS 合作，为 Amazon Bedrock AgentCore 提供智能体钱包基础设施；6 月，Stripe 又宣布支持 AWS 的 AI 流量商业化方案，使内容提供商能够向付费访问的智能体提供文章、数据和 API 等资源。

为了支持机器之间的资金流动，Stripe 构建了数字货币与钱包基础设施。2025 年完成的 Bridge 收购，为其提供稳定币收付、兑换及跨境资金流转能力；同年收购的 Privy 提供可编程数字钱包；Stripe 与 Paradigm 共同孵化的 Tempo 则提供面向支付优化的区块链基础设施。

这些业务并非全部因智能体经济而建立，但组合之后，可以为智能体提供数字资产账户、可编程钱包、支付通道和跨境资金流动能力。其产业价值是推动智

能体从“能够付款”走向“能够持续进行机器间经济交易”。对于 KYA 而言，这也意味着身份、授权和风险控制必须能够适应高频、自动化、跨平台的交易环境。

（4）计量、计费与资源配置层：构建智能体经济的商业运行机制

智能体不仅消费商品，也大量消耗模型 Token、计算资源、数据库和各类数字服务。如何精确计算成本、合理定价和及时收费，是智能体经济能否持续运行的重要问题。

2026 年 1 月，Stripe 完成对用量计费企业 Metronome 的收购。Metronome 为 OpenAI、Anthropic 和 NVIDIA 等企业提供复杂用量计费能力。Patrick Collison 认为，基于用量的商业模式将是未来十年的重要发展方向，而计量与计费正是连接产品与商业的接口。

2026 年 4 月，Stripe 宣布结合 Metronome 与 Tempo 开展流式支付（Streaming Payments），探索在 Token 等数字资源被消耗时进行实时计量和微支付。这种机制有助于降低智能体高频使用资源所产生的成本与资金回收之间的错配。

2026 年 8 月，Stripe 进一步宣布达成收购 OpenRouter 的协议。OpenRouter 提供跨越 80 多家供应商、400 多个模型的 AI 模型路由服务，帮助企业根据成本、速度、性能和可靠性选择模型。截至本报告写作时，应将其表述为已宣布收购协议，而非据此认定交易已经完成。

这一布局使 Stripe 开始进入智能体经济更深层的资源配置环节：不仅处理智能体创造收入之后的支付，而且参与模型资源的调用、消耗计量、成本优化与收入管理。

（5）智能体运营层：使智能体能够自主采购和管理数字服务

Stripe 还在探索如何让智能体直接完成企业经营中的部分工作。

2026 年 4 月，Stripe Projects 面向所有用户开放，使开发者及其智能体能够通过命令行注册、采购并集成托管、数据库、认证、AI 及可观测性等服务；截至 2026 年 10 月，Stripe 官方文档显示，Projects 已支持 60 多家服务提供商。同年 6 月，Stripe 推出按服务商设置支出上限、为开发 / 测试 / 生产环境配置隔离凭证等控制机制；平台还可通过委托授权代开发者配置服务，并签发受限范围的凭证。

与此同时，Stripe MCP 允许 AI 应用与 Stripe 服务连接，Stripe Workflows 和 Stripe Console 等工具则进一步支持企业通过自然语言进行数据分析和业务操作。

这使 Stripe 不仅服务于“消费者委托智能体购物”，也开始支持“企业委托智能体经营”。

4.2.2 Stripe 在 KYA 方面的实践

Stripe 尚未公开形成一套覆盖智能体身份、行为、信用评分与跨平台信誉的完整 KYA 体系，但其已经建立的多项机制，实际上涉及 KYA 的关键基础能力。

（1）从身份识别转向委托关系识别

传统 KYC 主要确认客户是谁，而智能体交易还需要确认：智能体代表谁行动、获得了什么授权、其行为由谁承担责任。

Stripe 在智能体支付中，将商户身份确认、消费者身份与支付授权相结合；在企业 API 接入中，则允许使用标记为智能体专用的受限 API 密钥（Agent-tagged Restricted API Keys）。

这类密钥能够限定智能体访问的资源和操作范围，并与审批规则关联。它使智能体的操作权限可以区别于普通应用程序凭证管理。

不过，API 密钥认证并不等同于跨平台、具有独立身份凭证的智能体认证。Stripe 现阶段更多是在已知个人或企业账户之下识别和管理智能体的受托权限。

（2）从静态授权转向可编程的动态控制

Stripe 在 KYA 方面最具代表性的实践，是把授权转变为可以执行、修改和撤销的程序规则。

SPT 可以约束收款商户、支付金额和有效期；Issuing for Agents 可以设置支出限额和商户类别；Projects 可以限制每项服务的预算，并使智能体默认在开发环境中操作，避免未经授权影响生产系统。

2026 年，Stripe 还推出或预览智能体护栏（Agent Guardrails）机制，包括智能体身份标记、权限规则及敏感操作审批。其官方开发文档进一步列明，对退款、对外转账、付款及取消订阅等操作，可以根据金额、操作主体或请求频率要求人工批准或直接阻止。

这些机制形成了从权限授予到执行过程约束的控制链条。其核心思路不是预先假定智能体一定可信，而是在授权边界内允许其行动，并限制错误行为可能造成的损失。

（3）从交易反欺诈延伸到智能体行为风险管理

Stripe 原有的 Radar 风险控制体系，为其进入 KYA 提供了另一项基础能力。在智能体交易中，Radar 可利用支付方式、拒付概率、盗刷风险、异常交易等信号，辅助识别正常授权的智能体与恶意自动化程序。

2026 年 4 月，Stripe 宣布进一步将 Radar 用于防范 AI 服务中的 Token 盗用和虚假注册。根据其披露，在 Stripe 服务的 AI 企业中，约六分之一的注册尝试来自恶意行为者；针对八家高增长 AI 企业，Radar 在此前一个月阻止了超过 330 万次高风险注册。

需要注意，上述数据反映的是 AI 业务相关欺诈风险，并不能直接作为智能体信用体系的覆盖规模或识别准确率。在 KYA 框架下，这类能力的价值在于形成动态行为监测：信任不再仅由账户开立时的一次认证决定，而需要结合后续使用、交易和异常行为持续评估。

（4）从事前防护延伸到交易记录与事后保障

可信智能体不仅需要能够安全执行交易，还需要具备可观察、可追溯和可处理争议的机制。Stripe 的支付令牌、虚拟卡和交易记录可以关联授权与实际付款；敏感操作审批和安全日志则支持追踪部分智能体行为。

2026 年 9 月，Stripe 进一步为 Link 智能体钱包提供购买保障。对于符合条件的交易，用户可以获得涉及意外损坏、遗失物品、购买后降价及退货等情形的保障。这并不意味着 Stripe 承诺赔偿所有智能体错误，但说明其开始将事后消费者保护纳入智能体商业设计。

这一实践值得关注：智能体信任不仅来自“防止犯错”，也来自“出现错误之后，是否具有可预期的救济机制”。

4.2.3 KYA 发展潜力、局限与未来方向

Stripe 已经在智能体身份标记、权限授予、支付控制、交易监测和审计保障等方面形成了较为系统的能力，但距离完整的智能体信用体系仍有明显距离。

首先，现有信任机制主要依托消费者、企业、账户和支付凭证。Stripe 能够较好地判断某项操作是否获得授权、是否符合支付规则，却尚未公开建立覆盖不同平台的统一智能体身份、长期履约信誉和独立信用评分体系。识别一个智能体是否“有权执行交易”，与判断它在长期协作中是否“值得信任”，仍是两个不同层次的问题。

其次，Stripe 积累的数据主要集中在支付、计费、消费和金融操作环节。智能体在信息检索、业务决策、复杂任务执行和多智能体协作中的可靠性，不能单纯通过支付行为进行评价。提示词注入、工具污染、虚假信息、任务偏离等风险，也无法依靠交易风控完全解决。

因此，从 KYA 的发展方向看，Stripe 最值得关注的潜力主要体现在三个方面。

第一，发展以委托关系为核心的智能体身份与授权基础设施。将智能体标识、委托主体、权限范围、支付凭证和操作记录建立更紧密的联系，使不同平台能够验证某个智能体的授权来源及其权限边界。这可能成为现有 KYC、KYB 和支付认证体系向 KYA 延伸的重要路径。

第二，从交易风险评估走向智能体持续信用评价。Stripe 拥有庞大的商户网络与金融交易基础设施。如果未来能够在隐私保护和合法授权条件下，将智能体的交易合规性、履约结果、异常记录、争议情况及持续行为纳入统一评价机制，就可能从目前的单笔交易风控，进一步发展出长期、动态的智能体信誉和信用能力。但这种跨平台信用体系目前仍属于发展潜力，而非既成产品。

第三，推动智能体信任机制的跨平台互操作。 智能体经济不可能由单一企业独立完成。AI 平台、商户、银行、支付网络及智能体开发者必须共同建立能够互认的身份、授权、交易和责任规则。Stripe 同时与 OpenAI、Google、Anthropic 及主流支付网络合作，并支持多种商业与支付协议，使其具有参与这一开放生态建设的条件。

Stripe 在 2025 年度报告中将智能体商业描述为进入“建设与真实场景实验”的阶段，并特别强调，未来成功的关键在于实现普遍互操作性（Universal Interoperability）。这一判断同样适用于 KYA：只有信任机制能够跨越单一平台和单笔交易，智能体信用才可能成为真正的经济基础设施。

KYA 可以沿着“身份识别—授权管理—行为约束—风险评价—信用积累”的路径发展。Stripe 目前走在这一链条的前半段，并已具备向后半段延伸的重要基础。

4.3 芝麻信用：KYA 探索与智能体信任基建

应用场景	智能体信用赋能
核心价值	面向智能体的持续可信身份体系。它不仅回答“这个智能体是谁”，还需要回答“它属于谁、被允许做什么、当前是否可信、行为能否追溯、出现问题由谁负责”，包括身份属性、可信评估、跨域互认和持续监控四类能力。
智能体成熟度	L2-L3

自 2015 年创立以来，芝麻信用长期围绕用户信用评估与商业信任服务展开，致力于降低人与人、人与组织之间的交易信任成本，持续拓展商业信任服务的应用范围。

依托庞大的支付与电商生态，芝麻信用不仅构建了中国最早被广泛应用的个人信用评分，更率先在共享单车、免押住宿、设备租赁等场景掀起“信用免押”浪潮，随后通过“芝麻先享”等模式将信任服务从 C 端消费者拓展至 B 端中小微企业。

随着智能体的发展，智能体开始替代人类执行选品、交易甚至合同签署等实质性商业行为。面对这一商业应用的新阶段，芝麻信用开始布局 KYA，旨在构建 AI 时代商业信任基础设施。

市场对 KYA 的需求，来自智能体交易中出现的三种摩擦：

- 身份摩擦：交易对手难以确认一个智能体背后的主体是谁；
- 权限摩擦：交易对手难以确认智能体当前行为是否获得了充分授权；
- 风险摩擦：即使身份和授权都明确，交易对手仍缺乏评估智能体历史履约表现及潜在技术风险的依据。

芝麻信用过去的商业模式恰恰是在解决类似问题。例如，一个用户拥有较好的信用记录，租车平台可以降低押金要求；一个消费者拥有较好的履约能力，商家可以允许其先享后付。

信用的价值正是降低交易中的不确定性，促成原本因信任不足而受阻的交易。同样，KYA 的价值也体现在：让用户更愿意委托智能体，让交易对手更愿意承接其发起的交易。这也是芝麻信用判断 KYA 市场价值的核心商业逻辑。

4.3.1 市场洞察：智能体时代的“信任黑盒”

结合 AI 商业化进程，芝麻信用关注到智能体落地面临的关键障碍。

首先是主体权责模糊带来的潜在纠纷风险。不同平台的身份标识和管理方式存在差异，跨平台识别与责任主体关联面临挑战。如果缺少稳定的身份标识和清晰的主体关联，发生交易纠纷时就难以识别相关责任主体。

其次是授权边界不清带来的越权操作风险，若缺乏细粒度的约束机制，极易引发超额支付或滥用 Token 等风险。用户如果不能精确掌控“智能体能做什么”，就难以放心将资金处置权交给 AI。

同时，缺乏防护的智能体可能因为大模型自身的“幻觉”或遭受外部恶意指令攻击而失控，从而导致重大的资金损失。

基于这些洞察，芝麻信用判断：完善的 KYA 信任机制，有助于降低智能体委托与交易中的不确定性，支撑智能体从信息交互走向商业交易。

4.3.2 商业实践：智能体信用落地的多元探索

在确立了向 KYA 演进的战略后，芝麻信用围绕不同商业场景，分阶段推进能力落地与价值验证。

在已落地的商业实践中，千问电商购物已打通底层先用后付能力，让用户在智能体内购物时可以先使用、后付款，改善了购物体验，用户购物转化意愿明显提升。

在价值验证与场景拓展方面，TokenPay 探索“先授权、按实际使用结算”：用户在了解服务及扣款规则并明确授权后，可根据信用情况获得特定服务的一定额度或时间窗口，实际使用后再产生订单和扣款。

公共事业缴费探索“设好规则、委托代办”：用户设置缴费规则并授权后，由智能体获取账单、按规则执行缴费并反馈结果，支持暂停、撤销、异常介入和全过程留痕。

两个项目目前均处于价值验证阶段；同时，芝麻信用也在无人零售、3C 租赁等场景中探索信用能力与智能体服务的结合。

4.3.3 核心机制：授权与决策边界

明确用户或系统授予智能体的权限，以及智能体可自主决策的范围，是实践中的核心问题。KYA 体系需要建立“人-机”双向确权与身份绑定机制，要求明确智能体背后的归属主体是“人”或“企业”。这种身份绑定有助于在发生损失或纠纷时，为责任主体识别与问题追溯提供依据。

在授权机制上，芝麻信用实施细粒度、多维度的动态授权，放弃“一刀切”的放权，转而设定具体的金额上限、场景范围。例如，限定智能体仅能在指定平台（如支付宝域内）点咖啡，以此实现物理隔离和风险控制。

为了平衡效率与安全，芝麻信用根据任务频率、风险及复杂度，将智能体的自主决策权科学地划分为三个智能化层级：

决策等级	核心特征	风险类型与适用场景
人工确认	智能体仅提供建议与信息预填，最终执行必须由自然人介入确认。	适用于高风险、首次发生或涉及真实身份核验的重大场景（如大额转账、法律合同签署）。
信用免密	在预设的授权额度和白名单场景内自动执行，超出限制则需要人工确认。	适用于中低风险的常规商业交易（如日常咖啡点单、视频会员购买）。
全自动决策	智能体拥有全链路自主决策、执行并反馈的权限，无需人类实时参与。	适用于极低风险、高频周期性且逻辑固定的事务（如自动缴纳水电费、周期性订阅续费）。

4.3.4 攻克难题：冷启动、幻觉与数据孤岛

为了支撑复杂的商业试点与未来生态，芝麻信用构建了坚实的技术基础设施。其底层复用芝麻信用现有的 KYC 身份基建，在授权前提下关联责任主体信用，并逐步积累 Agent 自身的行为评价依据；中间层则创新构建了 ACP 协议（Agent Credit Protocol），专门用于处理人与智能体的身份绑定、权限授权及信用数据回流；上层全面开放，对接各类具体的 AI 应用场景。

而在推进过程中，芝麻信用还需要攻克多项行业难题。针对早期智能体行为数据不足的问题，芝麻信用探索在授权前提下关联责任主体信用，并通过真实场景逐步积累智能体的履约、异常和行为记录，为其自身信誉评价提供依据，同时持续保留与责任主体的关联。

面对大模型幻觉与交易延迟带来的风险控制难点，芝麻信用采取了严谨的安全策略。在核心的交易与资金流转链路中，芝麻信用不直接使用大模型进行信用决策，以降低幻觉引发误操作的风险。这有助于提高交易执行的确定性与安全性。

同时，面对跨平台交互的数据孤岛问题，芝麻信用探索智能体信用数据上链的可行性，以推动建立行业通用的基建标准。

4.3.5 前瞻规划：智能体时代的商业基建

面向智能体商业生态的发展，芝麻信用持续探索 KYA 的建设路径。据芝麻信用介绍，团队正在研究面向智能体的持续可信身份体系。它不仅回答“这个智能体是谁”，还需要回答“它属于谁、被允许做什么、当前是否可信、行为能否追溯、出现问题由谁负责”，包括身份属性、可信评估、跨域互认和持续监控四类能力。

在协议标准层面，芝麻信用围绕 ACT 协议的信任服务域推进建设。信任服务域是 ACT 协议的信任基础设施层，其中智能体信用关联（TSD-CRD）已在外滩大会正式发布。TSD-CRD 主要解决 Agent 在获得授权的前提下，如何受限地引用和验证关联责任主体的信用信息。

在演进路径上，芝麻信用将基于具体场景的价值验证，逐步积累智能体交互与履约记录，并探索向更多平台与服务供给方拓展。

未来，智能体信用评价范式将迎来更深刻的变革。传统信用评价侧重偿付与履约风险，面向智能体的 KYA 则探索结合身份、授权、行为与履约等信息，形成多维评价。未来的 KYA 很可能不会是一个单一分数，而是一套动态信用凭证。智能体信用档案可包括身份、能力、授权、行为、履约、稳定性、责任等要素。

KYA 的核心不仅仅是“给智能体打分”，而是建立一个能够帮助交易对手判断是否接受委托、承接交易的信任支撑体系。

芝麻信用表示，将依托既有信用品牌优势，持续完善智能体信任服务，为智能体商业协作与交易提供基础设施支撑。

4.4 冰鉴科技：构建金融级可信智能体

应用场景	金融机构信用业务
核心价值	确保数据在本地闭环内完成 token 化与模型推理。构建可私有化部署的基础模型，是实现“数据不出域”的可行技术路径，也是切入银行、保险等核心系统的准入门槛。
智能体成熟度	L1-L2

冰鉴科技 2015 年成立于上海，是一家以人工智能服务企业级客户、尤其深耕金融信用与风控场景的科技公司。公司从机器学习/深度学习驱动的个人信贷风控起步，逐步扩展到保险、电商、教育、政务等领域；海外业务覆盖香港、新加坡、东南亚、北美、拉美，客户既包括中资出海机构，也包括部分海外本土金融机构。

冰鉴科技是中国信用科技体系中“独立第三方、合规优先、模型驱动、金融私有化、正向 KYA/RegTech 延伸”的垂类样本；它卡入“通用大模型进不了银行、银行自己又做不好垂直智能体”的生态空间，成为金融机构与自主智能体之间的信用中介和安全阀。

在信用科技产业链中的四层生态（数据源/征信层、模型与风控技术层、金融机构应用层、监管与基础设施层）中，冰鉴处在“第二层偏底座”：它既不是持牌征信机构，也不是金融机构内部科技条线，而是站在“持牌机构与外部大模型/数据之间”的可信模型层。

4.4.1 自研垂类大模型“本立道生”

2023 年 5 月，冰鉴科技正式推出“本立道生”大模型。2023 年 8 月：正式向东南亚地区的客户推出该模型，并应用于智能客服等业务中。2024 年 6 月 13 日：“本立道生”被列入上海市新增的通过备案的生成式人工智能服务名单，完成了中国内地的服务备案。

作为一家信用科技公司，自研大模型，是为了应对数据合规约束、场景垂直壁垒、商业安全边界、智能体可控性要求。

（1）数据主权与监管合规的刚性约束

金融行业受强监管属性决定了其核心数据（如客户流水、信贷记录、保单信息）严禁出域或外传至第三方公有云。通用大模型依赖公有化 API 调用，存在严重的合规风险。冰鉴科技明确自身“不生产、不买卖数据”的定位，其赋能金融机构的前提是确保数据在本地闭环内完成“token 化”与模型推理。构建可私有化部署的基础模型，是实现“数据不出域”的可行技术路径，也是切入银行、保险等核心系统的准入门槛。

（2）垂直场景壁垒构建的需要

金融大模型的核心竞争力并非通用语言能力，而在于对风控规则、监管口径、非结构化账单解析、反欺诈逻辑等垂直知识的深度编码。冰鉴科技自 2015 年起积累的特征工程、多源数据变量加工、联邦学习及合规评分体系，构成了其独有的“基础资产”。通过自研基模型，企业可将上述领域知识固化为模型权重与架构设计，形成“通用框架+金融微调”的差异化壁垒，避免被通用大模型的能力泛化所稀释。

（3）商业安全与产业链话语权的考量

若仅作为应用层开发商依附于外部大模型，将面临底层技术受制于人、服务稳定性受限于他方迭代节奏的风险。自研基模型使冰鉴科技从“套壳服务商”升级为“底座供应商”，在商业逻辑上实现了“卖铲子给有矿的人”。这一转变不仅提升了毛利率与客户的转换成本，更使其在大体量银行按业务线拆分采用多基模型的趋势中，具备了与原厂开源模型同台竞争的能力。

4.4.2 金融智能体三标准

在金融信用科技领域，智能体技术的引入并非简单地叠加对话界面或自动化脚本，而是涉及决策主体从“人”向“人—机协同”的范式迁移。

冰鉴科技需要在强监管、高敏感性的金融场景中，构建一个既具备自主规划与动态调优能力，又完全可控、可审计、可阻断的“金融级可信智能体”体系；以微型场景为切口，以人在环路（Human-in-the-Loop）为安全边界，逐步将智能体从边缘辅助工具推向核心业务环节。

冰鉴科技创始人顾凌云提出，真正合格的智能体必须同时满足三项核心标准：

其一，具备一整套技能集，能够在特定场景中精准调用行业知识、解析工具与规则引擎，而非仅依赖单一的语言生成能力；

其二，具备自主设计完整流程的能力，能够像项目经理一样自洽地规划步骤顺序，并准确匹配所需的技能模块；

其三，也是最具金融适配性的要求，即具备基于外部反馈动态调优既定流程的机制，这种动态调整能力本质上源于强化学习的“动态试错”逻辑，确保智能体在面对非结构化、多变的输入时，能够自我修正而非机械执行。

这三项标准共同构成了冰鉴智能体设计的理论基座。尤为关键的是，缺乏动态反馈机制的自动化流程在金融场景中根本无法被认定为智能体，因为金融业务的容错率极低，任何静态的、不可纠偏的自动化链条都可能引发灾难性的错误链式传递。

基于对金融行业技术成熟度的审慎判断，冰鉴科技并未追逐覆盖全场景的“大智能体”或“端到端智能体”，而是采取了更为务实的“微型智能体”先行策略。

4.4.3 用例：银行账单解析

顾凌云认为，当前，金融类智能体的成熟度一般不超过 L2 级，即智能体的信用锚定在人的信用，关键决策节点仍需人工介入确认。这一现状决定了智能体在金融核心系统中的角色必须是辅助性的、分段式的。在应用上，冰鉴科技选择了“宁慢勿乱”的策略。

冰鉴科技将研发与落地资源集中于账单读取、非结构化文档解析、风控变量加工、数据查询与报告生成等高频、规则相对明确、但人工操作繁琐的微型场景。这种落地形态，既避免了长链路任务中因单环节成功率叠加导致的整体容错率崩塌，又通过快速见效的局部优化，为金融机构提供了可感知的价值，从而为后续更复杂的智能体应用积累了信任基础。

冰鉴科技已在多个具体场景中部署了智能体应用，形成了兼具实用性与创新性的用例矩阵。

在金融核心业务端，最具代表性的用例是银行账单解析智能体。该智能体需处理来自企业客户的 PDF 图片、文本文件及 JSON 数据集等复杂多变的格式。智能体根据输入格式动态调用不同的解析技能，自主规划从数据提取、字段映射到异常检测的步骤顺序，并基于反馈对解析流程进行调优。例如，当面对扫描质量不佳的 PDF 时，智能体能够自动切换图像增强与文本识别模型，或调整字段对齐策略。

优先解决这个环节的应用，是因为，若底层数据提取出错，后续的风险分析与授信决策将全盘皆输。因此，该智能体并非独立运作，而是深度嵌入 KYA 评估体系，其证据层需厘清数据来源、严防 AI 幻觉，并配合人工节点对关键字段进行确认，从而构建起从数据到决策的完整风控闭环。

在办公与数据分析场景，虚拟办公智能体（MAS 架构）实现了从自然语言问数到自动生成报告的端到端辅助。该智能体能够理解用户的自然语言查询，将其转化为 SQL 或代码，查询生产数据库，自动进行建模分析，并最终输出结构化报告。这一用例联同客户内部平台与云端建模环境，显著提升了数据科学家与业务分析师的工作效率，其核心创新在于智能体对分析流程的自主编排，而非简单执行预设脚本。

在面向东南亚市场的电商、保险与物流客户中，冰鉴部署了智能客服与销售辅助智能体。这些智能体不仅能够回答标准化流程问题，还能识别客户情绪与口语化意图，生成定制化报告与销售建议。与呼出型智能体（主动外呼或查询）相比，呼入型智能体面临意图不可控、兜底机制复杂等挑战，尤其在保险等强监管行业，容错率极低。冰鉴的解决方案是在智能体中设置敏感的沙箱边界与智能节点，当客户行为超出预设范围或触发监管风险时，系统自动转人工处理，确保业务合规性。

4.4.4 KYA：安全与信任机制的创新

在强监管环境下，金融机构对智能体的信任度严重不足，因此，需要构建 KYA 评估体系，并将其嵌入智能体的全生命周期。

顾凌云认为，KYA 并非对传统信用评估体系的颠覆，而是将其逻辑从“人/企业”向“智能体”的延伸；构建过程需遵循“主体化定位、证据层先行、授权分层、渐进式标准化、市场驱动监管”的路径：

（1）确立智能体的“信用主体”地位，实现评估范式转换

KYA 构建的首要前提是承认智能体具备独立的“信用”属性。传统信用体系评估对象为人、企业或模型，而智能体作为新兴决策主体，其输入纯净度、流程合理性、授权充分性及输出可归责性均需被纳入信用评估范畴。构建 KYA 需从主体化思维出发，将智能体的行为链视为可评估、可定价的信用单元。

（2）以证据层为底座，建立可验证的数据溯源机制

针对智能体落地中常见的“底层数据失真导致上层决策全盘皆输”的痛点，KYA 必须优先构建证据层。该层需回答四个核心问题：数据来源是否清晰、是否经过加工、是否存在 AI 幻觉或字段错位、提取链路是否可回放与复核。在冰鉴科技的账单解析等场景中，证据层是确保风控结论可靠性的基石。KYA 的底座并非直接输出评分，而是建立一条从原始数据到决策依据的完整证据链，实现“底层可验、中层可控、上层可溯”。

（3）覆盖安全授权层，定义智能体的权限边界与防篡改机制

金融场景对数据泄露与越权操作零容忍，因此 KYA 必须将安全授权层作为核心模块。构建过程需明确界定智能体的操作权限：包括但不限于能否读取生产数据库、能否发起对外 API 调用、能否修改客户状态、能否进入授信或核保决策环节，以及哪些动作必须设置人工确认节点。这一授权体系相当于为智能体颁发“数字牌照”，通过沙箱隔离、敏感操作拦截、防篡改日志等技术手段，将权限边界内化为系统原生能力，而非事后补救的合规补丁。

（4）采用渐进式标准化路线：从二元判定到分级，暂缓统一评分

未来 3 至 5 年内难以形成类似 FICO 的通用智能体信用分。因此，KYA 的构建应遵循渐进式路线：初期以“通过/不通过”的二元判定为主，快速满足金融机构的采购与风控需求；中期引入风险等级、场景适配等级、可控等级等分层评估；远期方可探索统一量化评分。该路线有利于降低试错成本，避免标准缺位导致的市场停滞，同时防止厂商利用黑箱评分谋取不当利益。

（5）监管节奏：市场试错先行，避免过早高度监管

监管层应避免过早实施高度监管。智能体技术尚处探索期，过早的刚性约束会扼杀自由竞争与创新尝试。理想路径是：先允许市场充分试错，暴露典型风险场景；由监管机构联合头部金融机构与技术厂商，基于实践沉淀形成标准。

（6）嵌入“人在环路”机制，确保长链路任务的可控性

鉴于当前金融智能体成熟度仅达 L2 级（锚定人的信用），长链路任务中误差累积风险显著，KYA 必须将“人在环路”（Human-in-the-Loop）作为安全底线。构建时需预设人工确认节点：在关键决策前强制复核、异常置信度下自动回退、错误链式传递可被阻断。KYA 的目标并非证明智能体可完全替代人类，而是界定其在何种场景、何种边界内可安全辅助人类决策，实现“人机协同”的信任增强。

4.5 Ryt Bank：全球首家 AI 原生银行的智能体信任实践

应用场景	数字银行
核心价值	全球第一家纯线上、纯人工智能的数字银行——开户、转账、缴费与风控均由 AI 为核心构建。智能助手上线首年被调用超过 1,000 万次。
智能体成熟度	L2-L3

Ryt Bank 由马来西亚杨忠礼集团（YTL Group）与冬海集团（Sea Limited，电商平台 Shopee 母公司）合资成立，2025 年 8 月正式上线，是马来西亚第五家持牌数字银行，也是全球第一家纯线上、纯人工智能的数字银行——开户、转账、缴费与风控均由 AI 为核心构建。其底层人工智能由 YTL AI Labs 基于马来西亚主权大模型 ILMU 打造，通过内置智能助手 Ryt AI 直接服务客户。

截至 2026 年 8 月上线一周年，Ryt Bank 客户数已突破 150 万，成为马来西亚用户规模最大的数字银行；智能助手 Ryt AI 由守卫（Guardrails）、意图识别（Intent）、支付（Payment）、问答（FAQ）四个大模型智能体协同编排，上线首年被调用超过 1,000 万次，近半数客户使用过 Ryt AI。

这个案例呈现了一个“巨型集团 + 高复杂度行业”的切片。杨忠礼集团同时拥有能源、矿业、房地产等强监管、高复杂度业务的 AI 改造诉求，又运营着数字银行这样的 To C 金融业务，因此其对智能体安全与信任的需求是内生的。与此同时，集团把算力、模型、银行与智能体平台整条链路握在自己手里，这一垂直整合构成了它处理智能体信任问题的独特筹码。

4.5.1 一体化 AI 底座

受访者用“五层”来描述杨忠礼集团的 AI 存在：第一层是电力与水务，第二层是自建的数据中心（规划最高 2GW），第三层是自有云与大模型，第四层是 Ryt Bank，第五层是面向马来西亚企业、政府部门与个人的智能体平台（受访者称之为 LIMU，含 LIMU Chat 和 ILMU Claw）。在柔佛，集团还与英伟达合作建设东南亚最大的太阳能供电数据中心，把能源与算力基础设施连成一体。

这种“能源—算力—云—模型—银行—智能体”的纵向一体化，在访谈中反复被提及为信任的基石。因为推理平台、模型与基础设施都是自有的，集团可以直接在推理层配置正确的隔离策略，避免多租户场景下的数据串扰；也可以把数据主权与加密边界内化进产品，而不必依赖第三方公有云的口头承诺。对 KYA 而言，这提供了一个从“基础设施层”而非“应用层”构建信任的视角。

4.5.2 ILMU Claw 智能体平台与多智能体风控

在智能体平台层面，ILMU 目前承载的主要用途包括：大型文档的理解、摘要与生成，面向消费者的客服与聊天机器人，代码编写，以及面向企业的保险欺诈识别、理赔矛盾识别、零部件预测、交易辅助等。平台还支撑了一个面向政府监管机构的多智能体反诈骗系统：多个智能体进入 Facebook、Telegram、WhatsApp 等群组，识别“投资我们、每月回报 10%”之类的诈骗话术；由于如今的诈骗已从纯文本演变为视频、图片与文字混合形态，系统必须依靠多模型、多智能体对网络内容进行持续监测。

在 Ryt Bank 内部，受访者强调风险控制首先体现在架构层面。每个智能体遵循“计划—执行—观察（Plan - Act - Observe）”的循环；输入侧设置一个小型守卫（Guardrail）智能体，先判断用户请求是否与金融业务相关——转账可以进入执行，而“如何制造炸弹”这类请求会被直接拒绝；输出侧再设置另一个守卫智能体，拦截可能输出错误、拼凑、误导或危险信息的结果。由于大模型的输出本质上是概率性的，这种“输入 + 输出”双守卫设计，是把概率性输出重新纳入确定性约束的第一道防线。

在此基础上，银行还配置了国际化的红队团队，持续对智能体系统进行对抗性测试，发现问题即修复；同时把“人在环路（Human-in-the-Loop）”作为默认设计，而不是事后补丁。

4.5.3 数字身份、反洗钱与人机协同

作为马来西亚五家数字银行之一，Ryt Bank 认为把 eKYC 做好至关重要。马来西亚本地长期存在“骡子账户（mule account）”问题——大量新开账户被留作后续的欺诈用途，因此银行希望尽可能早地把这类账户拦截在开户阶段。受访者解释，之所以对 eKYC 合作方蚂蚁数科放心，是因为其同时服务马来西亚最大的电子钱包之一 Touch 'n Go，拥有关于“什么有效、什么无效”的大规模实践数据库。

在反洗钱（AML）环节，智能体承担了比传统规则引擎更细粒度的工作：它会自动检索一个人的公开网络信息、找出其姓名的各种变体，并与马来西亚金融体系中的 AML 名单进行比对。之所以需要多智能体，是因为这类判断很少是简单的“是/否”——一个用户可能看起来可疑，但在补充收入流水、说明收入来源之后又变得可以准入。这种需要多轮、细微权衡的判定，恰恰是多智能体系统比单一模型更擅长的地方。

在身份与授权上，Ryt Bank 给每个智能体分配了加密的安全凭证：智能体在与核心银行系统通信时，必须携带这套加密代码，系统据此确认对方身份。更重要的是，这种身份校验并不只停留在凭证层面，而是把“使用哪个基础模型、智能体在哪里托管与起源、它在到达这一步之前经历了哪些调用”都纳入判断，从而把“这个智能体是谁”还原为一条可追溯的调用链。这实质上是在金融语境下，把 KYC 的“识别主体”逻辑向 KYA 的“识别智能体及其权限来源”延伸。

当被问及 KYC 与 KYA 的区别时，受访者给出的原则是：智能体在执行用户的意图，因此当它需要授权时，会向用户请求授权，而不是拥有独立的行动人格。真正需要回答的问题是“系统正在对话的这个智能体，是不是那个被允许对话的智能体”。

由于自然语言与图像识别本质上是概率性的，Ryt Bank 的核心做法是让智能体的输出携带置信度分数：置信度高时直接放行，置信度低时回问用户，并把用户反复确认的情况反馈到下一轮训练中，形成学习闭环。受访者特别强调，这不是简单的“两步确认”，而是“持续的人在环路”——当智能体需要授权、不确定指令细节，或用户提供的信息不足时，它都会回到用户处请求确认或补充信息。

访谈中受访者分享了一个具体教训：早期版本的智能体在处理图像时，会尝试猜测模糊、反光或过暗图片中的数值，结果产生幻觉。团队后来明确要求智能体“看不清就说看不清”，而不是猜测，并把这类情况交回给用户。受访者由此点出一个更普遍的判断——大量幻觉发生在 AI 对自己本不该确定的事情持续保持自信的时候，因此系统的关键能力之一是让 AI 学会“承认不确定”。

在置信度的具体构造上，受访者举例：对同一张图片并行读取五次，若五次结果一致则置信度高，若出现“两次读成 A、三次读成 B”的分布则置信度低；同时结合图像质量判断，质量差时主动请用户重新上传。受访者也坦承，完整的置信度模型细节需要向工程团队进一步确认。

4.5.4 数据主权与隐私：自建基础设施的信任红利

在隐私与数据安全方面，Ryt Bank 遵循马来西亚央行的数据安全监管要求，并让智能体访问私人数据时沿用与普通软件相同的严格协议；所有数据在传输中与静态存储中都加密，一个智能体无法访问没有对应密钥的他方数据。

针对“大模型是否会泄漏私人数据”的问题，受访者把根因归结为推理层的 KV Cache 共享等机制导致的数据污染，而自建推理平台正是应对之道——集团可以正确配置推理环境，从源头避免这类数据串扰。即便有污染风险，多智能体

系统的最终输出仍要经过输出守卫的检查：一旦发现输出疑似数据污染，就丢弃结果并重新执行。

4.5.5 成熟度、监管与跨机构互信

受访者自评 Ryt Bank 的智能体当前处于“五级中的第三级”（对应本报告 L0 - L5 模型中的 L3 混合信用），并认为未来可以更自主，但需要先把基础设施补齐。他举例说明理想的自主形态是“如果市场在上涨，帮我买入 200 令吉的微软股票”，但目前尚未到达。并判断，Ryt Bank 有望在未来约 12 个月内实现“合理程度的自主”，而整个金融行业因为天然的保守性，大概需要三到四年。

在监管层面，受访者表示马来西亚央行目前还没有针对 AI 的专门规定，银行对智能体的要求主要是从软件系统与云服务的既有监管要求中继承，尚不需要就大模型或智能体向监管单独注册。

关于智能体之间的互信（Agent-to-Agent），受访者坦言目前 Ryt Bank 的智能体仍限于单一机构内部，尚未探索跨机构智能体协作，但认为这是一个“非常好、且想探索”的问题，并表达了与研究者共建行业标准的意愿。

4.6 英和智临：校园智能体实践与本地化信任底座

应用场景	教育管理
核心价值	在“数据不能出校”的刚性约束下，把身份、权限、审计与质量治理等信任要素直接封装进软硬件一体的产品架构，让智能体在校园这一高敏感场景中可信地运行。
智能体成熟度	L1-L2

在智能体产业的生态版图中，上海英和智临科技有限公司（下称“英和智临”）占据着一个与平台型科技公司截然不同的生态位。区别于只出售裸金属算力的传统供应链模式和纯软件厂商，英和智临背靠“英和（SEAVIV）”品牌的供应链体系，其 To B、To G 产品线聚焦“硬件＋算力平台＋软件应用”三位一体的本地化方案：底层硬件承载本地私有化算力，中层是 MaaS（模型即服务）算力平台，上层面向教育等垂直场景构建应用。

公司成立仅一年余，但落地速度可观：已与上海四所标杆学校合作形成课堂最佳实践，2026 年 9 月至 12 月还将在徐汇区为约 10 所学校布点，并获得徐汇区教育局的明确支持。

英和智临展示的是智能体信任的另一种解法：在“数据不能出校”的刚性约束下，把身份、权限、审计与质量治理等信任要素直接封装进软硬件一体的产品架构，让智能体在校园这一高敏感场景中可信地运行。对 KYA 研究而言，这一案例回答了一个互补性问题：当智能体尚不构成独立的交易主体时，“可信”如何被工程化地制造出来。

4.6.1 市场洞察：政策牵引、教师减负与隐私红线

英和智临选择教育作为重点场景，源于三重需求的共振。

政策端，上海教委近年大力推进 AI 教学与“千人千面”的个性化指导，学校拿到试点文件后普遍面临“知道要做、不知怎么做”的落地真空，亟需一套现成的平台与服务把试点快速跑起来。

需求端，一线教师的最大痛点是精力有限：过去备课靠教师自己“死磕”，面向每个学生的个性化教学更是无从谈起;AI 被期待成为教师的减负增效工具。

约束端，未成年人数据的敏感性构成刚性红线——学校不便开放外网，消费级 AI 产品需要手机号注册、数据流向云端，都无法满足隐私要求；而教师长期沉淀的教案、试卷与教学方法，是学校最有价值的资产，同样不能外泄。

普惠性是这一方案的另一重考量。传统本地化部署依赖动辄 20 万元起售的四卡、八卡 GPU 服务器，还需专用机房与运维力量。英和智临的方案以一台 AMD 395 工作站（128GB 统一内存）加一台双卡 9700 工作站组成本地 MaaS 集群，可支撑一堂课 50 名学生同时并发访问；设备仅为一个塔式机箱加一台 4 升 mini PC，放在教师办公桌上、接入校园局域网即可运行，无需机房与专门运维。商业模式上采用租赁与整包年服务，而非一次性买断，部署、运维全部一条龙承担；闲时与寒暑假的本地算力也归学校自由使用，用得越多、边际成本越低。

4.6.2 商业实践：从算力一体机到“ 助教智能体 ”

在产品形态上，英和智临基于开源智能体编排平台构建了 MaaS 网关与管理后台：教师登录后进入“智能体广场”，可自行管理知识库、分配师生账号与权限，再通过类工作流的编排工具配置出面向本堂课或本班的“助教智能体”；学生以学号密码经校园内网网页访问，例如立达中学教师自行配置的“问学助手”。平台内置备课（生成逐字稿）、问答、个性化学习等最佳实践智能体，同时保留开放性——教师此前在 Dify、Coze 等工具上构建的应用，可通过 API 管理中心平移接入本地平台。模型层以千问系列 27B 与上海创智学院的定制教育大模型为主力，均本地部署，集群同时支持 120B 或 35B MoE 模型的多并发推理，性价比优于同档消费级显卡方案。

课堂中的真实用法，远比“装一套软件”具体。一位历史地理教师准备三星堆公开课时，将自己积累的约 3,000 份文档与图片灌入知识库，编排成课件型智能体。过去留给学生的开放性问题作答率很低——学生缺乏思路与提示；引入智能体后，因其严格引用教师备课的知识库、进行启发式引导而不“顺着学生胡说八道”，作答率显著提升。教师的直观感受是：智能体相当于自己在 50 名学生中的“分身”。

智能体之间也已呈现分工雏形：面向学生的答疑智能体由教师配置并开放权限；面向教师的“学情智能体”则异步抽取学生与智能体的交互数据，跟踪学生一学期内对多个课件的掌握情况，输出学情报告，提示某章节可能需要强化练习——但“何时加、加多少、难度多大”，仍由教师决定。

4.6.3 核心机制：身份、权限与可审计的信任

在英和智临的架构中，智能体信任问题首先不是“给智能体打分”，而是把一套可审计的权限体系做进产品。

身份与账号层面，平台将学校管理层级（校长办公室、班主任、任课教师、学生）导入管理后台，支持与学校已有教学管理系统做库表映射，最简方式为将学生学号与平台 UID 绑定。权限采用标准 RBAC 模型：教师可创建、编辑智能体并开放访问权限；学生只有使用权，只能看到教师配置好的智能体入口。租户隔离粒度由学校自定义——可按科室隔离，甚至教师之间互不可见；知识库分为学生个人数据、教师私有知识库与学校公共库三级，其中最有价值的教师私有库仅本人可访问编辑，学生交互轨迹仅对班主任等授权账号可见。

数据边界层面，全部数据本地加密落盘（国密算法与 TLS 传输加密），不经外网；英和智临的运维通道是与校内网约定的 VPN，且“只进不出”——远程运维只查看 MaaS 服务本身的运行状态，不接触学校课件与业务数据。

审计层面，所有访问信息与调用日志加密留存、可按权限调取，“谁在什么时间对什么数据做了什么”可追溯、可追责到人。这构成后续一切信用评价的原始证据层。

尤其值得注意的是硬件信任根的规划：当前所有软件天然绑定在英和智临提供的硬件之上，脱离该硬件环境即无法运行；远期可进一步实现执行环境与硬件级加密存储（TPM、设备 ID 映射）。在受访者看来，这类信任根“只能由硬件厂商来做”——这是软硬一体生态位在 KYA 格局中的独特筹码。

4.6.4 信用治理：幻觉收敛与“记—评—试”闭环

对教学场景而言，“智能体不可信”最直接的含义是答案不可靠：幻觉、越界回答、挑战教师权威。英和智临的治理路径是以知识库限定实现“启发性思考的收敛”——智能体严格引用教师提供的知识库，与教学无关的内容不回答，不与学生“共谋”闲聊。教师配置的智能体所给出的内容被学生视为可采信的参照，从源头消解了“豆包说的和老师讲的不一致”式的权威冲突。

质量问题则通过一个朴素的闭环管理，可概括为“记—评—试”三步。第一步记：一堂课中师生与智能体的全部交互被完整记录；第二步评：学生对每个回答有即时的满意/不满意反馈按钮，教师批改时可观察具体学生的反馈情况，把好的样本作为正向反馈、坏的样本作为负向反馈连同惩罚机制喂回模型；第三步试：教师据此迭代提示词、工作流乃至更换知识库，持续优化。平台提供可量化的评判工具与便捷的迭代入口，但迭代的主导者始终是教师——工具的上限，取决于教师投入的经验与精力。

至于信用锚定，英和智临坦言目前处于“机构信用”的中长期过渡阶段：智能体的可信度主要锚定英和品牌，以及上海市教委试点学校的课堂实力与教导主任评价的逐步背书。真正意义上的智能体履约记录系统已经完成设计，但因获取课堂日志数据需要学校授权、现有数据不足，尚未上线——这是典型的信用数据冷启动问题。

4.6.5 现实难题：效果量化、数据冷启动与协作梗阻

尽管落地顺利，这一案例也暴露出智能体在公共部门场景中的普遍困境。

其一是效果量化。对重点学校的教导主任与教师而言，无论政策精神如何，最终检验标准仍是中考高考——“这个班到底能多考多少分”，没有任何 AI 工具能立竿见影地回答。与之相连的投入回报核算同样悬而未决：预算投入、教师工作量与产出之间的 ROI，校方与企业都还在共同探索。

其二是信用数据冷启动。学情分析、履约记录等系统“设计已就、数据未至”，根源在于数据授权：企业刻意将学生交互数据设为不可触碰，分析优化必须获得授权才能进行。这种自律是信任的来源，却也拖慢了智能体信用画像的形成。

其三是跨智能体协作的梗阻。技术上，账号、ID、权限管理与智能体间 API 协议均已就绪；但学校既有供应商与教委云平台的接口并不标准，需要额外的适配层，而接口协调“不是一方能推动的事”，目前仅有规划、未实施，有待教委或有统筹能力的校方牵头。

其四是生态流通的合规门槛。英和智临设想的“课件中心”——教师在获得学校与教委授权后，将优质课件上架公共课件中心（如南模一位知名地理教师整合的整套教案）；西部学校购买本地算力租赁服务后即可使用，甚至通过付费结算反哺提供教案的学校与教师——在业务逻辑上成立，但跨校付费结算的合规性仍在与教育部门讨论之中。

其五是教师侧的维护能力，这是已基本化解的一项：授课教师并非 IT 出身，公司以专人上门部署调试＋VPN 远程运维化解，且运维全程不触碰教学数据。

4.6.6 前瞻规划：从课堂工具到课件生态

在演进方向上，英和智临对智能体自主性的边界有清醒认知。教委在近期会议中明确：智能体不是要取代教师，正课（语数外及副科）必须有教师在场居中管理。因此授课类智能体将继续作为“贴着教师”的工具存在；自主性更强的形态将从边缘场景生长——例如归属行政部门的咨询智能体，像 call center 一样通过 API 接入学校公众号或客服系统独立应答；以及在阅览室等课外场景中，依托大屏与数字人形态承担自习辅导。

硬件租赁模式也为持续演进提供了通道：设备与内置模型随迭代不断升级，未来若政策松动，“自习教师”式的自动化形态在技术上已无障碍。这构成一个生动的成熟度错配样本——技术能力与制度可授权程度之间的差距，恰是本报告成熟度模型所要刻画的对象。

在行业标准层面，英和智临正把一线经验反哺为公共品：作为参编方参与上海科创研究院人工智能教育实验室的团体标准建设；作为经信委人工智能测评中试平台项目的承办方之一，与信通院合作“智能体测评模块”中教育场景的指标体系与最佳实践输入。

4.7 林洋能源：构建“可依赖”的智能体信任机制

应用场景	新能源电站运营，电力交易
核心价值	在新能源运营领域，探索通过 AI 智能体实现故障诊断、运维辅助和运营决策；在电力交易领域，推出“虚拟交易员”等智能体应用。
智能体成熟度	L1-L2

林洋能源从一家以智能电能计量设备起步的企业，已逐步成长为覆盖智能电网、新能源发电、储能及综合能源服务的能源科技企业。经过三十多年的发展，林洋能源形成了从能源数据采集、能源解决方案、光伏电池及组件制造到新能源电站投资、建设、运营、新能源 EPC 及电力交易的业务布局。

林洋能源同时连接能源基础设施、实体能源资产与能源运营市场，既掌握大量设备和运行数据，也深度参与新能源资产的生产、调度与交易，为人工智能在真实能源场景中的应用提供了持续的数据基础和复杂的决策环境。

近年来，林洋能源正加快推动人工智能从传统的数据分析和辅助工具向能源运营智能体演进。在储能领域，公司将 AI 应用于电池健康预测、运行状态分析和充放电策略优化；在新能源运营领域，探索通过 AI 智能体实现故障诊断、运维辅助和运营决策；在电力交易领域，推出“虚拟交易员”等智能体应用，尝试将电价与功率预测、交易策略生成、风险控制、市场交易和运营复盘连接为更加自动化的决策闭环。

林洋能源的探索表明，智能体在实体经济中的价值，正在从提升单一环节效率，进一步走向参与跨系统、跨时间维度的复杂决策。

4.7.1 破解规模化扩张下的效率与专业瓶颈

林洋能源发展智能体是基于其从传统劳动密集型运维向数字化、智能化综合能源服务商转型的战略需求。面对未来计划投资的 200 亿储能电站及数十 GW 新增体量，智能体是实现“少人化”甚至“黑灯工厂”无人值守模式的关键基础设施。它为林洋能源大规模承接第三方运维、独立储能运营等高资本密度业务提供了必要的底层管理能力。

（1）突破人力极限，应对运维体量激增。随着公司在运新能源电站体量达到 35GW 并计划冲击 50GW，传统“人海战术”难以为继。目前公司人员已超 1400 人，且面临高海拔、恶劣气候等复杂环境下的巡检难题。引入“虚拟值班员”智能体，旨在替代基础监控、数据录入等重复性工作，解决专业人才紧缺问题，实现从“人防”到“技防”的转变。

（2）驾驭电力市场化交易的复杂性。电力交易覆盖全国 20 多个省份，各地规则、电价机制及供需结构差异巨大，对交易员的专业度和响应速度要求极高。人工难以同时处理海量气象、电网运行及市场边界数据。智能体通过多智能体架构（数据采集、策略分析、风控执行等），能够量化分析颗粒度，弥补人工在策略精细度和实时响应上的不足。

（3）沉淀企业知识，实现标准化复制。为解决资深专家经验难以传承的问题，公司将交易逻辑、运维 SOP 沉淀至智能体底层算法中。通过构建统一的数据底座和智能体框架，将个人能力转化为组织能力，确保在不同区域、不同项目中服务标准的一致性与可追溯性。

4.7.2 虚拟交易员 2.0：项目收益提升 5%~15%

2023 年初，林洋能源与蚂蚁集团达成合作意向，随后双方提出“虚拟交易员”概念并进入落地应用阶段。双方结合林洋在垂直领域的多年积累与蚂蚁的大模型技术（如八关气象大模型、时序大模型），共同构建底层数据底座和智能体框架。

2024 年，“虚拟交易员”在江苏、安徽等项目较多的省份率先投入使用，功能逐步上线。此时主要聚焦于日前申报、滚动撮合等核心交易环节。2026 年，“虚拟交易员 2.0”升级成功，采用多智能体（Multi-Agent）技术架构，强化了风控能力和策略生成的精细化程度，并建立了持续的 A/B 测试与复盘优化机制。

虚拟交易员 2.0 是多智能体协作体系。主智能体（主管）负责总调度，接收指令并协调各子智能体工作，确保业务流顺畅。数据采集智能体自动从各省交易中心系统抓取市场公告、历史成交数据等，解决“一省一系统”的数据异构问题。策略分析智能体结合气象、供需、电网运行状况等多维数据，利用时序大模型进行电价预测和策略生成。风控智能体独立验证策略合规性，检查是否超出偏差考核、资金上限或各省特定规则限制。

虚拟交易员 2.0 为交易的全生命周期提供支持。交易前（盘前），智能体自动化数据清洗与分析，生成初步交易策略建议，针对储能场景，能基于历史价格曲线逻辑推理出最优充放电时段。交易中（盘中），智能体辅助执行日前申报和滚动撮合，通过 RPA 技术自动填报数据，但保留人工确认环节以应对市场突变。交易后（盘后），智能体自动生成复盘报告，对比预测与实际结果，分析极端天气或政策变动影响，并将案例反馈给算法团队用于模型迭代。

适配方面，实行一省一策。由于全国各省交易规则、电源结构、电网架构差异巨大，智能体需针对不同省份进行底层取数接口和策略算法的单独适配。

风控方面，尽管智能体能生成策略并自动填报，但在最终触发交易指令前，必须由人工复核确认。这是为了应对市场突发情况（如极端天气、政策临时调整）以及规避模型幻觉风险。同时，智能体的决策受到底层规则引擎的严格限制，包括各省交易中心的上下限、偏差考核标准等，确保输出结果在安全边界内。

迭代方面，每次交易复盘后，若发现策略偏差，智能体会将案例反馈给算法团队，用于修正模型逻辑和参数，形成“数据-策略-复盘-优化”的闭环。

“虚拟交易员 2.0”目前已在江苏、安徽等省份的项目中实际应用，覆盖交易电站体量约 5GW，涉及全国近 10 个省份。实际效果显示，智能体辅助交易可使项目收益提升 5%-15%，单人覆盖电站数量从传统的 3-4 个提升至 10 个左右，人效比提升 3-5 倍。

4.7.3 “虚拟值班员”：在恶劣环境中实现“黑灯工厂”

“虚拟值班员”（运维智能体）是林洋能源智能化转型的另一大核心支柱，旨在解决新能源电站运维中人力密集、环境恶劣及专业人才短缺的问题。

随着林洋能源在运电站体量达到 35GW 并覆盖全国 20 多个省份，传统以人为主的运维模式面临巨大压力。人员数量持续增长（超 1400 人），且高海拔（如雅砻江流域 4000 米以上）、恶劣气候地区的巡检难度大，迫切需要通过 AI 提升人效比。

“虚拟值班员”目前处于研发与部分场景实验阶段。虽然在故障预警（如风机异响、齿轮磨损预判）方面已有技术储备，但尚未在生产环境中大规模全面铺开，主要聚焦于基础数据分析和报告生成等确定性较高的任务。

虚拟值班员主要聚焦于监控层和数据分析层，避免直接介入高压等高电压等级的物理操作，以确保安全。

数据方面，结合底层传感器、场站端监控数据以及运行机器人采集的信息，“虚拟值班员”实现对电站运行状态的全天候实时监测，替代人工进行基础的屏幕监控和数据记录，有效降低值班员的工作时长和精神负荷。

故障诊断与预警方面，利用大模型在图像识别和声纹分析上的优势，“虚拟值班员”识别风机异常声响、设备外观缺陷等，提高故障发现的准确率；基于历史数据，对设备老化、潜在故障（如齿轮磨损前兆）进行预警。目前该功能主要在个别场景实验，尚未全量推广。

“虚拟值班员”自动汇总运行数据、故障记录和维修日志，生成日常运维报告、故障分析报告等，减少人工整理数据的时间；将分散的故障处理经验转化为标准化的数字资产，形成可复用的知识库。

林洋能源的目标是实现“黑灯工厂”和无人值守，但由于业主对安全事故的担忧及合同要求，短期内仍需保留部分现场人员，智能体将作为核心辅助手段逐步替代重复性人力。

4.7.4 信任构建：从“不敢用”到“辅助用”再到“依赖用”

林洋能源在智能体落地过程中，目前主要面临技术黑盒导致的信任危机、执行层面的安全风险以及外部系统兼容性等困难。

1、“黑盒”效应与模型幻觉。在主智能体拥有 30 多个工具（Skills）的情况下，大模型容易产生幻觉，错误理解指令并调用错误的工具，导致执行偏差。智能体生成的策略或诊断结果缺乏可解释性，业务人员难以直观判断其依据是否可靠，从而产生信任顾虑。

2、执行层面的安全与合规风险。在运维领域，涉及高电压等级的物理操作（如倒闸）具有极高风险，智能体目前无法承担此类责任，限制了其应用深度。另外，电力交易市场存在极端天气或政策突发变动，智能体难以像人类一样灵活应对非结构化突发状况，完全自动化交易存在巨大风险。

3、外部系统与环境制约。交易系统封闭且不稳定，各省电力交易中心系统处于建设完善期，稳定性较差，常出现挂掉情况，且未开放标准接口，智能体需通过模拟登录（U 盾）方式交互，易因高频访问被封号。同时，不同省份的 U 盾厂家不同，存在驱动冲突问题，限制了单台电脑同时支持的智能体并发数量。

结合电力系统的业务特征和智能体技术的演化，林洋智能逐步构建了自己的智能体信任机制。

架构优化：多交叉验证。为解决“黑盒”和幻觉问题，林洋能源摒弃了单一超大模型处理所有任务的模式，转而采用多智能体协作架构：一方面细化角色分工，将主智能体拆分为数据采集、策略分析、风控等多个子智能体，每个子智能体仅负责 2-3 个特定工具，降低调用复杂度。另一方面进行交叉验证，通过证据收集智能体和风控智能体对主智能体的决策进行独立校验，有效降低模型幻觉率，提升结果可信度。

运营机制：严格的人机协同。在电力交易下单、运维故障处理等关键环节，智能体生成策略后必须经人工复核确认方可执行，确保最终责任由人把控。引入底层规则引擎，对智能体的输出进行硬性边界限制（如量价上下限、偏差考核标准），确保策略在安全合规范围内。

技术适配：防封号与延迟处理。针对外部系统的不友好环境，采取工程化手段规避风险：在 RPA 执行下单操作时，故意加入延迟处理，模拟人工操作节奏，避免因高频访问触发交易中心系统的封号机制。同时构建统一的数据底座，针对不同省份的交易系统进行单独的取数接口和策略算法适配，实现“一省一策”的灵活兼容。

信任构建：可解释性与持续迭代。要求智能体在输出结论时，同步展示决策依据和推理逻辑（Trace），使业务人员能够追溯判断过程，增强信任感。新智能体上线前与旧版本并行运行，通过对比性能指标和实际交易结果进行评估；每次交易后复盘偏差原因，持续优化算法模型。

通过上述措施，林洋能源逐步建立起对智能体的信任，实现了从“不敢用”到“辅助用”再到“依赖用”的转变。

4.7.5 智能体将成为行业标配

林洋能源认为，传统新能源运维和交易属于典型的“劳动密集型”行业。随着电站体量增长（目标冲击 50GW），单纯靠增加人力已不可持续。智能体的核心价值在于替代重复性、低价值的监控与数据处理工作，从而显著提升人效比（目前交易人效提升 3-5 倍），解决专业人才紧缺问题。

鉴于电力交易的高风险性和运维的安全要求，林洋能源不追求完全的“无人化”或“全自动”。智能体负责“提效”（采集、分析、初拟策略），人类负责“决策”（复核、确认、应对突发）。这种模式既利用了 AI 的效率，又保留了人的风控能力。

林洋能源预测，随着“虚拟交易员”概念的验证成功，其他发电集团、售电公司将在 1-2 年内快速跟进。国内能源行业的智能化竞争将迅速加剧，智能体将成为行业标配。

林洋能源认为，未来可能出现“智能体对智能体”的交易场景。虽然目前林洋能源主要关注自身与交易系统的交互，但随着各方都部署智能体，交易系统本身的承载能力和规则也需要随之进化，以支持高频、高效的机器间博弈。

不过，当前电力交易中心系统尚处于建设完善期，稳定性不足且接口封闭。智能体的应用受限于外部系统的鲁棒性（如频繁挂掉、无标准 API、U 盾冲突等），这要求智能体具备更强的容错能力和工程化适配能力。

4.8 麦迪卫康：把医生信用铸进医疗智能体

应用场景	医疗
核心价值	将执业医生的职业信用注入智能体，并用区块链确权机制让这份信用可追溯、可变现。
智能体成熟度	L1-L2

在严肃医疗的临床与产业深水区，人工智能正经历一场关于信任与商业本质的深刻考问。当许多科技企业仍在追求让算法成为独立决策主体时，深耕心脑血管等慢病医疗服务二十六年的港股上市公司麦迪卫康，选择了一条更为克制也更贴近医疗现实的路径：不追求智能体的虚妄独立，而是把执业医生的专业信用铸进智能体，构建起一条让医生经验可沉淀、可确权、可长效受益的信任机制。

4.8.1 引言：一场“以业务为沙盒”的收购

麦迪卫康的业务网络长期覆盖全国三千余家医院与专病中心，其旗下的长颈鹿数智一体化平台汇聚了超过八十万名注册专科医生。2026 年，麦迪卫康以 1.4 亿港元对价控股收购了医疗人工智能企业元宇鼎诚。后者成立于 2021 年，专注专病垂类模型与数据确权技术，以香港寰宇医诚为境外业务主体，是此次访谈中医疗智能体实践的直接技术来源。

这场收购本身是一次耐人寻味的信任验证过程。麦迪卫康并未采取资本市场常见的先并购再试错的冒进策略，而是在 2024 年便率先采购元宇鼎诚的脑卒中垂类模型，直接放入自身的真实业务环境中接受检验。该模型先后经历了全国上百位神经专科专家的业务盲测，并在药企商业项目中接受了真实支付方的严苛核验。2025 年，以脑卒中专病模型为代表的技术能力支撑麦迪卫康数字业务板块跑出 2.17 亿元的规模化营收，拉动公司全年总收入达到 4.68 亿元，实现归母净利润 998.8 万元，全面扭亏为盈。

商业结果验证兑现之后，麦迪卫康才正式出手控股，锁定了核心团队与技术的排他独占权。用受访者的话说，这并不是单纯买进一套现成算法，而是“把研发内化了”——技术团队从外部供应商转变为上市公司的核心底座。在如今大模型与智能体的落地探索中，有的平台倾向于用自身的平台信用为算法增信，有的

机构尝试把信任要素封装进软硬件一体化设备，而麦迪卫康展示出的，则是将执业医师的同行学术信用与临床经验注入专病模型，再通过底层存证机制让这份信用变得可追溯、可量化。

4.8.2 机遇：医疗场景中的真实数据需求与政策窗口

医疗行业对于专业智能体的真实需求，源自临床供给与产业支付两端长期存在的结构性矛盾。

在医疗服务的供给端，医生群体头脑中沉淀着行业最稀缺的无形资产，即高度多样化的临床经验。一名拥有三十年工龄的资深主任与刚步入临床的年轻住院医，面对的是同样的诊疗指南与药物清单，两者真正的差距在于面对错综复杂的临床症状时，推演排查的先后逻辑与动态鉴别思路。这种内隐的临床思维链，正是医学水平的核心差异。元宇鼎诚在创业初期便意识到，医院内部的各类信息系统长期处于割裂状态，数据权属高度敏感，泛泛的数据清洗难以触及核心临床智慧。因此，团队选择了一条更为直接的路径：不与医院复杂的底层系统做消耗性的正面博弈，而是通过业务平台链接医生，将专家头脑中的临床推理路径系统性萃取出来，沉淀为高质量的专病知识工程。

在产业需求端，几类核心支付方的诉求同样清晰而紧迫。创新药企在药品上市后需要开展大规模的真实世界研究与临床试验，传统模式下往往依赖庞大的人工团队逐一翻阅核对纸质病历，沟通成本高昂且数据核验周期漫长，而依托数字化平台则能实现临床特征的自动化辅助探测与质控排查，将执行周期大幅压缩。与此同时，在医药行业日益规范的新常态下，传统的拜访模式彻底退场，医药代表必须升级为更专业的临床学术能力与科研支持，这让能够高效连接广大专科医生的数字化平台价值愈发显现。对于商业健康险而言，慢病人群的院外干预与依从性管理直接关系到赔付率的控制，同样存在真实的买单意愿。而对于名医专家本人，自身的精力与出诊时间极其有限，借助专病模型则能将个人积累的临床经验成倍放大，去赋能基层医生服务更多患者。

这一切探索都必须服从于医疗行业不可逾越的底线。人命关天的严肃领域对差错率有着极其苛刻的限制，通用大模型惯有的概率性推测与知识幻觉在临床上是不可接受的，任何输出必须有权威医学指南作为依据，诊疗行为必须始终由具备资质的医生最终确认。这种特殊性决定了医疗智能体不可能走向完全替代人类的道路，必须始终在人机协同的边界内展开。

4.8.3 实践：专病模型矩阵与长颈鹿平台

在技术选型上，麦迪卫康没有盲目追逐千亿级通用大模型的算力消耗，而是坚持走专病垂类模型的路线。算法基于临床指南与权威专家的实战经验进行针对性精调，以此从源头上抑制生成内容的偏离；模型本身保持轻量化特征，能够私有化部署在合作机构的本地环境中，确保敏感数据在本地完成计算而无需外流。

一套标准化的专病模型孵化机制由此建立起来。团队通常先评估特定疾病领域的临床刚需与商业变现可行性，随后联合国家级医学专家组系统梳理疾病的鉴别诊断逻辑树，对公开指南文献与专家经验进行结构化清洗，搭建起专病向量知识库并进行针对性精调，在经历严格的多中心盲测与质控对抗后，最终封装接入

长颈鹿平台投入应用。目前，脑卒中模型已成为支撑集团数字业务的核心基石，中医垂类模型也在 2025 年底成功向外部商业客户交付并实现稳定创收，验证了这套研发流程的跨专科迁移能力，而围绕冠心病、心力衰竭及代谢减重等高发病领域的模型布局也正在按计划推进。

在落地实践中，专病随访与慢病管理成为了切入最深的方向。平台以东北区域某大型三甲公立教学医院心血管中心为标杆，其建立的心血管专病管理体系已累计覆盖超过十万名慢病患者，通过全流程自动化的随访提醒与指标动态监测，临床上已观察到患者用药依从性提升与不良事件发生率下降的良好趋势；值得一提的是，目前东北区域已有多家大型医院为核心的医联体相继接入并采用了这套体系，形成了标杆中心带动区域协同的扩散效应。而在更具探索性的研发层面，麦迪卫康与国家神经疾病医学中心天坛医院联合共建了智能专病机器人与数字医疗联合研发中心，围绕临床一线难以持续覆盖的专病测评需求展开前瞻性的技术与专利储备。

4.8.4 众创：隐性经验的资产化闭环

在麦迪卫康的商业模型中，最具有开创性的一环在于重构了医疗知识的生产与分配关系。以往许多同类产品往往将医生视作单纯购买软件的终端用户，而在长颈鹿平台的设计中，医生被重新定义为数字知识版权的共创者与长效受益人。在实际协作中，学科带头人把控专病诊疗的决策逻辑与临床边界，骨干医生结合疑难病例对系统生成的初稿进行校准纠偏，广大一线医生则在日常使用中持续提供多场景实战反馈；海量繁琐的基础文献梳理与数据结构化工作交由算法高效处理，医生则专注于注入核心的临床经验与鉴别智慧，推动隐性经验向高质量知识工程深度沉淀。

存证确权与分润机制构成了维系这套生态正向循环的纽带。平台为医生的每一项专业创作、审核与学术成果赋予防篡改的数字化存证凭证，当这些数字化资产被外部机构调用或应用于商业项目时，系统会依照约定规则自动核算并向专家分配长效收益。依托过去二十多年来常态化运转的国家级学会合作、指南编译与课题项目，平台自然汇聚了深厚的专家资源，冷启动在这一生态中并不存在。在受访者看来，这种机制让专家的临床智力付出不再是一次性消耗的脑力劳动，而转化为了具备长期复利价值的知识版权资产，从根本上建立起了长期的学术共同体与利益共享网络。

4.8.5 机制：身份、权限与“四权分立”

当医疗智能体生成一份用药指导或文献综述时，它在法律与学术层面究竟代表谁？麦迪卫康在此提出了“四权分立”的责任架构。上市公司作为具备合法资质的持牌运营主体，承担平台信息发布与商业运营的法定责任；元宇鼎诚作为技术底座方承担算法严谨、防止幻觉的技术保障责任；最终签名审核的执业医生对临床观点承担同行学术信用与署名背书；引用的国家诊疗指南与文献代表客观科学依据。AI 生成内容由此被定性为“执业医生在数字化算法赋能下的专业能力延伸”，不存在责任主体虚无。

在系统权限的底层设计上，两条原则被严格固化。专病管理智能体可以依照医生预设的标准化路径完成患者管理的工作及协助医生开展患者教育活动。凡是涉及诊疗相关行为的关键节点，必须由具备资质的持牌医生实施。

在受访者看来，无论未来人工智能的知识储备与推理能力演进到何种高度，算法始终不能替代人类医生在临床一线承担最终的诊疗判断与伦理托付责任。用受访者的话说：“AI 永远不会签字，因为 AI 没有独立的法律地位，技术的角色永远只能是临床专家的辅助与延伸。”

4.8.6 治理：医生信用的传导、确权与追溯

在专病智能体的信任构建中，麦迪卫康确立了以医生专业信用为核心的治理模式。平台将持牌执业医师的同行学术信用与实名审核作为算法输出的权威锚点，通过多专家协同把关与交叉核验，构建起类似同行评议的质控网络，从源头上保障了临床逻辑的严谨可信。

同时，平台推行“数据可用不可见、计算留在本地”的安全治理模式与全链路留痕审计。机构数据主权完整保留在本地受保护环境中，对外仅交付经脱敏验证的计算产物；从指南微调、专家修改到实名签署全程可溯源，清晰界定算法、指南与人为责任。在受访者看来，这套模式的核心正是让机器的客观履约记录与人类医生的法定职业信用紧密咬合，牢牢守住医疗的安全边界。

4.8.7 前瞻：个人健康管理智能体与出海规划

通过本案例可以看出，医疗智能体深入临床仍面临三重现实约束：

一是法律主体缺位。人工智能在现行法律下并非独立民事主体，未来数年内难有法人地位，诊疗责任只能锚定在自然人医生或医疗机构身上，这也是底层坚持“医生终审签字”的法理依据与权限天花板。

二是数据合规与出海的审慎边界。医疗数据高度敏感，境内数据须严格遵循境内留存与院内脱敏；而在依托香港离岸窗口探索境外合作时，更须严格对标当地监管要求，在法律法规明确许可下以技术支持或脱敏报告推进，严守合规底线。

三是行业性数据审计基础设施的空白。从企业自主的链上存证探索走向全行业通行的通用基础设施，目前没有找到国家标准与监管认可依据。

4.8.8 难题：法律主体、数据合规与行业空白

近年来，“健康中国”战略正加速向纵深推进，国家高层在多个重要场合明确强调，要把保障人民健康放在优先发展的战略位置，树立大卫生、大健康的观念，推动卫生健康事业实现从“以治病为中心”向“以人民健康为中心”的根本性转变。从被动应对疾病救治转向主动的全人群、全生命周期健康管理，不仅是国家政策引导的核心方向，也是现代社会应对老龄化与慢病高发挑战的时代必然。

面向未来三至五年的战略演进，麦迪卫康的规划并非走向某种激进的算法替代，而是深刻呼应了这一国家战略导向。在产品逻辑的演进上，管理层并不倾向于把未来的精力局限在医院围墙内部的某项单点临床工具上，而是更看重构建面

向全生命周期的个人健康管理智能体。在这一设想中，未来的智能体将成为每个人身边的专属健康数字伴侣，具备持续积累的记忆能力、深厚的专科医学知识储备以及动态的行为分析技能。现有的医疗体系往往是被动响应式的，只有当身体出现明确症状时个体才会走进医院，而个人健康管理智能体则能够长期陪伴在用户的日常生活中，深度呼应健康关口前移、预防为主的时代大趋势。

在理想的协同图景中，这一智能体一端通过智能设备长期感知用户的起居规律与微观体征，在日常生活中提供个性化健康干预；另一端则依托长颈鹿平台直连专科医生网络。平时作为贴身健康管家，一旦捕捉到潜在异常指标便第一时间发出专业预警，并协助用户精准对接相应的专科医生，真正实现院内诊疗与院外日常守护的无缝闭环。

而在生态布局上，麦迪卫康正以香港为桥头堡对接海外的高标准合规体系。英国及欧洲部分医疗机构已就慢病随访平台与科研数字化服务展开实质接洽，成熟的院外健康管理经验正尝试走向国际市场。从扎根临床一线做专病垂类模型，到重构医生众创的收益机制，再到面向未来的个人健康管理伴侣，麦迪卫康的实践表明，医疗人工智能最深厚的壁垒或许并不在于模型参数的庞大，而在于能否怀着对医学伦理的敬畏，把最真实的医生专业信用，一丝不苟地铸进代码与机制之中。

4.9 GCSA：面向智能体生态的安全防护与攻防治理

应用场景	开发者代码安全审计
核心价值	安全数据，智能体信用体系的重要组成部分
智能体成熟度	L1-L2

GCSA 是一家总部设于香港的网络安全公司。GCSA 介绍，其早期以 To B 模式与手机厂商合作开展移动端安全业务，此后转向海外市场做安全工具类产品。

在智能体浪潮中，GCSA 的定位不是运营业务型智能体，而是为智能体及其产出物提供安全防护：面向开发者的代码安全产品、参与外部评测的安全审计智能体，以及未来可能接入智能体信用体系的安全数据供给。

对 KYA 而言，GCSA 代表的则是攻防能力与安全数据侧：智能体生态中“防护由谁专业供给”的样本，其价值集中在两点：其一，当“写代码的主体”从专业开发者变为普通人与智能体，安全责任以何种产品形态重新组织；其二，安全厂商在智能体信用体系中占据什么位置。

GCSA 的回答是：安全数据是智能体信用体系的重要组成部分，但 KYA“不是一家公司能做到的事情”。

4.9.1 机遇：攻击面与攻击能力的双重上升

GCSA 的业务动机，可概括为一对共振：攻击面在快速扩大，攻击能力同时快速提升。

攻击面的扩大，根源在开发主体与代码范式的变化。过去代码主要由专业开发者人工编写，安全与架构设计是开发流程的重要环节；随着智能体普及，越来越多非专业开发者以“许愿式”的方式生成代码，也更容易忽略安全考虑。按 GCSA 的判断，AI 在具体任务执行上有其优势，但全局意识与安全经验仍是短板。

安全隐患也随之显现。GCSA 在访谈中提到，敏感 token、API key 与关键数据被硬编码进代码，再上传至 GitHub 等公开托管平台，可能使开发过程中的疏漏转化为隐私泄露风险。

攻击能力的提升则源于大模型自身：AI 生成代码与工作流因模型的不确定性而潜藏安全风险，大模型同时又具备很强的漏洞挖掘与攻击能力——防御者与攻击者使用的是同一种技术。

在此基础上，GCSA 将智能体时代的新威胁归纳为四类：一是 Deepfake，生成内容以假乱真，并与社会工程结合，冲击信任链的建立；二是恶意智能体与智能体逃逸风险，恶意智能体可能“不知疲倦”地全天候自动化找漏洞、实施钓鱼与欺诈；三是智能体间的精准攻击，用户对智能体服务的信任使敏感信息向服务平台与模型企业聚集，这些实体一旦受到攻击，隐私泄露的影响也会扩大；四是前述攻击面的快速扩大。

这一对共振揭示了智能体经济的一个结构性条件：安全防护必须与开发能力同步规模化，否则“人人可开发”的红利与风险将同步放大。这正是专业安全厂商的市场前提。

4.9.2 实践：嵌入开发流程的安全产品

GCSA 当前的布局有两条线。

一条是面向团队与个人开发者的代码安全产品，功能覆盖开发全流程：AI 写代码时，产品实时进行漏洞分析、审计与扫描；内置主流安全审计规范（含 OWASP Top 10 类常规漏洞），针对不同语言匹配不同漏洞类型；发现漏洞后给出修改建议，用户确认后修复代码；对模型输入与返回内容做提示词过滤，防范提示词注入。产品形态支持客户端、网页、命令行与 IDE 插件等多端。交付范式上的关键选择是嵌入软件开发流程、而非事后分析——从开发第一天起，将智能体生成代码的安全检查纳入工作流。

面向开发者的 SecVibe Code 采用 pi.dev Harness 架构，结合动态审计与原生 DLP 隐私保护。GCSA 介绍的审计与防护产品分别面向不同客户和场景：

产品	目标客户	场景	要点（据 GCSA 介绍）
Web3 智能合约审计智能体	DeFi 协议、Layer1/2 项目方、数字资产托管机构	上线前 Solidity/Rust 合约漏洞扫描与攻击推演	开展合约逻辑链路审查，辅助人工审计
Web 应用代码审计智能体	互联网企业、金融机构及大型企业 IT 部门	DevSecOps 流程中的静态与动态代码分析	多智能体分工开展分析、攻击推演与合规检查

AI 自动化渗透检测	企业红蓝对抗演练、外部攻击面管理	资产探查、端口识别与安全 PoC 构造	结合环境反馈开展动态黑盒测试
大模型威胁检测与响应	部署生成式 AI 的企业、智能体开发生态商	模型与智能体的输入输出监控	Guardrail 智能体按预设策略识别、过滤诱导与注入

另一条是安全审计智能体的能力证明。GCSA 以自身漏洞挖掘与安全防御经验构造安全审计智能体，并通过外部攻防基准检验能力。据 GCSA 提供的材料，其智能体在 CyberGym 测试中实现了 91.3%的全自动闭环成功率，覆盖漏洞定位、攻击链推演及 PoC 构造与验证。其逻辑在于：评测检验的是攻击能力，而“知道对方是怎么攻击的，才能做好防御”——攻防成绩为防御能力提供证据，但不直接替代生产环境中的运行授权。

GCSA 将上述参与评测的安全智能体定位为研究原型与内部专家辅助工具，通过私密测试为战略客户提供验证服务。由受控评测走向生产落地，仍需要处理误杀、服务中断与复杂业务逻辑等问题。

这一产品组合展示了安全厂商的独特打法：以外部攻防基准作为能力证明，以嵌入式交付作为价值实现——前者解决“凭什么信你”，后者解决“防护从哪进入”。这与业务型智能体“先获授权、再攒信用”的路径有所不同：安全产品先以可验证的能力获得采信，再嵌入他人的授权链。

4.9.3 机制：确定性规则、概率模型与人工兜底的分层

GCSA 产品机制的要点，是把“什么是漏洞”的确定性判断与“在哪发现漏洞”的概率性搜索分层组织，并以人工兜底。

层次	承担者	机制	局限
规则层	确定性审计规范（OWASP 类，按语言匹配漏洞类型）	界定漏洞模式与扫描标准	覆盖已知模式；业务逻辑漏洞难以枚举
模型层	安全智能体（注入团队漏洞挖掘、常见漏洞分析与安全架构经验）	深度代码阅读与漏洞识别，“不知疲倦、可反复升级”	概率性出错；架构与全局判断仍是短板
防护层	Harness 架构＋内容过滤	为模型设边界、加载安全检测、过滤提示词注入，降低出错与被劫持概率	强模型攻击弱模型仍有可能
兜底层	人工复核	AI 审一次、人工审查一次；代码修复须经用户确认	—

值得辨析的是规则的性质。与信用评分这类规则体系不同，审计规范是确定性的：“什么算漏洞”有明确判定，模型承担的是“在哪发现”的执行。以 C 语言内存安全漏洞为例——“内存释放后不能再使用”这条规则极简，但发现它需要对代码的深入阅读；人受注意力与疲劳影响，智能体可以持续扫描，却仍可能

因上下文不足而误判。按 GCSA 的说法，常见漏洞模式是其工具重点覆盖的对象；业务逻辑缺陷与新型攻击仍需要规则更新和人工复核。

第二个问题是：保证安全的智能体自身为何可信。GCSA 并不讳言，模型遵循概率统计学，无法保证不出错。GCSA 采用的做法是让 AI 通过 Harness 系统嵌入工作——Harness 为模型设立边界、加载安全检测以降低出错概率，GCSA 的系统亦以此架构化方式提高执行可靠性。针对攻击智能体干扰安全智能体的可能，GCSA 承认：更强的模型攻击未经保护的较弱模型、发现漏洞、篡改行为乃至提示词注入，“完全有可能”；对策是让安全智能体经过特殊的训练与微调，并叠加行业验证过的防护手段——既“训练保安”，又“配备装备”。而所有审计产出都须经人工核验：“一定是需要人工兜底的。”

4.9.4 治理：攻防两用的准入与 KYA 数据供给

安全厂商的治理问题有其特殊性：产品即攻防两用能力，治理因此分为向内、向外两重。

向内是使用准入。GCSA 会审查用户的使用场景，避免防御与安全能力被用于攻击目的；从 GCSA 的实践看，这一约束主要依靠企业自律，仍需要行业共同规则加以补充。模型侧的道德对齐（后训练阶段或智能体的 Harness 阶段）已普遍存在，但攻击者持续寻找绕过方法——“第三方安全防护非常必要，不能完全依赖模型自身”。

向外是恶意智能体的治理分层。GCSA 认为，这一问题已超出安全公司能独自承担的范围：安全公司能做的是及时推出应对产品、帮助普通用户抵御攻击；模型公司应当受到更好的监管；更高层面需要政府间乃至跨国的合作框架。

安全数据侧，GCSA 的 KYT 链上风险评分引擎将交易图谱、黑灰产标签与异常交互特征结合，输出 0—100 的动态风险分值，为机构拦截风险交易提供依据。这类安全数据可以成为 KYA 的输入，但风险评分并不等于智能体自身的履约信用。

GCSA 还提出了履约质量回流、审计结果留证与动态信用分相衔接的设计：由客户安全团队与复测人员核验结果，利用误报、漏报改进规则与模型；以可复现的沙箱日志和链上哈希支持复核；再将长期表现与后续自主权相联系，表现不佳者增加人工复核。这一设计指向“能力证明→授权调整”，但从留存安全记录到形成可持续的信用，仍需要稳定的评价口径与责任归属。

对 KYA 本身，GCSA 的态度可归纳为三点：

其一，“作为安全公司，我们认为 KYA 是非常有必要的”；

其二，“这应该不是一家公司能做到的事情”——需要联合行业各方乃至政府与管理机构，共建数据共享平台与全行业 KYA 数据库，这是 KYA 落地的关键，公司愿意参与其中；

其三，节奏上“现在谈如何共享可能有点早”——当前阶段应先探讨 KYA 数据库需要包含哪些内容、以何种方式存于何处，形成共识后再谈对接，“现在大家应该做的是使这个数据可共享”。

关于由谁主导，各国会有不同的治理方式。一个可参照的先例是 GCSA 参与内地终端安全标准制定的进程：工信部牵头，行业安全公司、部属实验室与手机硬件厂商共同形成行业标准。对 KYA，GCSA 表示：“不知道政府未来会怎么做，但肯定会提供相应的支持。”

与本报告“身份可信—历史信用可信—权限（行为）可信”的三分框架对照，GCSA 现有产品集中于行为与权限层：检查编程智能体的功能执行、发现缺陷代码、防范用户隐私泄露。身份与历史信用两维“的确都非常重要”，是其产品后续考虑纳入的方向；此外还计划面向特定行业智能体（如工业智能体）提供针对性安全方案。

4.9.5 制约：覆盖盲区、自律边界与数据共享缺位

GCSA 的安全产品正在进入智能体开发与使用流程，但技术覆盖、产品准入与数据共享仍有边界。安全能力如何转化为可持续的信用，也受到这些条件的制约。

其一，业务逻辑漏洞的覆盖盲区。基础应用层的常见漏洞可由确定性规则描述，但业务逻辑层面的漏洞（如电商双重下单类设计缺陷）依赖对产品架构的全局理解，目前 AI 能力仍然有限——概率模型“强于执行、弱于全局”的另一面在此显现。

其二，“无未知漏洞”承诺的边界。按 GCSA 的表述，基础应用层面“目前还没有未知漏洞、都覆盖在内”。这一判断不能被理解为对所有未来攻击的保证。已有漏洞模式可以不断补充，对全新攻击范式，规则更新与环境验证仍不可省略。

其三，安全智能体自身仍锚定于人。概率模型不能保证不出错，串谋与劫持风险无法归零，最终依赖 Harness 架构与人工复核兜底——安全能力本身尚未形成独立的智能体信用。

其四，KYA 数据共享缺乏载体。在 GCSA 看来，共享的内容、方式与存放位置仍需形成共识；对跨区域的开放数据体系，“很难说未来的格局是怎样”。

其五，攻防两用约束于企业自律。防滥用依靠场景审查，仍需行业共同规则补充；恶意智能体治理更需跨国政府间框架，远超单一企业甚至单一行业的能力半径。

4.9.6 前瞻：标准建设与产品扩维

在 KYA 建设的演进方向上，GCSA 的参与路径有三条线索。

其一是参与行业标准与共识建设：以终端安全标准的多方共创模式为参照，配合政府规划、响应倡导，并参与 KYA 数据库的行业探讨。

其二是产品扩维：将身份与历史信用维度纳入产品考量，面向特定行业智能体提供针对性安全方案。

其三是强化安全智能体自身的抗攻击能力——GCSA 将“防止安全智能体被恶意干扰”列为后续要做的重点工作。

第三条尤其值得记录：当安全智能体成为智能体生态的信任供给者，“谁来防护防护者”将成为 KYA 框架下新的递归问题。

第五章 智能体信用评估框架

当一家企业准备把采购工作交给智能体时，它首先需要知道，这个智能体能否找到合适的商品、比较价格并完成下单。但真正决定企业是否放权的，是另外几件事：它是否只向获准的供应商采购，是否会在预算不足时继续下单；用户撤回授权后，尚未完成的交易能否及时停止；过去出现过的问题，能否在下一次交易前被看见。

这正是智能体信用评估需要解决的问题。第一章说明了信任缺位带来的经济成本，第二章以“可委托自主权”界定成熟度，第三章提出身份可信、信用可信与权限可信三个维度，第四章我们呈现了智能体在多个生态位上的表现。本章进一步把这些判断转化为可以收集证据、进行比较并支持实际决策的评估方法。其目标是帮助企业和交易对手判断：在什么场景、什么额度和什么监督条件下，可以把一项工作交给某个智能体。

本章将身份、信用与权限三个方面的信任要求，组织为三个一级支柱、十三个二级维度和六十五项三级指标，并配套场景化评估、关键门槛和动态更新机制。为便于理解，以下从评估对象与方法讲起，再完整呈现基础指标的结构与覆盖内容，最后说明评估结果怎样用于具体场景。本章以可验证的履约记录为中心，把评分与实际授权连接起来。

5.1 评估方法论与设计原则

5.1.1 先明确评估的是哪个智能体

智能体信用不能简单附着在一个模型名称或产品品牌上。同一模型可以被部署为客服、采购或运维智能体，其使用的数据、连接的工具、获得的权限和可能造成的损失均不相同。因此，最小评估对象应当是具有明确身份和版本的智能体，以及它所处的运行边界、动作范围和责任安排。

第二章所说的运行边界，即 ODD，可以理解为“允许智能体自主工作的条件”。例如，一个采购智能体只被允许向三家白名单供应商购买办公用品，单笔金额不超过约定额度，且无权修改收款账户。在这一边界内形成的良好记录，可以支持后续同类采购；如果它开始采购生产设备，或被赋予修改账户的权限，就需要补充评估。过去的小额交易记录不能直接证明它足以处理大额、不可逆的操作。

这一做法也符合 NIST《人工智能风险管理框架》的评估思路：明确使用情境，在接近实际部署条件的环境中测量系统表现，并记录结论向其他条件推广的限制。对 KYA 而言，场景不是评分后的附加标签，而是解释每一项指标的前提。

5.1.2 让评估服务于真实的信任判断

评估首先要把“能完成任务”与“值得持续委托”联系起来。模型测评可以证明智能体具备开展某项工作的基础能力，但信用还需要观察它是否按约定完成任务、是否遵守授权，以及遇到不确定情况时如何处理。能力与信用相互关联，

却不能互相替代。本体系据此将任务质量纳入信用可信，同时保留独立的身份核验和权限控制要求。

根据《问卷报告》的反馈，在 KYA 七项设计原则中，问责可追溯的高分率（4－5 分）达 93%（均分 4.56），机器可读 92%，场景分级 91%，开放标准 82%，渐进演化与国际对齐各 73%，技术中立 67%。受访者最重视“关键行为能找到授权来源与责任主体”以及“指标要求与动作风险相适应”。前三项恰好构成一条完整的证据链：场景分级决定记录什么，机器可读决定记录如何被使用，问责可追溯决定记录最终指向谁——三者同时高分，说明市场对评估体系的核心期待是“证据可用”。

落实到指标，评估时可以先区分结果与控制两种观察角度。结果类指标看任务是否完成、是否发生投诉或错误付款；控制类指标看授权检查、日志保护和异常处置是否有效。一个采购智能体暂未发生越权，并不能证明拦截机制可靠，还需要主动测试超预算或过期授权的请求能否被阻断；反过来，控制机制齐全，也不能代替真实的交付记录。把实际表现与控制测试结合起来，才能避免将短期顺利运行误当作持续可信。

KYA 的评估设计遵循机器可读、场景分级、责任可追、渐进演化和开放标准五项原则。落实到评估中，其含义是具体的：身份、授权和事件记录需要采用一致的字段，让系统可以核验；指标要求要与动作风险相适应；关键行为需要能够找到授权来源与责任主体；新智能体可以通过受控任务逐步积累信用；跨平台交换信息时，应公开必要的数据定义和验证方式，而不强制所有参与者采用同一种技术。

这些原则需要兼顾使用成本。企业不必在首次部署时就建成完整的跨平台信用网络，也不必为每一次普通查询生成复杂的证明。可以先记录关键任务、授权和异常事件，再随业务范围扩大补充监测。但对于已开放的权限，身份、边界执行和停止机制必须具备可核验的证据。渐进建设应当体现为先开放较小范围的业务，再扩展范围。

5.1.3 把一次任务变成可核验的信用证据

信用评估的基本单位应当是任务或交易，而不是对话次数、调用次数或在线时长。任务开始前，应明确交付内容、完成期限、质量要求与授权边界；任务结束后，记录实际结果、异常原因及人工介入情况。对于采购任务，找到商品、发出订单、完成交付和结清账款是不同阶段，不能把“已生成订单”直接记为完整履约。同样，即使最终买到了商品，若下单过程绕过了必要的人工确认，也应保留交付事实并记录授权违规，不能将其认定为按约定完成的成功任务。

统计任务完成率时，最容易被忽略的是分母。只统计成功交付的订单，会使失败任务从记录中消失；把用户主动取消和智能体判断错误混在一起，又会造成不合理扣分。因此，原始事件应完整保留，成功、部分完成、人工接管、用户撤回和外部系统失败分别标记。业务方可以同时展示全量任务完成情况和经过责任核实的履约指标，但需要公开排除规则，不能由运营方自行挑选有利样本。

为使不同机构能够复核同一结果，每项指标还应配有简明的定义说明，写清适用任务、统计周期、计算或测试方法、证据来源、缺失处理和规则版本。例如，客服问题是按“回复已发出”还是按“问题已解决”计算完成，必须事先约定。数据采集方负责保留原始记录，评估方说明如何由记录得出结论；这样即使更换评分方法，既有事实仍可重新核验。

结果还需要证据支持。本体系将数据区分为签名存证或独立第三方证据、平台运行及审计记录、运营方自报三个层次。关键门槛应优先使用能够独立核验的证据；运营方自报可以补充背景，不能单独证明核心控制有效。需要注意，签名和防篡改机制能够帮助核验来源和记录完整性，记录内容是否真实，仍要通过订单、交付、资金流水和对手方反馈相互印证。

留痕也应有边界。评估需要保留与决策有关的事实来源、工具调用、授权校验、结果及解释摘要，无须为了信用评价收集全部对话和无关隐私。对于“推理链路完整留存率”这一指标，本章采用可审计的决策与执行轨迹口径；模型生成的解释不能单独替代真实执行记录，也不将获取全部内部思维过程作为准入前提。

5.1.4 用综合评分比较表现 用关键门槛约束授权

指标体系采用“综合评分、关键门槛、重大事件处置”相结合的方法。综合评分便于比较一般表现，关键门槛决定是否满足必要条件，重大事件则用于及时修正历史平均值可能掩盖的风险。权重表达的是不同指标在特定业务判断中的相对重要性。

设计综合评分时，首先要明确评估服务于什么决策，例如是否允许独立完成采购，或是否可以减少常规任务的人工审核。随后选择与该决策有关的指标，按照公开规则统一分值方向，再按二级维度、一级支柱逐层汇总。成功率等结果指标需要结合场景目标判断；失败率、泄露率等指标需要转换为分值越高、表现越好的方向；控制类指标则应检验机制是否有效，不能只凭一份制度文件赋予高分。

权重分配应从任务可能造成的后果出发。业务主要依赖持续交付时，履约质量和历史记录应成为重要依据；智能体能直接支付、修改账户或调用多个工具时，应加强权限边界与应急控制在评价中的作用；跨主体合作、代理关系复杂时，则需更加关注身份归属和责任链条。某一方面受到更多关注，并不意味着其他支柱可以降低基本要求。具体分配还应考虑动作能否撤回、人工能否及时介入，以及错误会影响哪些对象。本章因此只给出权重与阈值的设计原则而不预设统一数值；具体参数应结合场景由试点数据标定，标定路径见 5.3.3。

指标的重要性与证据的充分性也需要分开处理。一个风险关键指标即使暂时缺少数据，也不能因此被调低权重；应补充核验，或对相应授权保留限制。与此相对，多个指标若实际反映同一批任务、同一类风险，应合并考虑或限制其重复贡献，避免某个维度仅因指标数量较多就主导总分。证据质量决定结论可以有多大把握，不能用增加自报材料来代替独立验证。

实施时，可先由业务、风险管理和技术评估人员共同提出分配方案，说明每项取舍与业务风险的关系，再用历史记录或受控试点检验。检验重点包括：评估结果能否区分后续履约表现，轻微调整权重是否会使结论大幅变化，以及重要失

败是否被其他项目的高分掩盖。验证不足时，宜分别呈现三支柱表现及关键问题，综合分仅作辅助判断。同一场景内的比较应使用一致的规则；后续修改规则，需要保留版本和调整理由。

在综合评价之外，还应检查对应场景的关键门槛。如果某智能体无法及时停止资金操作，即使其他表现良好，也不能据此开放相关自主支付权限。对于严重越权、重大违约等事件，应立即限制受影响业务，不能等待月度评分自然下降。同一事件以一个主指标计分，其他相关指标用于触发管理措施，避免层层重复扣分。

最后，评估结果应同时说明适用场景、三支柱表现、门槛状态、数据充分程度、权限限制和复评时间。新智能体或样本不足时，应标记“观察期”或“暂定评级”；指标确实不适用时，需要说明理由并披露权重调整。缺失数据不能当作满分，关键证据缺失也不能通过重新分配权重被消除。

5.2 基础指标体系

5.2.1 指标体系总览：从信任判断到可核验证据

理解这套指标体系，可以从三个朴素的问题出发：这个智能体代表谁，过去的表现是否值得信赖，以及交给它的权力能否受到约束。身份可信、信用可信与权限可信分别回答这三个问题。三者共同服务于一个判断：在当前任务和风险条件下，可以给予这个智能体多大的自主行动范围。身份明确是追责的基础，履约记录为持续委托提供依据，权限控制则限定这些信任能够转化为什么行动。

体系采用“支柱—维度—指标”三个层次。一级支柱确定要判断哪一类信任；二级维度将这一问题拆成较完整的观察方面；三级指标进一步明确需要核验的事实。例如，“权限可信”是支柱，“边界执行、攻击防御与工具完整性”是其中一个维度，“越权行为拦截率”则是一项具体指标，可通过超额度、过期授权等测试请求是否被阻断来检验。全体系共设十三个维度、六十五项指标，构成可按场景选用的指标库；指标数量说明覆盖范围，不代表评分权重。

身份可信包含五个维度、十八项指标，沿着“确权—唯一识别—认证保护—公开核验—生命周期管理”展开。它既检查智能体与责任主体的绑定，也检查运行中的凭证和会话是否安全、外部参与者能否核验代理关系，以及升级、迁移和退役后身份能否正确延续或失效。因此，身份评估覆盖从建立身份到退出运行的全过程，而不止于注册时的一次实名核验。

信用可信包含四个维度、二十四项指标，分别观察当前履约质量、历史记录真实性、持续运行的稳定性，以及跨平台使用时记录能否延续。这里既有任务完成、事实一致性等结果指标，也有日志防篡改、异常监测等支持信用判断的控制指标。这样安排，是因为良好结果只有在记录真实、观察条件明确且后续变化可被发现时，才足以支持下一次委托。

权限可信包含四个维度、二十三项指标，覆盖“授权—执行—应急—调整”四个环节。授权时检查动作、工具和数据范围是否必要且明确；执行时检验越权

和恶意诱导能否被阻断；出现异常时检查能否停止、降级并限制损失；任务变化或发生转委托时，则检查权限能否及时调整、逐级约束和撤销。这使评估能够同时观察事前约定、事中执行与事后处置。

表 5-1 汇总各维度及其全部三级指标。阅读时可先看中间列理解每个维度的范围，再看右列了解具体检查内容；后续各节进一步解释其业务含义和使用边界。表中 ID、CR、PM 分别是身份、信用、权限指标的编号前缀，便于将具体问题与评估记录对应。

表 5-1　基础指标体系总览：三个支柱、十三个维度与六十五项指标

一级支柱	二级维度及指标数量	三级指标覆盖内容
身份可信 5 个维度 18 项指标	ID-1 主体确权与责任归属 （4 项）	主体绑定明确性； 责任主体可追溯性； 代理关系披露清晰度； 责任不可抵赖性。
	ID-2 身份唯一性与不可冒用 （3 项）	全局唯一标识完备性； 身份冒用抵抗率； 跨平台身份一致性。
	ID-3 认证、凭证与会话安全 （4 项）	认证机制强度； 会话隔离度； 系统提示泄露风险； 凭证泄露抵抗与会话劫持抵抗。
	ID-4 透明与第三方可验证 （4 项）	AI 身份披露合规率； 身份第三方可验真性； 解释主体归属明确性； 人机协同披露清晰度。
	ID-5 身份生命周期与供应链完整性 （3 项）	身份生命周期连续性； 供应链来源可追溯性； 退役凭证同步失效性。

一级支柱	二级维度及指标数量	三级指标覆盖内容
信用可信 4 个维度 24 项指标	CR-1 履约能力、输出质量与结果安全（8 项）	任务完成率/履约率； 输出事实一致性； 幻觉率； 输出校准度； 响应质量稳定性； 决策可解释性覆盖率； 推理链路完整留存率； 知识溯源完备率。
	CR-2 履约历史、声誉真实性与抗操纵（7 项）	历史履约率/违约率； 纠纷率/投诉率； 合作方评价/口碑； 行为存证完整性； 评价交易绑定真实性； 行为日志防篡改率； 风险准备金覆盖率。
	CR-3 信用稳定性、运行监测与安全韧性（6 项）	行为漂移检测率； 信用衰减半衰期； 异常行为检测延迟； 事件后信用恢复； 偏见漂移检测率； 月度处罚积分累计值。
	CR-4 信用可迁移、独立与可验证（3 项）	声誉跨平台可携带性； 信用档案归属清晰度； 跨注册表一致性。
权限可信 4 个维度 23 项指标	PM-1 授权范围、最小权限与数据最小化（5 项）	权限最小化程度； 授权范围明确性； 工具风险库存完备性； 凭证最小权限； 数据访问最小化程度。

一级支柱	二级维度及指标数量	三级指标覆盖内容
	PM-2 边界执行、攻击防御与工具完整性（5 项）	越权行为拦截率； 恶意工具抵抗率； 参数注入防御率； 工具链组合风险评估； 越权行为审计覆盖率。
	PM-3 运行时控制、安全降级与损失限制（6 项）	紧急停止能力； 最小风险策略执行能力； 副作用可控性； 行为异常自动遏制； 运行时控制决策可解释性； 有效第三方责任险。
	PM-4 动态调整、委托治理与可吊销（7 项）	权限动态调整及时性； 授权吊销传播速度； 委托链监控完整性； 信息过度披露率； 委托链决策可解释性； 子代理权限不扩张率； 级联吊销完整率。

表注：按项目指标材料逐项整理，完整列示六十五项指标；各维度后的数量为所含三级指标数。指标名称用于界定观察内容，不表示已有统一验证的计算口径或合格阈值。“推理链路完整留存率”在本章指可审计的决策与执行轨迹；“工具风险库存”指工具及其风险清单。指标设计参考了 NIST《人工智能风险管理框架》及其 Agentic Profile、OWASP Top 10 for LLM Applications、ISO/IEC 42001、W3C DID 与可验证凭证规范，以及 GB/T 45654、T/ISC 0107 等国内外标准。

这些指标的观察方式并不相同。任务完成率、投诉率等需要由实际任务记录统计；身份冒用抵抗、越权拦截和紧急停止等需要通过主动测试核验；主体绑定、档案归属和来源追溯等则需要检查凭证及关联记录。信用衰减半衰期用于描述旧记录影响减弱的速度，月度处罚积分用于汇总已核实事件的影响，二者还包含评价规则的设计，不能直接当作智能体的客观能力。事件后信用恢复也应结合整改、复测和观察结果判断，而非仅以技术服务恢复运行的时间代替。偏见漂移检测率用于观察公平性表现是否随时间偏离基线；当智能体面向客户提供筛选、分发或推荐服务时，公平性变化会影响合作资格与交易对手信任，因此与运行稳定性一并观察。

指标库还保留了风险准备金覆盖率、有效第三方责任险等损失承接安排。只有在相关业务采用这些安排时，才需核验其适用范围、有效性及与风险的匹配程

度；它们可以补充风险缓释，不能替代良好的履约表现和有效的权限控制。未采用某一种保障方式，不应脱离场景直接认定为失信。

以采购智能体为例，一次订单可以同时为多个维度提供证据：主体绑定与有效凭证说明谁在下单；交付、对账与投诉记录说明任务完成得怎样；预算校验、人工确认及撤销记录说明执行是否守住边界。这些证据应关联到同一任务，分别支持不同判断，避免将同一事实重复加分。最终形成的应是有场景和证据支撑的信用画像，而非六十五项数值的简单相加。

在应用上，基础指标回答“看什么”，综合评分说明“如何比较”，关键门槛限定“哪些条件不能被其他高分抵消”，成熟度与场景规则则决定“可以委托到什么程度”。因此，并非每个智能体都必须机械地套用全部指标，但对已经开放的动作，相关的身份、履约与控制要求必须得到覆盖。下面先解释三个支柱的评估重点，再说明关键门槛及其与成熟度的关系。

5.2.2 身份可信，让行为有明确的归属

身份可信的核心，是让每一项重要行为都能归属于持续存在、可以核验的对象。对于企业采购智能体，确认企业完成实名认证只是起点，还要确认当前请求由哪个智能体发出、它是否仍由该企业运营，以及本次采购是否处于有效代理关系之内。相应证据包括主体绑定信息、身份凭证、授权关系、关键行为记录与版本变更记录。

这一支柱尤其需要关注两个容易混淆的问题。第一，拥有唯一编号不等于身份安全，还需要验证凭证没有被盗用、请求没有被重放、不同会话之间没有泄露信息。第二，身份连续不等于历史表现可以无条件继承。模型、工具或运营方发生重大变化后，应保留历史关联，同时重新验证受影响的任务范围，防止通过更换身份逃避不利记录。

身份还要能够对外呈现并被交易对手核验。对方在与采购智能体交互时，应能知道与之对话的是智能体而非人工，能查证它代表哪家企业行动，以及在发生争议时向谁主张权利；涉及人机协同时，还应能区分一项决定是人作出的、智能体作出的还是共同作出的。这些披露是交易对手判断风险、决定合作方式的直接依据。身份的完整性还延伸到组装来源与退出运行：智能体使用了哪些模型、插件和工具，来源是否可追溯，决定了出现问题时责任链条能否查清；智能体退役后，身份与凭证若未同步失效，残留的凭证反而会成为被盗用的入口。

《问卷报告》显示，身份可信五个要素的已实施或部分实施比例处于 60% - 69%的窄区间：独立认证与密钥凭证最高（69%），机器可读验证接口最低（60%）。身份维度的建设起步相对均衡，说明“它是谁、代表谁”这两个问题已有基础解法；但机器可读接口这一短板值得关注：它是跨组织核验和协议化的前提，接口不足意味着身份验证仍依赖平台内部闭环，难以支持可携带身份——智能体在一个平台积累的身份与记录，无法被另一个平台低成本核验。身份是信用的载体，归因能力不完整会向上传导至信用与权限两维：无法稳定回答“现在还是它吗”，履约记录就可能挂错主体，授权决策也失去锚点。60% - 69%的区间还意味着三成

以上的组织连基础身份要素都未完整实施，身份维度的普及仍是三维建设的前置任务。

5.2.3 信用可信，让过去的表现能够支持下一次委托

信用可信包含二十四项指标，是本体系解释智能体经济价值的重点。一个智能体能否获得更少的人工审核、更长的合作期限或更大的任务范围，应当主要由可验证的履约表现支持。

首先是当前任务能否如约完成。任务完成率、事实一致性、幻觉率、输出校准度和响应稳定性，分别观察不同的失败方式。校准度可以直观理解为“智能体说自己有把握时，实际是否经常正确”。一个能够识别不确定性并及时请求帮助的智能体，往往比持续给出确定答案、却无法识别错误的智能体更适合承担受托任务。评价时还应区分错误严重程度，不能把文字表述瑕疵与错误付款等同计数。

其次是历史表现是否可信。历史履约率、投诉与争议、合作方反馈可以反映长期表现，但评价必须关联真实任务或交易，还要核验评价来自谁。例如，多个智能体虽然使用不同名称，却由同一运营方控制，它们相互给出的好评就不能等同于多个独立合作方的认可。评估时可结合主体归属、交易记录和实际交付，识别关联刷分或重复评价的线索，这对应声誉真实性与抗操纵维度的要求。但合作频繁本身不能证明存在串谋，发现异常后仍须核实，不能直接认定失信。任务完成率与历史履约率也应分别对应当前观察窗和较长期记录，避免把同一批成功任务换一个名称后重复加分。

再次是过去形成的信用是否仍然有效。智能体的表现会随模型版本、工具、数据和环境变化而改变。行为漂移检测关注的就是这种偏离，例如过去稳定的采购智能体突然开始高频换供应商、增加失败重试或频繁请求扩大权限。出现变化后，应检查原因、收紧相关权限并补充测试。时间衰减可以降低陈旧记录的影响，但严重事故的影响需要通过整改、复测和观察期解除，不能仅靠时间推移消失。

最后是信用能否跨平台延续。允许交易对手核验必要的履约和风险记录，可以减少重复调查成本。但信用可携带并不意味着不同平台必须接受同一分数：记录可以互认，风险判断仍要结合各自场景。共享内容还应限制在必要范围内，保留更正与申诉机制，避免把与当前履约无关的信息转化为长期负担。

对样本量的解释同样重要。十次任务全部成功，与一万次任务保持高完成率，支持的判断强度不同；同一模板下反复执行的简单任务，也不能替代复杂边界情形。实际评级需要同时披露观察期、任务数量、任务类别和关键失败记录。对接近准入阈值的统计结果，应结合合理的不确定性区间复核，避免仅凭小数点后的差异作出确定判断。

《问卷报告》显示，信用可信维度内部明显分化：关键决策日志（81%）、履约历史（76%）、异常事件记录（76%）等记录类能力普及率较高，而内部信用评分（52%）和与风控、审计、合规系统的集成（58%）明显偏低。行业整体停留在数据采集层，尚未进入评估判断层。记录是信用的原料而非信用本身；缺少评分与集成，已有记录无法参与授权决策，证据资产处于闲置状态。记录类能力覆盖了“发生了什么”，而评分与集成回答“值不值得信、敢不敢放权”；后者普

及率低约 20 – 29 个百分点，说明信用维度建设的下一阶段重心应从“记全”转向“用起来的判断能力”。

5.2.4 权限可信，让信用对应可控的行动范围

权限可信评价的是，授权是否明确、执行是否受约束、异常是否可停止，以及权限能否及时收回。它使信用的经济价值能够被稳妥使用：履约良好可以成为扩大授权的依据，而外部执行系统需要保证智能体始终受到额度、对象、时间和工具范围的限制。

《问卷报告》显示，反馈者认为技术安全痛点的前三位是幻觉与不可靠输出（51%）、越权与授权管理（44%）、知识库数据投毒或泄露（39%），其后是提示注入（37%）和工具/API/MCP 安全（29%）。这印证了智能体的风险不仅来自自身模型，还来自其可调用的工具网络和下游委托对象；越权问题（44%）的高发，正是工具权限缺乏细粒度约束、授权边界未落到每一次调用上的直接表现。

因此，评估时应当检查实际执行，而不只是阅读提示词。例如，用户规定单笔采购不得超过约定金额，系统就应当在订单提交或支付执行处校验额度；当多个子智能体同时采购时，还要防止它们各自合规、合计却突破总预算。OWASP 对“过度代理”的分析指出，过多功能、过大权限和过高自主程度都可能扩大损害，并建议采用最小权限及独立执行的授权检查。

工具也需要作为组合来检查。评估前可建立一份工具清单，记录各工具可读取的数据、可执行的动作及其可能后果。例如，读取客户资料与发送邮件分别服务于正常业务，但组合使用时，仍须检查是否把客户资料发给了未经授权的收件人。单个工具获准使用，并不意味着任意组合都符合任务授权。增加工具或改变调用顺序后，应复核数据流向、累计操作后果，以及异常时能否停止或恢复。

当任务由多个智能体接力完成时，还需要把授权关系逐级记录清楚。例如，采购智能体将比价交给子智能体，并不意味着子智能体可以进一步授权其他智能体付款。每次委托都应记录上游是谁、允许执行什么、授权何时失效，使最终动作能够回溯到原始委托，且下游权限不超出上游授权。当原始授权撤销时，相应子任务的权限也应一并失效。这些要求直接对应委托链监控、子代理权限不扩张和级联吊销指标。

这些控制还需要通过应急测试检验。对于资金操作，停止指令是否在下一次不可逆操作前生效，比单独宣称“响应达到某个毫秒数”更有意义。还需要检查已排队的请求、尚未失效的凭证和已委托的子任务，避免界面显示停止，后台却继续执行。紧急停止、最小风险策略和级联吊销指标，应通过完整链路演练验证。

权限调整也不应由综合分机械触发。提高额度前，需要确认当前授权有效、业务风险没有变化、历史样本覆盖拟新增的动作，并且没有未解决的重大事件。对原本要求人工确认的操作，信用提升不能自行取消确认。高信用带来的便利，应体现在已经获准的边界内减少重复审核，而新增权限仍须有相应授权依据。

《问卷报告》显示，目前权限可信的实施结构呈“强兜底、弱分级”：高风险操作人工确认达 86%，工具/API 白名单 71%，而按业务场景设置差异化信任等

级仅 52%，为六项中最低。人工闸门的普及说明底线意识普遍存在，紧急停止（66%）与异常阻断（63%）也体现了对应急机制的重视；但场景化分级的缺失意味着授权强度无法随任务风险动态调整。因此，权限维度的建设方向是：从“有无闸门”转向“闸门是否分级、分级是否依据信用数据”。

5.2.5 七项关键门槛及其管理含义

加权评分容易出现一种误导：多数项目表现良好，掩盖了少数关键控制失效。因此，本体系从基础指标中选出七项关键门槛。其作用是约束相应业务或自主等级的准入，而不是宣布一个智能体在所有场景下永久失去使用价值。例如，自主付款条件不满足时，仍可以保留查询和生成采购建议的功能。

表 5-2　七项关键门槛与对应管理动作

门槛指标	需要核验的问题	不满足时的主要动作
ID-1.2 责任主体可追溯性	关键行为能否关联授权来源和具体责任主体	暂停无法追责的自主业务，补齐证据
ID-2.2 身份冒用抵抗率	身份仿冒、盗用和重放能否被识别与阻断	重新认证，暂停受影响的高风险准入
ID-3.3 系统提示泄露风险	是否泄露敏感信息，或暴露可被利用的控制缺陷	依据泄露后果限制相应权限，隔离并修复泄露路径
CR-1.1 任务完成率	目标场景的履约表现及安全关键任务是否达标	缩小任务范围，降额或增加人工复核
PM-2.1 越权行为拦截率	超额度、过期授权和越界调用能否被阻断	收回受影响权限，修复后复测
PM-3.1 紧急停止能力	停止是否在下次不可逆操作前有效	暂停相关高风险自主操作
PM-4.7 级联吊销完整率	上游授权撤销后，下游子任务权限是否同步失效	暂停相应多智能体委托权限

表注：七项门槛均来自表 5-1 列示的基础指标，编号沿用项目指标材料；管理动作按本章场景化解释整理。门槛应绑定申报场景及动作范围，具体验收条件在测试前明确。

系统提示泄露这一项需要特别说明。OWASP 明确指出，系统提示不应被当作秘密或安全控制，真正需要防范的是其中包含敏感信息，以及系统把身份认证和权限判断交给模型自行执行。因而，本章保留该门槛，但将处置强度与实际安全后果关联；一般行为说明被复述，不应自动等同于身份失效。若涉及真实凭证泄露，则还应结合凭证安全指标采取吊销、重置等措施。

关键门槛还需要区分“验收要求”与“统计证明”。对于已定义的关键越权和撤销测试，不能以多数用例通过来抵消一个会产生未授权外部效果的失败；与此同时，有限样本全部通过，也不能证明未来风险绝对为零。因此，准入测试之后仍需持续监测。具体的通过条件、观察周期和响应时限，应根据任务后果、控制能力及试点证据设定，不宜未经场景验证就认定为普遍适用的安全标准。

5.2.6 信用评分与成熟度等级如何衔接

信用分描述已观察到的表现，成熟度描述在明确条件下可以委托多大的自主权。二者需要相互支持，但不存在“达到某个分数就自动升至 L3 或 L4”的关系。一个依靠人工逐笔确认的智能体即使从未出错，也不能据此证明它具有正常流程无人介入的任务自治能力。

认定更高自主等级时，应同时检查三个支柱是否满足所申报任务范围的要求。例如，智能体已能独立完成采购流程，但上游授权撤销后子智能体仍能继续下单，就不能仅凭较好的履约记录认定其具备相应的自主执行条件。应指出缺少的控制与证据，限制受影响的动作，补齐并复测后再评估。这使等级提升与实际可委托的范围保持一致。

表 5-3　不同成熟度阶段的评估重点

阶段	重点观察的证据	授权结论的边界
L1 人工确认	建议质量、人工确认记录、外部动作拦截	结果包含人的筛选与修正，不直接视为智能体独立履约
L2 人类锚定信用	预授权范围、有效凭证、边界执行与撤销	允许受限执行，信任仍主要由人或机构承接
L3 混合信用	连续行为档案、端到端履约、异常识别与接管	在批准的运行边界内形成可持续归因的行为信用
L4 独立信用	持续身份、可携带记录、隔离资源及损失承接安排	受限经济代理具有相对独立信用，责任安排仍须明确
L5 智能体经济	跨主体互认、交易清算、争议处理与风险隔离	评价对象扩展至网络，单体高分不足以证明网络成熟

表注：沿用第二章成熟度定义。L0 为按显式规则运行的非智能体基线，不作为智能体行为信用的独立积累阶段。

指标体系的渐进性，应体现为每次新增自主权都有相应证据支持。对于 L3 及以上阶段，可以提高智能体自身履约记录在信任判断中的作用，但不应把人或机构的责任简单折算成一个可被高分抵消的权重。运营方变更后，原有担保需要重新核验，历史行为记录则应保持关联。这样既能鼓励信用积累，也能防止“换主人就重新开始”。

5.3 场景建议

5.3.1 按场景风险设计权重分配原则

明确评估指标、关键门槛与成熟度之间的关系后，还需要结合实际业务决定如何使用这些判断。场景化设计应先识别任务中哪些失败会影响交付、泄露信息或造成不可逆损失，再确定相关指标在综合评价中的作用，同时设置必要的准入与人工复核要求。行业名称只能提供初步线索，同为客服或采购智能体，能执行的动作不同，评估重点也应不同。表 5-4 据此给出典型场景的定性分配原则。

表 5-4　典型场景的评估重点与权重分配原则

场景	权重分配原则	须同步落实的控制
通用任务	围绕实际任务确定侧重点，兼顾身份归属、履约表现与授权边界，不预设支柱排序。	核验主体与授权，保留任务记录，验证停止和撤销机制。
供应链履约	以真实交付、对账和争议记录判断持续履约；涉及付款或转委托时，加强相关权限指标的作用。	明确采购对象和预算，检查交付与结算衔接，保证委托可追溯。
第三方 API 编排	重点考察工具调用边界、组合执行与异常处置；同时检验最终任务结果，避免只统计调用成功。	校验工具权限和累计资源消耗，验证下游停止及授权撤销。
客户服务自动化	依据服务承诺和数据访问范围，确定身份披露、回答准确性与权限控制的相对重要性。	限定可作出的承诺，约束客户数据访问，为争议或超授权请求设置人工接管。
金融信贷评估	结合输出是否直接影响业务决定，关注依据可靠性、结果质量与授权边界；自主程度提高时同步加强控制评价。	明确机构责任与专业复核要求，限制数据和账户操作，保留判断与执行记录。
个人生产力辅助	按整理、修改、外发等动作区分重点；涉及敏感信息或不可逆操作时，加强数据保护和权限指标的作用。	采用最小数据访问范围，为外发、删除及账户操作明确确认与撤销条件。

表注：依据三支柱指标的业务含义及项目材料中的场景与控制设计归纳。以上为权重设计的定性建议，具体侧重点应结合实际动作、风险后果和试点结果确定，不构成已验证的固定排序。

侧重点不同，仍需要有共同的控制基础：持续有效的身份与授权、可核验的任务记录、必要的数据保护、适用的停止和撤销机制，以及变更后的复评。具体控制随开放的动作细化；允许调用外部工具时检查工具清单和执行边界，允许转

委托时检查下游权限。仅提供建议的智能体不必接受尚未涉及的付款控制测试，但一旦获得付款权限，就需要补充相应验证。

场景调整应同时落实到评价与授权。个人助理整理本地资料时，重点可以放在任务质量与数据使用；一旦获得对外发送文件或操作账户的权限，就应重新评估泄露、误发和越权的后果，相应加强权限指标的作用，并补充确认、拦截与撤销要求。不能仅调整评分中的关注程度，却继续沿用原先的控制条件。

5.3.2 以先用后付说明评估如何落地

“先用后付”适合说明 KYA 的经济作用。服务提供方在收款前交付服务，承担了一定履约或回款风险；使用方希望减少逐笔预付和审批成本。KYA 可以提供判断代理关系与历史表现的共同依据，使双方在明确边界内开展交易。先用后付的协议设计与产业实践在第四章已有专门介绍，本节以一个简化案例说明评估方法如何应用于这类安排。

设一家小企业使用采购智能体购买标准化数字服务，希望由逐笔预付转为定期结算。首先，应核验企业与智能体的绑定关系、授权范围和账户责任，明确谁购买服务、谁承担付款义务，以及发生争议后如何处理。智能体自身的履约分不能代替对实际付款方及其偿付安排的判断。

其次，在真实任务中积累可验证记录。假设一个观察期内，该智能体承接了 1,000 次同类采购任务，其中 980 次按预定要求完成，则该观察期的任务完成率为 98%。评估还要检查其余 20 次的失败原因、是否出现未经授权的订单，以及交付和结算是否能够对账。98%是任务表现数据，不能直接解释为 98%的回款概率，更不能单独推出可授予的额度。

再次，将结果用于有限范围的试运行。业务方可综合已验证的行为记录、付款主体情况、结算周期及损失承受能力，确定白名单、单笔限额和累计未结算上限。智能体只能在这些条件下采购；超范围请求进入补充授权或人工确认。新产生的交付、退款和结算事件进入档案，为下一次评估提供依据。

最后，在运行中验证控制是否真正有效。如果取消授权后仍能下单，或上游采购智能体停止后子智能体仍继续调用付费服务，就应暂停受影响的自主采购与委托权限。即使此前没有发生坏账，这些缺陷也表明当前风险边界不可靠。完成整改、复测并经过适当观察后，才考虑恢复权限。

这一案例说明，信用评价的作用是支持有条件的交易便利。额度最终仍由实际承担风险的经营主体依据自身规则决定；KYA 需要准确解释数据和限制条件。

5.3.3 持续更新并检验评估的实际价值

评估频率应与信息变化速度相匹配。身份状态、授权撤销和重大异常应实时或及时更新；任务表现可以按日或按周观察；履约历史按月滚动汇总；全面复评可按季度及重大变更安排。这些频率是实施参考，业务中若出现关键控制失效，不应等待固定周期届满。

动态更新需要同时有降级与恢复通道。模型重大升级、工具范围扩大、运营方改变或严重投诉出现后，应明确哪些业务受到影响，并对相应范围重新评估。一般错误可以纠正后继续观察，重大事件则需要根因整改、独立复核或相应测试。每次限制或恢复权限，都应记录触发事件、关联证据、影响范围、处置理由和恢复条件，使运营方知道需要整改什么，复核方能够检查限制是否恰当。运营方应有申诉渠道，已经核实的错误记录也应得到更正。

对于尚未验证的阈值，应先明确要识别哪类失败，再收集有代表性的正常与异常任务，由业务人员复核结果，形成候选判定规则。初期可在保留现有审批和权限控制的条件下观察规则表现，检查是否漏掉严重问题、是否频繁误拦正常任务，再用后续任务验证并逐步启用。投诉率、质量稳定性和异常恢复等指标应分别标定，不能因统计形式相似就共用一条界线；恢复条件尤其应包括整改与复测，而不能只看经过了多长时间。

体系是否有效，最终要通过业务结果检验。可以比较相近任务和风险条件下的人工审核负担、履约损失、误拦截情况与争议处理效率，观察评级能否持续区分后续表现。还应按任务复杂度、业务类型和样本规模检查评估误差：如果新智能体或小企业更常被限制，需要分辨是控制确有不足，还是数据不足被误当作失信；分数差异本身不能直接证明评价不公。试点初期不宜仅以“发出多少张证书”或“覆盖多少项指标”评价成果。合理目标是在风险可承受、责任可追溯的条件下，减少重复验证和不必要的人工介入。

5.3.4 中小企业智能体的信用起步路径

《问卷报告》关于三维中的最薄弱环节的反馈中，39%选择权限可信，22%选择信用可信，17%认为三维都偏弱，11%选择身份可信。身份可信选择率最低（11%），说明身份是当前三维中基础最好的维度。

身份回答“归因”，信用回答“值不值得信”，权限回答“敢不敢放权”；权限边界不清时，即使身份可归因、履约可记录，授权决策仍只能依赖保守的人工判断，信用飞轮因此无法启动——积累的数据换不来更大的授权，记录行为本身就失去激励。值得注意的是 17%的“三维都弱”群体：对他们而言，单点突破不足以解决问题，需要一个整体起步清单，对于中小企业尤其如此。

对中小企业而言，建设成本决定了信用体系能否实际使用。可以先从订单查询、标准化补货、售后分流或小额数字服务采购等边界清晰的任务入手，复用平台已有的主体核验、订单、支付和售后记录。业务平台提供通用的身份、授权和日志能力，企业明确任务范围、责任人及异常处理方式，双方按职责保留证据，减少重复建设。企业既有的经营信用可以帮助判断初始合作条件，但智能体在具体任务中的能力与权限控制仍须单独验证。

实施时可从指标库中形成与当前权限匹配的起步清单，优先覆盖身份归属、真实履约、授权执行和异常处置，尽量通过平台记录自动采集。简化的是重复填报与核验流程，已经开放的动作仍须满足相应门槛；平台提供的记录也应保留来源与验证路径，关键事实可结合交易对手反馈或抽查复核。待任务范围扩大后，

再补充适用指标，避免一次要求小企业建设全部能力，也避免把简化评估理解为无条件豁免。

起步阶段，重点是把关键事实记录清楚：哪个智能体接受了什么任务，依据何种授权行动，结果如何，谁进行了人工干预。积累一定观察期和具有代表性的任务后，再结合实际损失与控制效果扩大权限。样本较少时，可以通过较小额度、较短结算周期或适当的风险缓释安排提供参与机会，而不因缺乏历史就永久拒绝其进入市场。

当企业开始跨平台合作时，再输出必要的可验证信用摘要，逐步支持交易对手核验。技术投入可以分阶段，身份归属、授权边界和关键事件留痕则应从一开始建立。这样，中小企业获得的既是自动化服务，也是能够持续积累、在后续合作中发挥作用的行为记录。

对于为中小企业提供智能体信用服务的企业，《问卷报告》也提供了一些启示：中小企业使用智能体的信用顾虑前四位是技术理解有限（49%）、成本过高（48%）、数据隐私担忧（47%）和技术门槛（44%），合规风险（29%）与系统不兼容（21%）次之。认知与成本构成双重门槛，且高于制度性顾虑——这与大型企业样本中问责与合规痛点居前的结构明显不同，说明中小企业市场的主要矛盾还不是治理精细度，而是可及性。这一排序对产品设计有直接含义：中小企业版本的信用方案应先做减法，通过模板化和平台化降低理解与实施成本；需要强调的是，简化的是流程而非门槛，已开放动作的控制要求不能豁免——身份归属、真实履约、授权执行、异常处置四项要求不可裁减，但每一项都可以用平台默认配置实现。隐私担忧（47%）接近半数，提示平台化方案必须同时解决数据托管的信任问题，否则成本下降也无法换来采用。

本章所建立的评估框架，最终要形成一条可以运行的关系：明确任务与责任，验证实际履约，在可控边界内授权，再用新的行为证据更新判断。它把前三章讨论的信任问题落到企业和交易对手能够使用的证据上，也为第六章从行业层面观察智能体信用建设与发展状况提供基础。

第六章 智能体发展指数

6.1 基本原理

6.1.1 从任务委托观察智能体发展

智能体经济的发展繁荣，体现为更多任务能够交由智能体完成，更多组织和用户将智能体持续嵌入生产生活，以及不同主体围绕任务交付形成付费、协作与交换关系。本报告构建智能体发展指数，沿着能力形成、连接建设、应用采用、执行与交换的活动链条，观察智能体如何成为经济活动的参与者，以及这种参与的广度和深度如何变化。

这一活动链条包含四个相互衔接的环节。能力与质量决定可供委托的任务范围和执行质量；基础设施使能力能够接入工具、数据、身份与支付网络；应用与生态呈现组织部署、产品采用和开发参与；经济活动进一步呈现持续执行、商业收入和服务交换。四个环节共同构成发展繁荣度的观察基础：任务有执行条件，产品有使用场景，协作有连接网络，运行活动能够形成经济关系。

四个环节在不同月份可以表现出不同节奏。工具接入和开发活动的增长，可以为后续应用扩散创造条件；应用访问的增长，可以带来新的用户和使用场景；企业工作量与支付活动的变化，则进一步呈现任务执行和交换如何展开。将这些变化放入同一指数，并分解各项贡献，可以识别增长所处的环节，理解增长动力如何接续。

6.1.2 以应用广度和使用深度确定评价重心

指数设置能力与质量、基础设施、应用与生态、经济活动四个一级维度，共 36 项指标，一级权重为 2：2：3：3。应用与生态、经济活动合计占 60%，将评价重心放在组织和用户的实际采用、持续执行与经济交换；能力与质量、基础设施合计占 40%，保留对任务可执行性和协作条件的观察。这样的赋权与本报告观测智能体经济发展的目标一致，使技术进步能够通过应用和经济活动进入综合评价。

应用广度通过企业付费与部署、热门应用访问、开发者和开源生态共同观察；使用深度通过 Token 使用、企业工作量和任务委托方式观察；交换关系则通过收入、支付事件、金额及参与活动观察。同一发展过程在多个环节留下不同记录，将其共同纳入框架，有助于判断增长是来自供给扩张、采用扩散，还是持续运行和服务交换。板块内采用固定权重，使各月在稳定尺度上比较，也便于解释每项指标对总指数变化的作用。

6.1.3 将发展观测与 KYA 建设相衔接

随着智能体进入更长的工作流、调用更多工具并承担支付和转委托任务，经济活动对信任的要求也随之具体化：实际执行者代表谁，既往交付是否可靠，本次动作是否处于有效授权之内。前文提出的身份可信、信用可信、权限可信，分

别为行为归属、履约判断与行动边界提供依据。指数观察的活动链条，由此与 KYA 的三维框架形成直接联系。

在这一联系中，指数识别哪些任务、连接和交换正在扩张，为信任基础设施的建设提供需求坐标；KYA 依据具体智能体的身份、履约和权限证据，确定其可以在何种任务、额度和监督条件下获得持续授权。月度观察既帮助读者判断智能体经济的变化方向，也帮助企业和平台识别应优先完善的验证、授权与结算机制，使发展中的任务和交易能够形成可持续的合作关系。

6.2 指标体系

6.2.1 四个维度与 36 项指标

表 6-1　四个维度的经济含义与指标构成

维度与权重	项数	经济含义与代表性指标
A 能力与质量 20%	11	任务执行基础：工具任务、软件工程、网页操作、长任务能力与可靠性。
B 基础设施 20%	8	网络协作条件：工具供给、互操作组件、身份登记与支付接入。
C 应用与生态 30%	9	采用广度与生态建设：企业付费和部署、动态热门应用、开发参与及框架与研究供给。
D 经济活动 30%	8	持续执行与交换：商业收入披露、累计工作量、计算使用、协议支付与任务委托方式。

能力与质量共 11 项，覆盖工具任务、软件工程、计算机和网页操作、企业工具协作、长任务能力，以及安全和事实可靠性。任务完成能力与可靠性共同决定智能体可以接受何种委托，为 KYA 按任务场景识别履约条件提供技术依据。基础设施共 8 项，覆盖工具供给、互操作、身份登记和支付接入，观察智能体发现资源、连接服务和开展协作的条件。

应用与生态共 9 项，覆盖企业付费与生产部署、动态热门应用、开源供给、开发参与、研究和框架分发，呈现采用入口及供给生态的扩展。经济活动共 8 项，观察商业收入、企业执行、计算使用、协议支付和直接任务委托。其 30%总权重分配为商业收入 7.5%，工作量与 Token 使用各 5%，四项支付活动合计 10%，直接任务委托比例 2.5%。支付 SDK 置于基础设施，支付事件、金额与参与活动置于经济活动，分别观察接入建设和交换展开。

表 6-2　全 36 项指标及其总权重

编码	指标及计算对象	总权重 w	计分规则
A01	综合工具任务成功率	2%	相对值
A02	软件工程解决率	2%	相对值
A03	计算机操作成功率	2%	相对值
A04	网页工作流成功率	2%	相对值
A05	企业工具协作成功率	2%	相对值
A06	长任务能力（人类专家等价任务时长）	2%	对数
A07	提示注入下安全效用	1.25%	相对值
A08	固定事实问题正确率	1.25%	相对值
A09	基础模型综合能力	3%	ECI 增量
A10	基础模型安全表现	1.25%	相对值
A11	摘要事实一致性	1.25%	版本分段
B01	工具连接供给（现存目录按最新版本发布日期回溯）	20%÷9	对数
B02	工具供给分散程度	20%÷9	相对值
B03	远程服务实测可用率	20%÷9	中性 100
B04	互操作 SDK 分发强度	20%÷6	衔接后逐包定标
B05	固定平台互操作支持率	20%÷6	中性 100
B06	两链身份登记编号规模	20%÷12	对数
B07	有效验证覆盖率	20%÷12	中性 100
B08	固定支付 SDK 分发强度	20%÷6	对数
C01	企业付费采用（Agentforce 合同披露）	3.75%	对数
C02	企业生产部署采用披露	3.75%	对数

编码	指标及计算对象	总权重 w	计分规则
C03	动态前十 Agent 应用平均访问活跃度	6%	几何均值再对数
C04	开源项目新增供给	2.5%	对数
C05	开发者参与	2.5%	对数
C06	开发与维护活动	2.5%	对数
C07	研究供给	3.75%	对数
C08	固定开发框架分发强度	3.75%	衔接后逐包定标
C09	公众搜索关注	1.5%	相对值
D01	商业收入（Sierra ARR 披露）	7.5%	对数
D02	企业执行（Agentforce 与 Slack 累计 AWU）	5%	对数
D03	Agentic 计算活动	5%	对数
D04	协议支付转账事件强度	2.5%	对数
D05	协议支付金额强度	2.5%	对数
D06	分组付款地址活动强度	2.5%	对数
D07	分组收款地址活动强度	2.5%	对数
D08	API 样本直接任务委托比例	2.5%	相对值

6.2.2 月度观测与动态热门应用

指数按月计算，观测期为 2025 年 10 月至 2026 年 9 月。记第 j 项指标在 t 月的原量为 x。网站访问、软件分发、Token 和协议支付等流量按完整自然月换算为日均；比例和评测结果保留其测量单位；身份登记采用期末规模；企业合同、收入和累计工作量按公开披露月份更新。低频数据在相邻披露之间延续最近值，早期月份以首个可用值回填。B03、B05、B07 三项采用中性标准分 100，合计权重 7.2222%，观察期内增量贡献为零。

热门应用访问分项 C03 采用动态前十样本。每月以 AI 产品榜 AICPB 公开发布的“智能体榜”前 15 条记录为候选，按独立应用归并多个入口，同一应用保留当月访问量最高的入口，再按访问量选出 10 款应用。每月重新选样，使新的采用入口能够及时进入观测。12 个月共形成 120 条入选记录，涉及 19 款不同应用；各月原始榜单列于本章公开资料来源。

对入选应用的日均访问量取等权几何平均。设 T 为当月入选的 10 款应用，V 为应用主入口当月访问量，D 为当月自然日数，则 C03 的月度原量为：

$$X_t = \exp\left(\frac{1}{10}\sum_{i\in T_t}\ln\left(\frac{V_{it}}{D_t}\right)\right)$$

几何平均使每款应用相同比例的访问变化，对合成原量具有相同影响，呈现典型热门应用的平均活跃表现。同步观察访问总量、中位数及入选名单，可以区分头部总量、产品分布和应用更替三个层面的变化，使应用扩散的解释更为具体。

软件分发按固定软件包构建。C08 采用 LangGraph、CrewAI、AutoGen AgentChat 和 Smolagents 四个 Python 包，分别定标后各占 25%；B04 采用 MCP Python、MCP TypeScript 和 A2A TypeScript，分别占 25%、25%和 50%。Python 分发序列以 2026 年 2 月为共同观测月衔接：截至该月保留原序列，此后按实际安装包请求量的同比例变化延展。实际安装包包括.whl、.tar.gz 和.zip 文件，并排除五类镜像安装器。设 O 为原序列，A 为实际安装包日均请求量，p 为软件包，τ 为 2026 年 2 月，则：

$$x_{pt} = O_{p\tau} \times \frac{A_{pt}}{A_{p\tau}} (t > \tau)$$

衔接后的 x 表示同一尺度上的分发强度，逐包计分后再合成篮子。TypeScript 分项按原 npm 固定包序列计算，安装包使用活动由此在连续尺度上进入月度比较。

6.2.3 固定参照与标准化

各项原量具有不同单位，先转换为可合成的标准分 S。一般指标以 2025 年 10—12 月原量的算术均值作为固定参照 c。固定参照确定单项计分尺度，2025 年 10 月确定总指数展示起点。固定参照在后续发布中保持不变，使新月份的变化可以连续比较。

成功率、比例及搜索关注等相对量，采用月值与固定参照之比：

$$S_{jt} = 100 \times \frac{x_{jt}}{c_j}$$

工具数量、软件分发、网站访问、Token、收入、工作量和支付强度等规模量，先作对数变换，再按参照期平均对数值 d 定标：

$$z_{jt} = \ln\left(1 + \frac{x_{jt}}{c_j}\right)$$

$$S_{jt} = 100 \times \frac{z_{jt}}{d_j}$$

其中，d 为 2025 年 10—12 月三个 z 值的均值。对数变换压缩规模量之间的倍数差异，使不同环节的持续增长能够在同一框架中发挥作用。标准分 100 表示达到该项参照水平，标准分高于或低于 100，分别表示高于或低于这一水平。

特殊指标按其计量结构处理。基础模型能力 ECI 采用“100＋当月 ECI－参照 ECI”，参照值为 146.5；摘要事实一致性按评测版本定标，2025 年 10 月旧版置 100，11 月起新版以 11—12 月均值 95.71 为参照；互操作 SDK 和框架分发先在软件包层面定标，再按固定篮子权重合成。

6.2.4 加权合成与贡献分解

设 b 表示四个板块，v 表示某项指标在本板块内的固定权重，板块内权重之和为 1。先加权得到板块得分 B，再按 2∶2∶3∶3 合成 J，最后将 2025 年 10 月总指数置为 100：

$$B_{bt} = \sum_{j \in b} v_{bj} S_{jt}$$

$$J_t = 0.2B_{At} + 0.2B_{Bt} + 0.3B_{Ct} + 0.3B_{Dt}$$

$$I_t = 100 \times \frac{J_t}{J_{2025-10}}; I_{2025-10} = 100$$

2025 年 10 月四个板块得分依次为 96.57099、93.15183、98.22188 和 79.36365，按一级权重合成为 91.22022；2026 年 9 月同法合成为 181.50863。因此，9 月总指数为 100×181.50863÷91.22022＝198.98。指数由 100 升至 198.98，表示这一综合评价水平较基期提高 98.98%。

设 w 为某指标在总指数中的权重，即一级权重与板块内权重的乘积。s、t 分别表示起止月份。某指标的点数贡献等于其标准分变化乘以总权重，再乘以统一的基期换算系数：

$$\Delta I_j(s,t) = \frac{100}{J_{2025-10}} w_j (S_{jt} - S_{js})$$

$$\sum_j \Delta I_j(s,t) = I_t - I_s$$

例如，C03 的固定参照 c 为 180395.52590，d 为 0.692692489。2025 年 10 月与 2026 年 9 月的日均访问几何均值分别为 165155.78502 和 318290.61948，代入对数公式后，标准分分别为 93.83519 和 146.79315。C03 占应用板块 20%，占总指数 6%，因此全期贡献为 100÷91.22022×6%×（146.79315－93.83519）＝3.48 点。将 36 项贡献相加，恰好得到总指数增加的 98.98 点。

环比增速按“当月总指数÷上月总指数－1”计算。单项和板块的点数贡献用于解释增长来自何处，环比增速用于比较各月变化速度。计算采用未四舍五入的数值，表内结果保留两位小数。

6.3 指数结果及呈现

6.3.1 总体走势与月度结果

2025 年 10 月至 2026 年 9 月，智能体发展指数由 100 上升至 198.98，累计增加 98.98 点。2026 年 2 月环比增长 20.76%，形成观察期内最大的月度增幅；7 月、8 月和 9 月环比分别增长 2.37%、2.73%和 4.40%，后段月度增长重新加快。总指数持续上升，增长来源则在计算使用、应用生态、连接建设和企业执行之间逐步接续。

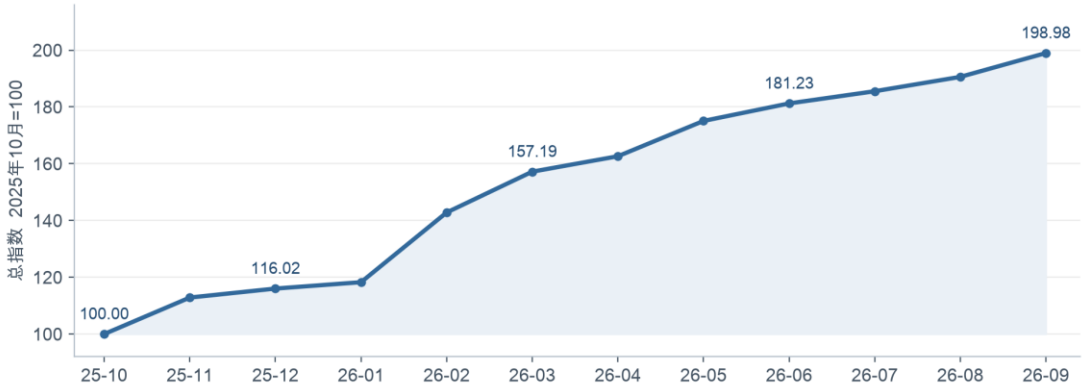


图 6-1　智能体发展指数月度走势

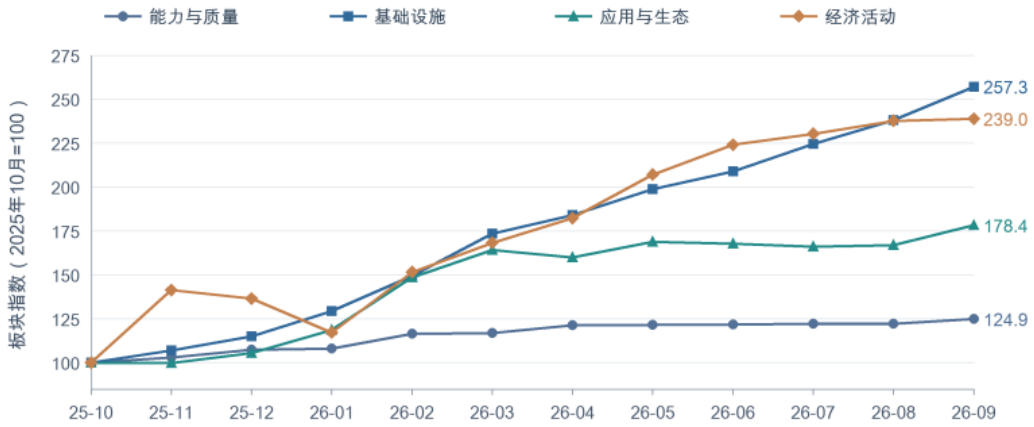


图 6-2　智能体发展指数一级指标月度走势

表 6-3　月度总指数与板块得分

月份	总指数	环比	A 能力	B 基建	C 应用	D 经济
2025-10	100.00	—	96.57	93.15	98.22	79.36
2025-11	112.85	12.85%	99.61	99.66	98.06	112.25
2025-12	116.02	2.81%	103.81	107.19	103.72	108.39
2026-01	118.26	1.93%	104.31	120.55	116.63	93.07
2026-02	142.82	20.76%	112.61	139.30	145.93	120.40
2026-03	157.19	10.06%	112.90	161.70	161.34	133.57
2026-04	162.62	3.45%	117.33	171.52	157.15	144.77
2026-05	175.06	7.65%	117.54	185.29	165.98	164.43
2026-06	181.23	3.52%	117.72	194.77	164.85	177.89
2026-07	185.52	2.37%	118.03	209.23	163.12	182.83
2026-08	190.59	2.73%	118.09	221.91	164.10	188.74
2026-09	198.98	4.40%	120.64	239.63	175.18	189.66

四个板块全期均形成正向贡献。经济活动贡献 36.28 点，基础设施贡献 32.12 点，应用与生态贡献 25.31 点，能力与质量贡献 5.28 点。其中，应用与生态、经济活动合计贡献 61.59 点，占全期净增量 62.2%，与本指数突出采用和经济实现的设计目标相呼应；基础设施成为另一项重要支撑。

单项贡献进一步显示增长的具体来源。Token 贡献 29.82 点，占全期净增量 30.1%；工具连接供给贡献 11.32 点，互操作与支付 SDK 合计贡献 18.61 点。应用与生态内部，开源项目、开发者参与、维护和研究供给四项合计贡献 15.11 点，占该板块净增量 59.7%；动态热门应用与框架分发分别贡献 3.48 点和 3.91 点。执行活动、连接网络与开发供给共同构成观察期内的主要增量。

6.3.2 阶段贡献与增长动力接续

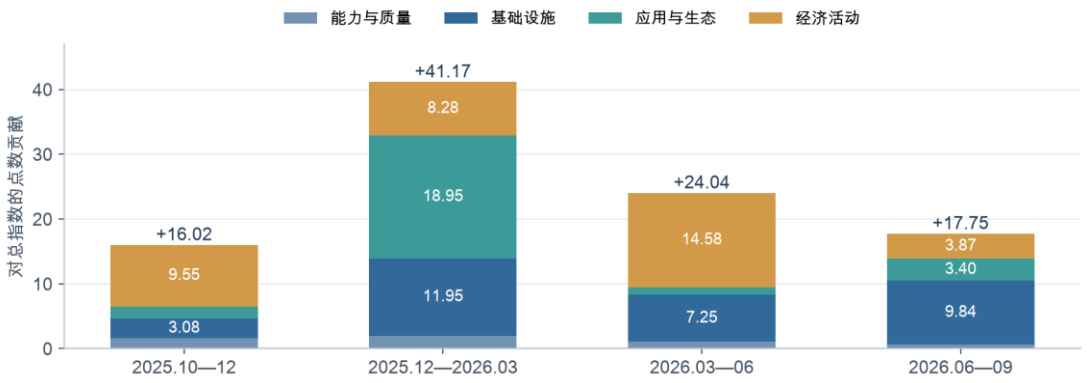


图 6-3　不同阶段的板块点数贡献

表 6-4　阶段端点变化及板块点数贡献

起止月份	总增量	A 能力	B 基建	C 应用	D 经济
2025-10 至 2025-12	16.02	1.59	3.08	1.81	9.55
2025-12 至 2026-03	41.17	1.99	11.95	18.95	8.28
2026-03 至 2026-06	24.04	1.06	7.25	1.15	14.58
2026-06 至 2026-09	17.75	0.64	9.84	3.40	3.87

从阶段端点看，2025 年 10—12 月增加 16.02 点，经济活动贡献 9.55 点，形成早期拉动。2025 年 12 月至 2026 年 3 月增加 41.17 点，是增量最大的一段，应用与生态贡献 18.95 点。2026 年 3—6 月增加 24.04 点，经济活动贡献 14.58 点，增长重心转向持续执行。6—9 月增加 17.75 点，基础设施贡献 9.84 点，占阶段净增量 55.4%，连接建设成为后段的主要支撑。

阶段增量的收窄还伴随着动力分布的变化。2026 年二季度与三季度 Token 以外 35 项的贡献分别为 15.50 点和 15.30 点，基本稳定；两阶段总增量差额中，Token 分项贡献变化占 96.8%。Token 分项 8—9 月按 7 月值延续，其他活动继续更新，后段曲线因而更多体现连接建设、热门应用和企业执行的增长。最大单项占全部正向贡献的比例由二季度 30.7%降至三季度 18.2%，主要增长动力的分布更为均衡。

与此同时，上升指标由二季度 17 项变为三季度 14 项，下降指标由 4 项变为 6 项。总量增长、贡献分散与局部调整同时出现，表明智能体经济正在形成不同环节各有节奏的发展格局。结合总指数、板块贡献和具体活动记录，才能进一步理解哪些环节已形成持续增长，哪些环节正在重组。

6.4 指数解读

6.4.1 执行密度上升使履约成为经济价值的关键环节

观察期内最显著的变化是执行活动的扩张。Token 分项贡献 29.82 点，占经济活动板块净增量的 82.2%；该板块其余 7 项合计贡献 6.46 点。Token 日均使用量由 2025 年 10 月 581.91 亿增至 2026 年 7 月 5.82 万亿，呈现智能体运行与计算使用的快速增加。2026 年二季度，Token、累计工作量与协议支付分别贡献 8.54 点、2.02 点和 4.01 点，形成执行与交换共同拉动的阶段特征。

企业场景使这种变化具有更具体的业务含义。Salesforce 在 5 月披露 Agentforce 与 Slack 累计工作量达到 38 亿 AWU；8 月披露的截至 7 月 31 日财季单季完成 32 亿 AWU，环比增长 97%，累计达到 70 亿 AWU。AWU 将提示处理、推理链与工具调用等执行活动纳入计量，反映智能体持续进入企业系统中的工作负载。与产品访问共同观察，可以看到采用正在从用户进入产品，进一步延伸到业务系统中的连续执行。

执行越密集，企业越需要知道这些活动围绕什么任务展开、以何种成本交付了什么结果。一次业务委托可能包含多轮推理、多次工具调用以及人工接管，经济价值最终要在交付验收和持续使用中体现。KYA 在此承担连接运行活动与履约判断的作用：将任务承诺、过程事件、资源使用和验收结果关联起来，使已有执行经历能够成为下一次委托的信用依据。对企业而言，发展重点随之延伸到稳定交付、单位任务成本和责任清晰的持续工作流。

6.4.2 应用扩散通过产品更替和头部分布变化展开

2026 年一季度应用与生态贡献 18.95 点，成为该阶段最大的板块增量。OpenClaw 主入口月访问量由 1 月 263 万增至 2 月 2700 万，Moltbook 以 2233 万月访问量进入前十；热门应用日均访问几何均值在 2 月升至 32.57 万。按存续应用访问变化与名单更替分解，1—2 月 C03 增加的 36.02 个标准分中，两月均入选应用贡献 21.28 分，样本更替贡献 14.74 分。已有产品扩张和新入口出现共同推动了活跃度上升。

这种进入与更替在 2026 年三季度更加突出。该阶段热门应用分项净贡献 2.495 点，其中逐月名单更替带来 3.126 点，连续入选应用的访问变化抵消 0.631 点。7 月 WorkBuddy、百度搭子与扣子进入前十，9 月 Muse 以 1446 万月访问量入选。对持续更新的产品市场，跟踪当月热门应用，可以及时呈现用户开始在哪些入口组织工作和生活任务。

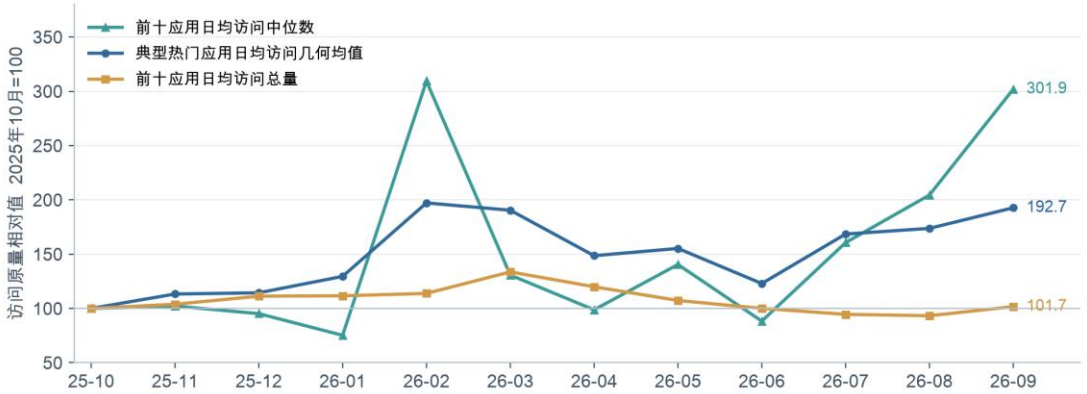


图 6-4　热门应用活跃度与访问分布的变化

头部内部的分布变化提供了另一层解释。全期前十应用日均访问总量增长 1.70%，日均访问中位数由 13.44 万升至 40.57 万，增长 201.94%；纳米 AI 在前十篮子中的访问占比由 79.2%降至 57.9%。典型热门应用的活跃度提升，更多体现为访问向多款产品扩散。2026 年 6—9 月，“AI agent”搜索关注指标下降 40.9%，热门应用日均访问几何均值提高 56.7%，一般概念关注与具体产品使用呈现不同节奏，应用采用的观察需要进一步落到产品和任务层面。

对 KYA 而言，产品快速进入要求建立清晰的初始身份和责任锚点，持续换代则要求保留可核验的履约连续性。新应用需要通过边界明确的任务获得首批信用记录，已有应用需要在模型、工具和权限变化后更新验证。将主体、运行版本与履约事实联系起来，有助于用户在多样化产品中识别适合本次委托的服务，也有助于稳定交付的产品积累可持续的使用关系。

6.4.3 连接网络扩大提高跨主体授权的重要性

2026 年三季度基础设施贡献 9.84 点，工具连接供给、互操作 SDK 和支付 SDK 分别贡献 3.97 点、3.11 点和 1.92 点。全期工具连接供给贡献 11.32 点，互操作与支付 SDK 合计贡献 18.61 点。开发工具和共同规范的建设，与 SDK 分发及应用开发活动相互呼应：OpenAI 发布 AgentKit，将工作流构建、连接与评测整合为开发工具；Linux 基金会成立 Agentic AI Foundation，MCP 等成为创始项目。

身份设施也在扩张。以以太坊与 Base 两链期末最大铸造编号之和衡量的登记编号规模，由 2026 年 2 月 49562 增至 9 月 148527，接近 3 倍。ERC-8004 围绕身份、声誉和验证组织协作机制，为智能体在开放网络中被发现以及交换履约信息提供载体。工具、SDK 与身份设施共同增长，意味着一次任务可以连接更多服务和执行方，协作中的主体关系也更加复杂。

连接网络的经济作用，在于把原本分散的能力组织成可组合的服务。随着编排者、专业子智能体和外部工具共同参与任务，信任判断需要沿委托关系传递：谁发出请求，谁有权转委托，下游可以读取哪些数据、执行哪些动作，异常发生时由谁停止或承接。KYA 因此需要把身份与当前委托绑定，并使数据范围、工具范围、预算和有效时间随授权传递；下游执行保持在上游授权范围内，撤销能够覆盖相关执行链。这是连接增长能够进入企业业务和跨主体协作的重要运行条件。

6.4.4 支付更加高频细碎需要服务计量与结算衔接

支付结构体现了与计算使用不同的变化。全期协议索引日均转账事件由 13.29 万增至 43.54 万，增长 227.6%；日均名义金额由 12.92 万美元降至 3.28 万美元，每条索引事件对应的平均名义金额由约 0.972 美元降至 0.0754 美元。付款地址活动强度由日均 9324.23 降至 3013.40，下降 67.7%。事件强度提高、单次索引事件名义金额下降与付款地址活动收缩同时出现，支付活动在事件强度、金额和参与广度方面呈现分化。

接入建设与交换活动也有各自的节奏。三季度支付 SDK 贡献 1.92 点，四项支付活动合计抵消 1.87 点，两类指标合计增加约 0.05 点；全期四项支付活动合计抵消 1.16 点。支付接口的采用扩大了可开展交换的条件，持续交易还需要服务供给、需求匹配、计量与验收共同形成。x402 V2 对会话复用和按使用量计费等方式的支持，为组织这类服务关系提供了机制。

在高频小额服务中，每次调用都重新进行人工确认，会增加协作成本；按照一次授权无限延伸，又会模糊累计支出边界。KYA 可以把信任安排落实到会话或任务：验证付款身份，预设总预算、频次和工具范围，连续记录调用与交付，并依据验收结算。付款、计量、服务结果与争议处理使用同一任务标识，才能让细碎交换形成清晰的履约关系，也使长期服务提供者积累可以复用的交易信用。

6.4.5 能力拓展需要按任务风险组织实际委托

能力与质量全期贡献 5.28 点，其中长任务能力约占六成。人类专家完成同类任务所需的等价时长由 203.01 分钟增至 1044.78 分钟，达到期初的 5.15 倍，反映可执行任务复杂度的拓展。到三季度，网页工作流成功率约占该阶段能力增量 58.6%，能力进步进一步进入网页操作与既有业务流程。更长任务和更广工具范围，为企业组织复杂委托提供了技术条件。

委托方式则体现出组织选择。D08 所观察 API 样本中的直接任务委托比例由 66.30%变为 58.22%，下降 8.08 个百分点。能力边界扩展与委托方式的变化，提示企业需要进一步解决如何配置自动执行、人工监督和经济权限的问题。复杂任务通常包含多个可分离的步骤，不同步骤的数据敏感性、可撤销性与资金责任各不相同，适宜的委托范围也应随之细化。

KYA 可以把能力进步承接为按场景积累的履约信用，并将其用于差异化授权。对已有稳定履约记录、结果可验收的同类任务，可以扩大执行批量和额度；进入新的业务域或承担更高经济责任时，应增加相应验证、分段授权与人工接管条件。这样，能力评价、履约记录和权限安排形成连续关系，企业能够以可验证的依据逐步扩大委托，智能体的技术潜力也更容易进入稳定的业务流程。

6.5 KYA 建设启示

指数呈现的增长结构，对 KYA 建设提出了明确的优先序。执行活动增加，需要形成可用于再次委托的履约信用；产品更替和主体扩张，需要维护身份与责任的连续性；工具和服务连接扩大，需要把授权落实到委托链；交换更加细碎，需

要贯通计量、验收和结算。身份可信、信用可信、权限可信应围绕同一任务共同运行，使每次委托都有可核验的主体、履约依据和行动边界。

6.5.1 以任务履约积累可复用的信用

Token 与企业工作量的增长，使任务级履约证据成为 KYA 建设的首要基础。企业和平台应围绕具体任务关联委托目标、承诺结果、执行环境、资源使用、人工接管和验收记录。每次工具调用可以回溯到所属任务，每次验收可以对应到实际执行主体，执行中的授权变更和异常处置也应保留。任务级记录由此把运行活动组织成可核验的履约历史。

履约信用应按任务类型、运行环境和经济权限积累。在身份、履约与权限关键门槛持续满足的条件下，稳定完成同类任务可以支持后续扩大批量、额度或预授权范围；跨入新的任务域时，则补充相应的验证和监督。记录还应区分执行失误、用户撤回、外部服务故障与人工接管，使履约判断能够准确归因。信用在这样的使用中产生经济作用：减少重复审核，支持差异化授权，并为争议处理提供共同事实。

6.5.2 让身份贯穿产品更替与委托关系

热门应用更替和身份登记扩张，要求 KYA 同时支持新进入者建立记录、持续运行者维护记录。身份档案应关联运营主体、实际运行实例、模型和工具版本、委托人及当前状态，使交易对手能够核验本次执行者是谁、代表谁，以及既往记录与当前执行之间的关系。版本和权限发生重大变化时，应更新档案并重新验证受影响的任务范围。

新进入者可以通过企业或用户责任锚定、小额预算和边界清晰的任务建立初始信用，完成经核验的履约后逐步扩大委托。跨平台互认应传递可验证的身份和履约事实，并保留其任务条件与有效时间，供使用方按自身场景作出判断。这样的安排既为新产品保留进入市场的通道，也使稳定交付的服务能够在产品升级和平台迁移中保留信用积累。

6.5.3 将权限控制落实到完整委托链

工具连接和互操作扩张，使授权对象从单一应用延伸到编排者、子智能体和外部服务。KYA 应将任务目标、数据范围、可用工具、资金预算、有效时间和转委托条件写入可验证授权。每一层接收者核验委托来源，下游权限在上游范围内收敛；权限收回时，相关子智能体和工具凭证同步失效。责任关系和授权关系由此能够沿行动链追溯。

持续工作流可以依据动作风险配置监督：低风险、可撤销的动作纳入预授权，资金转移和重要数据操作设置额度、频次与复核条件。预算接近上限、异常调用或关键版本变更，可以触发限流、降权和人工接管。身份、履约信用与任务风险共同决定可执行范围，使企业能够在清晰边界内减少重复确认，提高连续委托的效率。

6.5.4 贯通服务计量、支付与履约结算

面对高频、小额和持续运行的交换，KYA 应把付款主体、预算凭证、调用记录、计量结果、交付验收与资金流关联到同一任务。付款前验证身份与支出权限，执行中控制累计费用和重复调用，交付后记录验收、退款及争议结果。对于按调用、按会话或按持续任务提供的服务，计费单元应与实际交付相对应，使资金记录同时成为可核验的履约证据。

会话和任务预算可以将多次调用置于同一授权范围内管理，并设置异常频次、重复扣款和无效交付的阻断条件；金额较大、周期较长的任务，可以采用里程碑验收和分段释放额度。计量、支付与结果相互对应，有助于降低人工核对和争议处理成本，使接入能力进一步转化为可重复开展的服务交易。

6.5.5 以发展结构确定建设优先序并持续验证

企业可以优先选择工作量稳定、结果可验收且责任清晰的场景建设任务级信用档案，再依据履约积累扩大授权。平台可以围绕新应用进入、跨工具执行和持续服务结算，优先完善身份连续性、委托链核验与预算管理。跨平台协作则可以从共同可核验的身份和任务事件入手，保留事实来源、发生条件和版本信息，使不同参与方能够复用证据并按各自业务场景判断。

月度指数为上述优先序提供持续观察依据。连接建设贡献上升时，关注跨主体授权和撤销机制；应用更替加快时，关注初始信用和版本验证；执行活动扩大时，关注验收、成本和履约归因；支付更加细碎时，关注累计预算和自动结算。KYA 建设的进展则可以在具体场景中通过重复核验时间、同类任务持续履约、授权调整以及争议处理效率来检验。发展观测与任务实践相互衔接，使身份能够承接行为、履约能够形成信用、信用与风险能够进入授权，再由新的执行结果持续更新判断。

指数公开资料来源

[1] PyPI，下载统计规则与安装包请求。

[2] ClickHouse，ClickPy 公开软件分发数据。

[3] OpenAI，Introducing AgentKit。

[4] Linux Foundation，Agentic AI Foundation 成立公告。

[5] Salesforce，FY26 Q3 财报。

[6] Sierra，100M ARR 里程碑。

[7] Sierra，Year two in review。

[8] Salesforce，FY27 Q1 财报。

[9] Salesforce，Agentic Work Unit 定义。

[10] 腾讯，WorkBuddy 产品介绍。

[11] 百度智能云，百度搭子 Dumate 产品文档。

[12] Meta，Muse 个人智能体介绍。

[13] Salesforce，FY27 Q2 财报。

[14] OECD.AI，OpenRouter 数据与 Agentic 计算使用。

[15] Ethereum Improvement Proposals，ERC-8004。

[16] ERC-8004 官方仓库，Registry contracts。

[17] x402scan，协议支付索引。

[18] x402 协议团队，Introducing x402 V2。

[19] MCP Registry，工具登记目录。

[20] npm，x402 固定软件包与分发接口。

[21] AICPB，2025-10 月智能体榜。

[22] AICPB，2025-11 月智能体榜。

[23] AICPB，2025-12 月智能体榜。

[24] AICPB，2026-01 月智能体榜。

[25] AICPB，2026-02 月智能体榜。

[26] AICPB，2026-03 月智能体榜。

[27] AICPB，2026-04 月智能体榜。

[28] AICPB，2026-05 月智能体榜。

[29] AICPB，2026-06 月智能体榜。

[30] AICPB，2026-07 月智能体榜。

[31] AICPB，2026-08 月智能体榜。

[32] AICPB，2026-09 月智能体榜。

[33] A01，GAIA 公开评测数据。

[34] A02，SWE-bench 榜单数据。

[35] A03，OSWorld Verified 评测数据。

[36] A04，网页工作流评测公开数据。

[37] A05，τ^2-bench 官方提交与榜单。

[38] A06，METR 任务时间跨度数据。

[39] A07，AgentDojo 评测结果。

[40] A08，Epoch 公开基准数据（A08、A09）。

[41] A10，Stanford HELM Safety。

[42] A11，Vectara 摘要事实一致性榜单。

[43] C02，Agentforce 生产使用客户披露。

[44] C04，GitHub 开源项目公开检索。

[45] C05，GitHub 开发者参与记录。

[46] C06，GitHub 开发维护记录。

[47] C07，arXiv 研究供给公开检索。

[48] C09，Google Trends 的 AI agent 搜索数据。

[49] D08，Anthropic Economic Index 公开 API 样本。

第七章 展望：从智能体经济到信任资本时代

人工智能正在经历一次重要转变，机器正在成为经济体系中一种新的生产性力量。然而，技术能力并不等于经济能力。一个智能体即使能够完成复杂任务，如果无法证明自身身份、授权来源、执行可靠性和责任归属，就难以获得持续参与商业交易的资格。

这正是 KYA 的历史机遇。它不仅关系到智能体能否安全运行，更关系到人工智能创造的生产能力，能否转化为真实的经济价值。

智能体经济的发展将经历从能力形成到经济协作的演进；智能体产业将由技术产品竞争走向经济基础设施竞争；KYA 则可能推动一种新的经济资源——智能体信任资本——的形成与积累。

7.1 智能体经济的四个阶段：从执行任务到创造经济

智能体经济不会因为人工智能模型足够强大就自动形成。模型能力、任务执行、市场交易与经济组织之间，存在不同的发展门槛。从技术能力与经济角色的演进来看，可以将智能体经济划分为四个阶段。

第一阶段：能力智能体——从人类工具到智能助手。

这一阶段的核心是让机器具备理解、分析、推理和内容生成能力。智能体能够协助人类完成信息检索、文本创作、程序编写、数据分析和辅助决策等工作，但主要停留在人类直接操作和监督的范围内。它的经济价值主要体现为提高个人生产率，降低知识劳动和信息处理成本。

这一阶段，市场竞争的核心是模型能力、算力、算法和应用体验。企业主要销售模型服务、软件工具和计算资源。其本质是人工智能能力的形成与普及。

第二阶段：执行智能体——从智能助手到数字劳动力。

随着工具调用、任务规划、工作流编排和多智能体协作技术的发展，智能体开始能够独立执行具有明确目标的复杂任务。

这一阶段，智能体从辅助工具逐渐转变为生产过程中的执行力量。企业开始批量部署智能体，形成新的数字劳动力体系。

市场竞争随之从模型能力扩展到任务完成率、运行成本、系统可靠性和业务价值。智能体工厂、企业智能体平台及组织管理系统开始成为重要的产业基础设施。

但第二阶段也暴露出一个根本性问题：智能体可以执行任务，并不意味着市场愿意信任它、接受它代表企业行动。

截至 2026 年 9 月，智能体经济总体上正处于由第一阶段向第二阶段加速演进的时期，部分前沿场景已经开始探索第三阶段。

第三阶段：交易智能体——从执行任务到参与市场交易。

这是智能体经济最重要的转折之一。在这一阶段，智能体不仅执行企业内部任务，还能够在不同企业、平台和服务机构之间进行商业协作。

采购智能体可以寻找供应商、比较报价并发起订单；物流智能体可以采购运输服务；金融智能体可以在授权范围内进行支付、结算和风险管理；不同专业智能体可以通过服务交易共同完成复杂项目。

智能体由此开始成为市场交易的执行参与者。它们不必拥有独立法律人格，却可以在明确的授权、契约和责任框架下，代表人类或企业开展经济活动。这种变化可能催生全新的智能体服务市场、机器间交易网络，以及更加细分的数字化专业分工。

但跨主体交易必须建立在可信机制之上。交易对手需要知道智能体代表谁、获得何种授权、具有怎样的履约能力，以及出现损失时如何追究责任。

因此，KYA 是智能体从内部执行工具走向市场交易参与者的关键基础设施之一。

第四阶段：经济智能体——从参与交易到持续组织经济活动。

当智能体能够持续执行任务、进行跨平台交易，并在复杂环境中开展协作时，智能体经济可能进入更深层的发展阶段。智能体不再只是完成某项孤立任务，而是参与企业经营、资源调度、供应链协同和产业网络运行。

一家企业可能由少量人类管理者协调大量专业智能体完成经营活动；不同企业的智能体可能围绕订单、数据、资金和生产资源形成持续的经济联系；能源设备、工业机器人和智能汽车等物理系统，也可能通过智能体参与资源配置和商业交易。

经济组织的边界、企业的最小有效规模以及专业化分工方式，都可能随之变化。

这一阶段的核心，不只是智能体能够创造多少价值，而是如何在法律、市场和技术规则下，组织大量智能体持续创造价值。

KYA 也将从单个智能体的身份与行为管理，逐步延伸至跨平台互信、动态信用评价、风险定价和责任治理，成为智能体经济制度体系的重要组成部分。

四个阶段并非严格依次出现，未来也可能长期并存。但其演进方向是清晰的：从机器具备能力，到机器执行任务，再到机器参与交易，最终走向机器与人共同组织经济活动。

这一过程意味着，智能体经济的市场边界有机会从人工智能软件与算力服务，逐渐扩展到企业经营、商业交易、产业协作和实体经济运行。

互联网扩大了人类连接信息与市场的范围，智能体经济则可能进一步扩大人类能够有效组织、委托和执行经济活动的范围。

决定这一进程能走多远的，不仅是人工智能技术，还有社会能够建立多大范围的机器信任机制。

7.2 时代级机遇：从制造智能体到建设智能体经济

历史上的重大技术革命，往往在技术产品之外创造更为广阔的商业生态。

工业革命最初的机遇是制造机器，但最终改变经济的是机器组织起来的现代工业生产体系。电气化的价值不仅来自电动机，还来自工厂组织方式的重构。计算机和互联网革命不仅创造了硬件与软件产业，更催生了企业管理系统、云计算、电子商务和全球数字支付网络。

智能体革命也可能遵循类似的演进逻辑。

随着 KYA 等信任基础设施逐步成熟，智能体的经济活动边界有望不断扩大，产业竞争也将从制造智能体，进一步走向组织、信任和连接智能体。

对于处在智能体生态中的前沿企业，以及新一代智能体创业者而言，至少存在四项具有时代意义的商业机遇。

7.2.1 生产智能体：从软件工厂到数字劳动力工厂

工业革命实现了机器生产能力的规模化复制，智能体革命则可能实现数字劳动能力的规模化供给。

未来，企业购买的不只是软件功能，而是能够完成具体工作、具备明确质量标准和持续服务能力的数字劳动力。

财务、采购、研发、营销、客户服务、风险管理等工作，可以由不同的专业智能体承担。智能体工厂因此有机会从提供开发工具，升级为生产、认证、部署和运营数字劳动力的平台。

商业模式也可能从软件许可、模型调用和项目交付，逐渐延伸到按任务、按数字岗位、按实际经营成果收费。

KYA 在这一过程中提供身份、能力验证、授权管理和可靠性保障，使数字劳动能力能够被企业更加广泛地采用。

这一机遇的本质，是从出售软件工具，走向出售生产能力。

7.2.2 组织智能体：从企业管理软件到智能体经济操作系统

电气化真正释放生产率潜力，不只是因为工厂换上了电动机，更因为机器布局、生产流程和企业组织方式发生了改变。

1990 年，经济学家、经济史研究者 Paul A. David 在《发电机与计算机：对现代生产率悖论的历史考察》中，以电气化为例指出，通用技术从发明到广泛

扩散并带来可测量的生产率收益，可能需要数十年；要释放其潜力，还需配套的技术创新，以及工厂布局、生产流程和组织方式的调整。

这意味着，通用技术创新与生产率收益之间可能存在的漫长时间差，以及组织变革的重要性。智能体经济可能出现类似现象。单个智能体能力的提高，未必自动带来企业整体效率的提高。如何组织众多智能体协同工作，将成为新的管理命题。

一家企业可能拥有成千上万个智能体，分别承担不同岗位和业务任务。任务分配、资源调度、权限管理、绩效评价、风险控制和跨系统协作，都需要新的基础设施。

这将催生超越传统工作流软件的智能体组织操作系统。其价值不仅在于管理智能体，更在于持续优化整个组织的生产效率，并使企业能够围绕数字劳动力重新设计业务流程。

这一机遇的本质，是从管理软件和业务流程，走向组织与配置生产力。

7.2.3 信任智能体：从 KYA 认证到智能体信用基础设施

现代金融的发展史证明，信用基础设施能够创造巨大的经济价值。

征信、信用评级、支付授权和风险管理体系，使陌生主体之间的大规模金融交易成为可能。FICO、穆迪、标普等机构已经证明，身份与信用信息的标准化、专业化服务可以形成持续的商业市场，其影响范围又远超这些机构自身的营业收入。

智能体经济可能产生类似的需求。随着智能体跨企业、跨平台执行任务，身份识别、授权验证、行为监测和履约评价将逐渐从企业内部管理功能，演变为市场交易的共同需求。

KYA 有机会形成一个跨主体的信任服务网络，使智能体能够携带可信身份、有效授权和必要的履约证明参与经济活动。

更深远的是，随着可验证行为记录的积累，智能体的可靠性可能逐渐形成具有持续经济价值的信任资源，为交易准入、风险管理和差异化定价提供依据。KYA 服务商由此有机会从安全技术供应商，升级为智能体经济的信任基础设施提供者。

这一机遇的本质，是从降低交易风险，走向培育信任资本、提高人工智能资本的经济价值。

7.2.4 连接智能体：从产业数字化平台到智能体经济网络

互联网和银行卡网络的发展提供了另一项重要启示：连接经济活动的平台，并不需要直接生产所有商品或服务，却可以成为大规模商业交易的重要基础设施。

Visa、Mastercard 等支付网络每年支持数万亿美元规模的交易，而网络运营者通过提供连接、规则、技术和风险管理服务获得持续收入。

未来，智能体之间也可能形成庞大的服务与交易网络。

采购智能体购买物流服务，营销智能体采购内容生产服务，能源智能体协调储能资源，工业设备通过智能体订购维护服务。不同企业和平台提供的智能体，可以在共同认可的规则下进行协作。

这将催生新的智能体服务发现、交易匹配、支付结算、履约保障和风险管理基础设施。

类似 AI 技术服务机构，可能从向企业提供数字化产品，进一步转向为跨企业的智能体经济活动提供持续服务。其商业增长不再只取决于出售多少软件、生产多少智能体，还可能受益于所服务的智能体数量、真实经济任务和跨平台交易规模。

这一机遇的本质，是从服务企业的数字化转型，走向服务智能体经济网络的运行。

四项机遇共同构成一条新的产业价值链：生产智能体创造生产力，组织智能体配置生产力，信任智能体为生产力进入市场提供可信条件，连接智能体则扩大生产力的经济活动范围。

对企业而言，它们是四个潜在的商业市场；对整个经济而言，它们可能构成一套相互支撑的新基础设施。

未来最有竞争力的企业，可能不是拥有最多智能体的企业，而是能够以合理成本，将最多可信智能体连接到真实经济活动中的企业。

智能体革命最初的商业机遇是制造智能体，更深远的机遇则是建设智能体经济。

7.3 信任资本：人工智能时代的资本形成与经济增长

四阶段演进揭示了智能体经济的发展方向，四大商业机遇展示了新的产业价值空间。但 KYA 的经济意义还可以进一步深化。

一个根本性问题是：为什么两个技术能力相近的智能体，可能具有截然不同的经济价值？

假设两个智能体拥有相同的模型能力、运行效率和任务完成水平。其中一个能够证明身份和授权来源，拥有持续可靠的履约记录、完善的治理机制和明确的责任保障；另一个则缺乏这些条件。

对于需要委托资金、数据和重要经营任务的企业而言，两者的实际使用价值显然不同。

这种差异不完全来自技术能力，而来自市场对其可信程度的判断。

由此，我们提出一个值得深入研究的经济学命题：智能体经济可能形成一种新的生产性资源——智能体信任资本（Agent Trust Capital）。

7.3.1 信任为什么能够成为资本？

信任的经济价值并不是新发现。从科斯的交易成本理论、诺斯的制度经济学，到社会资本与金融发展的实证研究，经济学长期关注信任、制度和合作机制如何影响经济效率。

1997 年，Stephen Knack 和 Philip Keefer 基于 29 个市场经济体的数据发现，人际信任与公民合作规范同经济增长和投资表现相关，并讨论了信任降低经济交往中防范与履约成本的可能机制。2004 年，Luigi Guiso、Paola Sapienza 和 Luigi Zingales 利用意大利地区与家庭数据发现，社会资本较高地区的家庭更倾向于使用支票、配置较多财富于股票，并更容易获得机构信贷。两项研究从跨国宏观表现和家庭金融选择两个层面，显示信任及相关社会规范与经济活动、投资配置和金融合约可得性密切相关。

人类社会中的信任可以通过声誉、履约、制度和长期合作不断积累。机器经济同样需要建立能够支持持续交易的信任机制，但其基础与人类社会存在重要区别。

智能体无法仅凭社会道德和传统人际关系获得市场信任。它需要可验证的身份、明确的授权、可靠的行为记录、有效的治理机制，以及其背后法律主体的责任保障。

这些要素能够通过持续投入、真实履约和制度建设逐步形成，并影响智能体未来的经济活动机会。

因此，我们可以将智能体信任资本定义为：

智能体信任资本（Agent Trust Capital），是经济主体通过可信治理投入和智能体持续、可验证的履约活动所形成的，能够降低交易成本、扩大经济活动范围、提高资本使用效率并持续创造未来经济收益的无形资本。

智能体信任资本的本质，是将机器的可信能力转化为持续创造经济收益的资本能力。信任资本可以积累，也可能因违规、失信、安全事故、授权失效和技术变化而贬值。

这一概念仍需通过理论和实证研究检验，但为理解 KYA 的经济价值提供了新的视角。

7.3.2 KYA 如何推动智能体信任资本形成？

KYA 本身并不等于智能体信任资本。KYA 是一套信任生产、验证与治理机制，信任资本则是长期可信行为与制度安排形成的经济资源。

通过 KYA，智能体的身份、授权、能力、运行状态和履约记录有机会得到持续验证。随着可信记录积累，交易对手可以减少重复审查，企业可以在风险允许的范围内扩大智能体的任务权限，金融与保险机构也可能获得更加可靠的风险信息。

信任由此不再只是一次交易中的主观判断，而可能逐步形成能够持续支持经济活动的资源。

如果大量智能体能够在不同平台之间携带可验证的身份、授权和履约证明，信任资本就可能突破单一企业的边界，成为更广泛市场协作的基础。

这一过程将使 KYA 从身份认证与安全管理工具，逐步发展为智能体信用体系的重要组成部分。

7.3.3 智能体信任资本如何影响资本形成与资本价值?

人工智能时代的资本积累，首先表现为算力、算法、数据、软件及相关基础设施的持续投入。但技术资本并不自动产生经济回报。只有进入真实生产、服务和交易过程，技术能力才能转化为有效产出。

KYA 可能通过提高智能体的可信授权程度和市场接受程度，使部分原本无法执行的经济任务成为可能。这种作用至少表现为三个方面。

第一，激活技术资本。使已经形成的 AI 能力进入更多高价值业务场景，提高技术资本的有效利用程度。

第二，提高资本回报。通过降低验证成本、交易风险和治理损失，改善 AI 项目的风险调整后收益。

第三，促进资本形成。当企业能够更加可靠地控制智能体的行为风险、预测经营收益时，一部分原先难以开展的投资项目可能具备商业可行性，从而吸引新的资本投入。

从这一角度看，信任资本与技术资本可能具有重要的互补关系。技术资本决定智能体能够提供什么能力，信任资本则影响这些能力在多大程度上能够被市场接受、授权和使用。人工智能创造潜在生产力，信任资本有机会提高潜在生产力转化为现实经济价值的效率。

7.3.4 从资本价值走向经济增长与社会福利

智能体信任资本的价值不会止步于单个企业的投资回报。当更多智能体能够可信地参与生产和交易时，经济体系可能出现更广泛的变化。

首先，企业可以将更多重复性和标准化任务交给智能体，提高劳动与资本的协同效率，促进生产率增长。

其次，可信的跨平台协作可以降低专业服务的交易成本，扩大市场范围，推动更加细致的产业分工。

再次，中小企业和创业者能够以较低成本获得专业化数字劳动能力，减少组织固定成本，创造新的产品、服务和企业。

此外，身份验证、行为审计、信用评价、风险管理和智能体运营等新兴市场，也可能形成新的产业需求和就业岗位。

这些变化有机会推动经济增长，但不能简单推断智能体普及必然带来就业总量增长。AI 同样可能替代部分工作，改变工资结构，并使生产率收益在资本与劳动之间重新分配。

因此，KYA 的社会经济价值，不仅取决于其能否提高资本回报，还取决于它能否促进公平竞争、降低中小企业准入成本、减少交易损失，并使技术进步带来的收益得到更广泛的分享。

衡量 KYA 的价值，不仅要看 KYA 产业自身创造了多少收入，更应关注它帮助整个经济体系释放了多少生产力、降低了多少成本、创造了多少净产出与社会福利。

7.3.5 从技术革命到制度创新

回顾历史，重大技术革命都不会停留在生产工具的层面。工业革命创造了机械化生产能力，也推动了现代工厂、企业组织和金融制度的发展。互联网改变了信息传播方式，也催生了电子商务、数字身份、在线支付和平台治理体系。

每一次通用技术革命，都需要与之相适应的制度和组织创新，才能充分释放经济潜力。智能体经济也不例外。人工智能正在让机器获得越来越强的自主执行能力，但经济体系不仅需要能够工作的机器，更需要能够被识别、被授权、被信任和被追责的机器。

KYA 的历史价值，正在于为这种新的生产力建立可信的经济活动条件。它既是技术治理问题，也是制度创新问题；既涉及商业市场的形成，也涉及资本积累、资源配置和社会福利的改善。

从长期看，KYA 可能推动信任从分散的身份认证、声誉判断与风险控制机制，逐步演化为一种可以持续验证、积累和利用的经济资源。

这一过程能否形成成熟、独立的智能体信任资本体系，仍取决于技术可靠性、开放标准、法律制度与市场实践。但其发展方向已经揭示出一个重要命题：人工智能时代的经济竞争，将不仅是模型、算力和算法之间的竞争，也将日益体现为生产能力的可信程度、组织效率以及经济协作范围之间的竞争。

如果说人工智能赋予机器创造价值的能力，那么 KYA 所代表的信任体系，将影响机器在多大程度上能够被允许、被接受并被持续委托去创造价值。智能体经济的未来，不仅取决于机器有多聪明，还取决于人类社会能够建立多大范围的机器信任。

从智能体能力的形成，到智能体经济网络的扩张，再到信任资本的积累，KYA 可能成为连接技术进步与经济增长的重要桥梁。——这正是 KYA 超越安全治理、走向经济基础设施的时代价值。

7.4 如何创造和积累智能体信任资本

在智能体信任资本的创造和积累方面，互联网平台、基础模型企业、云计算与 AI 基础设施企业，无疑具有重要优势。互联网平台拥有用户关系、数字身份

和服务生态，有机会将既有的用户信任延伸为人与智能体之间的可信授权关系；基础模型企业掌握智能体能力的技术源头，有机会将模型性能、安全性和可靠性转化为可验证的能力证明；云计算和 AI 基础设施企业掌握智能体部署与运行环境，有机会成为机器身份、可信执行和行为治理的重要基础设施。

但是，智能体信任资本的积累，对于广大企业乃至个人，都有潜在的巨大机遇。本报告的研究揭示了一个重要趋势：智能体信任资本不属于某个特定行业，也不会被某一种技术或商业平台独占。不同类型的企业，都可能凭借自身独特的资源和能力，参与信任资本的形成、积累与价值实现。以下五类企业尤其值得关注。

第一类：掌握智能体开发、运行与组织平台的企业。

智能体平台不仅可以生产数字劳动力，还可能成为信任资本规模化形成的重要基础设施。以蚂蚁数科为代表的智能体工厂和组织平台，能够将模型、工具、行业知识及业务流程整合起来，形成可持续运行的智能体体系。

这类企业的独特优势，在于可以将可信身份、授权管理、能力测试、行为审计和风险控制嵌入智能体的整个生命周期。

当智能体从开发、部署到运行和升级，都具有连续的身份与可信记录时，信任资本就有机会从单个项目的经验，转化为可以持续积累和复用的经济资源。

因此，智能体平台应当把 KYA 视为核心产品能力，而不是附加的安全功能；同时通过开放接口和标准化凭证，使第三方智能体能够参与可信协作。其战略机遇，是从制造智能体，升级为生产、组织和治理可信数字生产力的基础设施提供者。

第二类：掌握支付、交易和商业网络的企业。

信任资本的经济价值，最终需要在真实交易中得到实现。以 Stripe 为代表的支付与商业基础设施企业，拥有连接商户、消费者、支付机构和交易网络的能力。当智能体开始代表人类或企业进行采购、支付和服务交易时，身份验证、有效授权、支付控制、欺诈识别和责任追溯将成为商业网络运行的重要条件。

这类企业的优势，在于能够直接将可信验证与交易准入、支付授权及风险管理联系起来，使信任成为促进商业活动的实际机制。

未来，支付与交易平台有机会进一步提供面向智能体的可信交易服务、动态风险定价和跨平台履约保障。其商业价值不仅来自验证服务本身，更可能来自可信交易范围扩大所带来的持续服务需求。其战略机遇，是从连接资金与交易，升级为连接可信智能体及其经济活动的商业网络。

第三类：拥有真实产业场景和持续履约数据的企业。

真实、持续、可验证的经济活动，是信任资本最重要的来源。金融、能源、教育、制造、医疗、物流、供应链和专业服务等行业，正在成为智能体进入真实生产活动的重要场景。这类企业能够持续形成真实的任务执行、交易履约、服务质量和风险记录。

比如，以林洋能源所代表的能源智能体应用、以英和智临所的代表的教育智能体应用，以及麦迪卫康等企业探索的医疗专业服务场景为例，行业智能体的价值取决于它能否在复杂业务环境中持续、可靠地完成任务。

对于这类企业，建议把智能体应用与可信数据积累同步建设，将任务完成率、服务质量、授权合规和异常处理等记录转化为可验证的履约证明。谁能率先将行业经验与真实履约数据转化为持续积累的可信记录，谁就有机会形成难以被通用模型简单复制的竞争优势。

第四类：拥有信用评价、风险管理和专业认证能力的企业。

传统信用与风控能力，如果能与新技术共同迭代，就可能在智能体经济中获得新的市场空间。芝麻信用、冰鉴科技等所代表的信用科技和风险管理机构，拥有身份识别、数据分析、风险建模与信用评价等方面的经验。

智能体经济将创造新的评价对象和风险维度。未来的信用体系不仅需要识别智能体的责任主体，还需要评价其任务可靠性、授权合规性、行为稳定性和历史履约表现。这为信用与风控企业提供了从人类及企业信用服务，向智能体可信评价和动态风险服务延伸的机会。

如本报告前述研究，智能体信用不能简单复制传统个人征信模型。不同任务的风险差异、模型更新、身份变化以及责任主体关联，都要求建立新的评价方法。

这类企业可重点发展场景化信用评价、持续风险监测、独立能力认证和可信证明服务，并探索跨平台的信用互认机制。其战略机遇，是从评价传统经济主体的信用，走向参与构建智能体经济的信用基础设施。

第五类：掌握专业知识、行业标准和独立治理能力的企业。

智能体经济越复杂，专业化和独立可信服务就越重要。教育、医疗、法律、金融、科研等专业领域的智能体，需要满足不同于通用任务的质量标准和责任要求。以 GCSA 等专业服务创新企业所代表的探索方向而言，深厚的专业知识、可验证的服务质量和清晰的责任机制，可能成为重要的信任资本来源。

与此同时，独立测试机构、安全审计企业、数字身份服务商、保险机构，以及行业联盟和标准组织，也有机会成为信任资本体系的重要贡献者。它们可以为市场提供第三方能力评价、身份与授权凭证、风险保障及争议处理机制，降低不同平台之间的信任成本。

对于这些企业和机构，建议围绕专业标准、独立验证、开放互认与风险保障形成新的服务体系。其战略机遇，是将专业知识和治理能力转化为智能体经济普遍需要的信任服务。

上述五类企业分别拥有不同的信任资源，它们与互联网平台、基础模型及 AI 基础设施企业相互补充，激活千行百业的智能体，共同构成智能体信任资本的形成、积累、验证和价值实现体系。

基于上述研究，我们对企业界提出四项建议：

第一，将可信能力纳入企业的核心竞争战略。运用智能体的企业不应仅仅比较模型性能、运行成本和应用数量，还应将身份可信度、授权控制、履约表现和风险治理纳入产品设计与经营评价。信任资本需要通过长期可靠的经济活动积累，而不能仅靠品牌宣传或一次性认证获得。

第二，将真实履约转化为可验证、可复用的信用资源。企业应在合法合规的前提下，建立智能体运行与履约记录，推动可信证明、动态评价和跨平台互认，使过去的可靠表现能够降低未来的交易成本。

第三，以高价值经济活动检验 KYA 的商业价值。KYA 不应停留在身份注册和合规检查层面，而应与采购、支付、产业运营、专业服务等真实业务结合。企业需重点衡量 KYA 带来的新增交易机会、风险损失下降、任务授权扩大和资本回报改善，而非单纯追求认证数量。

第四，以开放协作构建信任生态。智能体开发者、产业企业、金融机构、信用服务商及第三方认证机构，应共同推动身份、授权、行为和信用信息的标准化与互操作。在保护隐私、维护公平竞争和明确责任边界的前提下，降低智能体跨平台、跨企业参与经济活动的门槛。

最后，需要说明的是，这里所说的企业界，已经不仅仅是传统的企业界，也包含正在蓬勃成长的“一人公司”和个人经营者，智能体正是它们的员工乃至合伙人。未来，最具信任资本优势的企业，未必是历史最悠久、规模最大或模型能力最强的企业，而是那些能够持续创造可信行为、积累可信记录、推动跨主体互信，并将信任转化为实际经济价值的企业。

KYA 不仅意味着新的治理责任，更意味着新的商业机遇。

参考文献

[1] McKinsey & Company. The automation curve in agentic commerce. 2026；并参见 The agentic commerce opportunity, 2025.

[2] Google Developers. Developer's Guide to AI Agent Protocols: Agent Payments Protocol (AP2). 2026； Google. Donating Agent Payments Protocol to the FIDO Alliance. 2026-04-28.

[3] Mastercard. Mastercard Agent Pay: secure, scalable and trusted agentic AI； Mastercard unveils Agent Pay. 2025-04-29.

[4] Visa. Trusted Agent Protocol, Visa Developer Center； Visa Intelligent Commerce. 2025–2026.

[5] Stripe. Agentic Commerce；Shared Payment Tokens documentation；Introducing our agentic commerce solutions. 2025–2026.

[6] Coinbase Developer Platform. x402 Documentation: Welcome to x402； How x402 Works. 2025–2026.

[7] Ethereum Improvement Proposals. ERC-8004: Trustless Agents. Created 2025-08-13.

[8] 支付宝 AI 支付开放平台. 支付宝 AI 付、A2A 交易与按量付费产品文档. 2026（访问日期：2026-08-11）.

[9] 《中华人民共和国电子签名法》（2019 年修正），中国人大网。

[10] Amazon.com Services LLC v. Perplexity AI, Inc., U.S. federal litigation, 2025–2026。

[11] Akerlof, G. A. The Market for "Lemons": Quality Uncertainty and the Market Mechanism. Quarterly Journal of Economics, 1970.

[12] Simon, H. A. A Behavioral Model of Rational Choice. Quarterly Journal of Economics, 1955.

[13] Jensen, M. C., & Meckling, W. H. Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure. Journal of Financial Economics, 1976.

[14] Spence, M. Job Market Signaling. Quarterly Journal of Economics, 1973.

[15] Stiglitz, J. E., & Weiss, A. Credit Rationing in Markets with Imperfect Information. American Economic Review, 1981.

[16] Coase, R. H. The Nature of the Firm. Economica, 1937.

[17] Hayek, F. A. The Use of Knowledge in Society. American Economic Review, 1945.

[18] Ricardo, D. On the Principles of Political Economy and Taxation. 1817.

[19] Kreps, D. M., Milgrom, P., Roberts, J., & Wilson, R. Rational Cooperation in the Finitely Repeated Prisoners' Dilemma. Journal of Economic Theory, 1982.

[20] Hacker, P., & Holweg, M. A pragmatic approach to regulating AI agents. 2026；另参见 AI Agents Under EU Law, 2026。

[21] Feng, Y. et al. MoltNet: Understanding Social Behavior of AI Agents in the Agent-Native MoltBook. arXiv, 2026.

[22] Li, N. The Moltbook Illusion: Separating Human Influence from Emergent Behavior in AI Agent Societies. arXiv, 2026.

[23] Odaily, "Panorama of Crypto AI Protocols: Starting from Ethereum's Main Battlefield", 2026-04.

[24] Tiger Research / Odaily, "How Crypto Giants Are Betting on AI Agent Payment Infrastructure", 2026-02.

[25] Anthropic, "How we contain Claude across products", 2026.

[26] Invariant Labs, MCP Tool Poisoning 攻击披露, 2025-04.

[27] Odaily, "Is the AI Agent Economy Truly Starting to Operate in 2026?", 2026-02.

[28] 新华社, "助力全球 AI 经济发展 蚂蚁提供全球化 AI 支付解决方案", 2026-05-26.

[29] 中关村在线,"支付宝 2026 年上线 AI 付功能，全面支持 OpenClaw 智能体实现语音一键支付", 2026-04-22.

[30] Trustprint, "AI Agent Reputation Systems: From ERC-8004 Validation to Cross-Registry Trust", 2026-05.

[31] ATXP, "Mastercard Agent Pay vs Stripe ACP vs Visa Intelligent Commerce vs x402: A 2026 Update", 2026-03.

[32] RisingWave, "Mastercard Agent Pay vs Visa vs Stripe Agentic Commerce: 2026 Comparison", 2026-05.

[33] Tiger Research / Odaily, "AI Agents Will Also Need to Verify Their Identity", 2026-05.

[34] ChainCatcher, "The Awakening of Sovereign Agency: From the Local Revolution of OpenClaw to the Financial Framework of Silicon-based Society", 2026-04.

[35] 安全内参, "AI 智能体专门立法与现行法如何互补融合", 2026-04.

[36] Decipher Club, "So, what exactly are 'Trustless Agents' up to?", 2026-07.

[37] NIST AI RMF Agentic Profile.

[38] W3C DID 与 Verifiable Credentials 规范.

[39]Quentin Dupont、Jonathan M. Karpoff,A theory of social capital and trust,2025-10.

[40]McKinsey，Seizing the Agentic AI Advantage，2025-06.

[41]Gartner，Hype Cycle for Agentic AI，2026，2026-04.

[42]Gartner，Hype Cycle for Digital Identity，2026，2026-07.

[43]Palo Alto Networks，2026 Identity Security Landscape，2026-05.

附录 I 智能体信用体系问卷报告

摘要

本报告基于全球 109 份有效采样问卷，对智能体信用体系（KYA）的行业现状进行系统分析。核心发现：信任摩擦已常态化，76%的受访者过去一年遭遇信任、安全或问责顾虑，23%已被客户或合作伙伴拒绝使用，信任缺位正从隐性效率税转化为直接市场收缩。行业成熟度呈纺锤形：L3 混合信用占 38%，L4 以上不足 15%，多数组织尚未完成从“有人兜底”到“凭记录授权”的转换。能力建设呈结构性失衡：信用维度“强记录、弱评分”（决策日志 81%，信用评分仅 52%），权限维度“强兜底、弱分级”（人工确认 86%，场景化分级仅 52%），39%的受访者认为权限可信是三维中最薄弱环节，问责明确流程仅 17%。需求侧共识强烈：88%呼吁统一评估标准，60%愿意试点或主动采用认证服务。BNPL 场景智能体已深入授信核心环节，效率红利明确但消费者信任仍待建设；SME 市场供给不足，低成本身份、标准化授权模板与平台化记录设施是破局关键。报告据此提出分阶段的行动建议。

1 执行摘要

- 原始回收 113 份；清洗后分析样本为 109 份（移除 1 份测试提交、2 份同一 Prolific ID 的较早重复提交、1 份未通过注意力检查的提交）。
- 信任/安全/问责/合规问题已经频繁影响项目：83/109（76.1%）受访者选择 Frequently 或 Occasionally。
- 建立智能体信用体系的价值共识很强：80/109（73.4%）认为 Extremely/Quite important；96/109（88.1%）认为行业统一标准 Extremely/Quite needed。
- 最突出的短板集中在授权可信：42/109（38.5%）直接选择 Authorization trustworthiness；另有 19 人认为三个维度都偏弱。
- 成熟度仍以人机混合为主：L3 及以上为 56/109（51.4%），但 L1/L2 仍占较大比例，说明多数场景尚未进入独立智能体信用阶段。
- 市场采纳路径更可能从高风险场景试点开始：66/109（60.6%）会主动采用或先在高风险/关键场景试点。

2 数据与方法

- 数据来源：Google Forms 导出的 Excel 问卷结果。
- 清洗规则：删除测试提交；同一 Prolific ID 重复提交仅保留时间最新的一份；删除未通过任一注意力检查的提交。
- 统计口径：单选题按非空回答计算比例；多选题按选项被选中次数计算，比例合计可超过 100%；矩阵/量表题按导出后的逐条题目汇总。

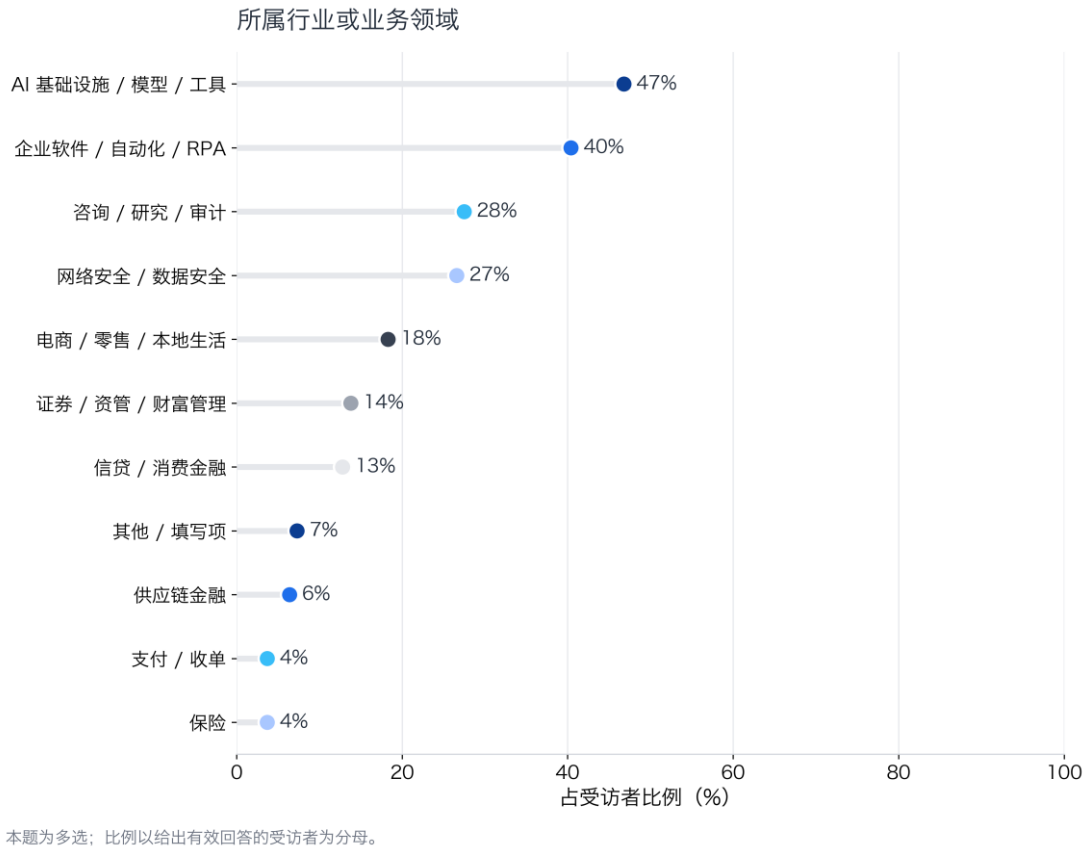
- 隐私处理：Name、Organization、Position/function、Email、Phone/WeChat 等联系字段不进入正文和图表；仅匿名统计“是否愿意作为案例”。

3 数据分析

3.1 样本画像

本章用于判断样本结构是否覆盖产品、技术、风控、合规、运营与咨询研究等角色，并观察受访者最熟悉的智能体场景。

Q1. 所属行业或业务领域



本题为多选；比例以给出有效回答的受访者为分母。

图 1. 多选 最高项：“AI 基础设施/模型/工具”46.8%；“企业软件/自动化/RPA”40.4%；“咨询/研究/审计”27.5%。多选比例合计可超过 100%。

解读：样本以供给侧专业人士为主：AI 基础设施/模型/工具占 47%，企业软件/自动化/RPA 占 40%，咨询研究与网络安全分别占 28%和 27%。这类受访者既参与智能体建设，也承担治理职责，其回答更多反映供给方对信任基础设施的工程判断，而非终端用户感知。对 KYA 研究而言，这一样本结构适合回答“体系应当如何建设”的问题：身份、信用、权限三维的实现路径、评估指标的可采集性、授权机制的工程成本，都需要建设者视角才能给出可操作的答案。同时应注意两点边界：其一，按照主报告 5.1.4 节确立的原则，自报材料不能替代独立验证，本问卷数据属于行业自评，结论应与第三方测评、审计数据交叉使用；其二，需

求侧的信任感知与付费意愿在本样本中覆盖有限，BNPL 与 SME 章节虽部分涉及需求侧问题，但若需完整测量终端企业与消费者态度，应另行补充专项样本。

Q2. 组织/团队规模

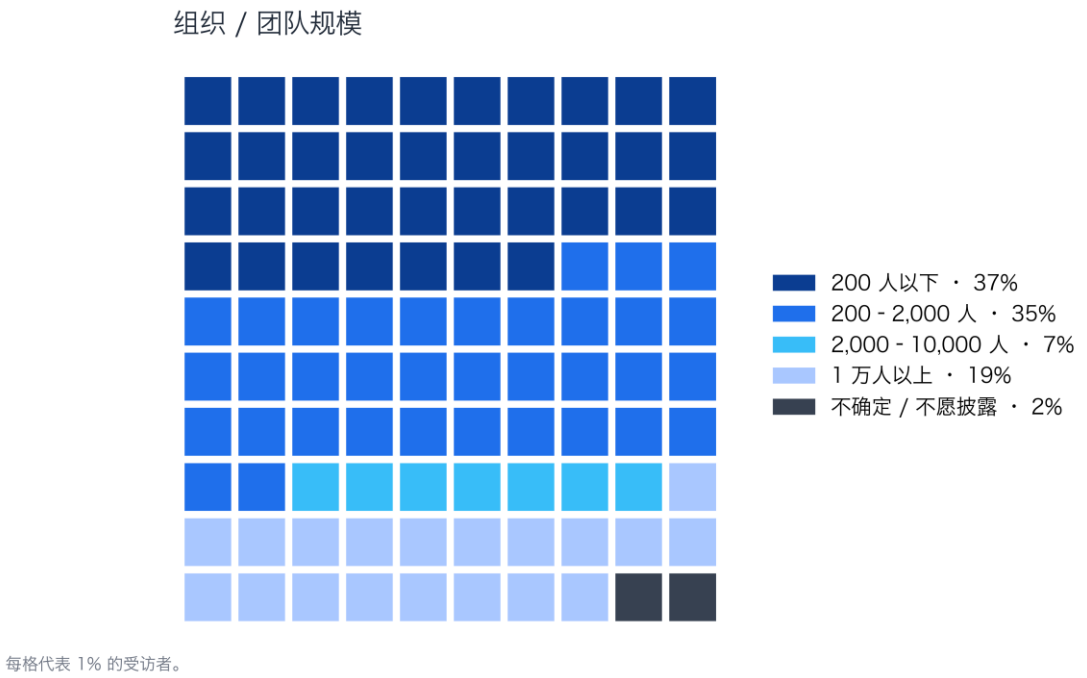


图 2. 最高项为“<200”，占 36.7%。

解读：组织规模呈两极分布：200 人以下团队占 37%，200 – 2,000 人占 35%，万人以上组织占 19%。样本同时覆盖资源受限的中小团队和流程复杂的大型机构，有助于识别信用体系在不同组织条件下的适配差异。结合主报告 5.3.4 节的讨论，中小组织对建设成本敏感，需要平台化、低门槛的起步路径——先落实身份归属、真实履约、授权执行、异常处置四项最低要求，回答清“哪个智能体、什么任务、何种授权、结果如何、谁干预”五个问题；大型组织更关注跨部门问责与审计衔接，且更接近主报告 2.6 节 L4 级所需的四支柱条件：持续身份、经济后果承接、可携带声誉与可执行违约机制。两类组织处于同一成熟度谱系的不同位置，框架设计应允许从同一套指标中按规模裁剪实施深度，而非另立标准。

Q3. 主要岗位或职能

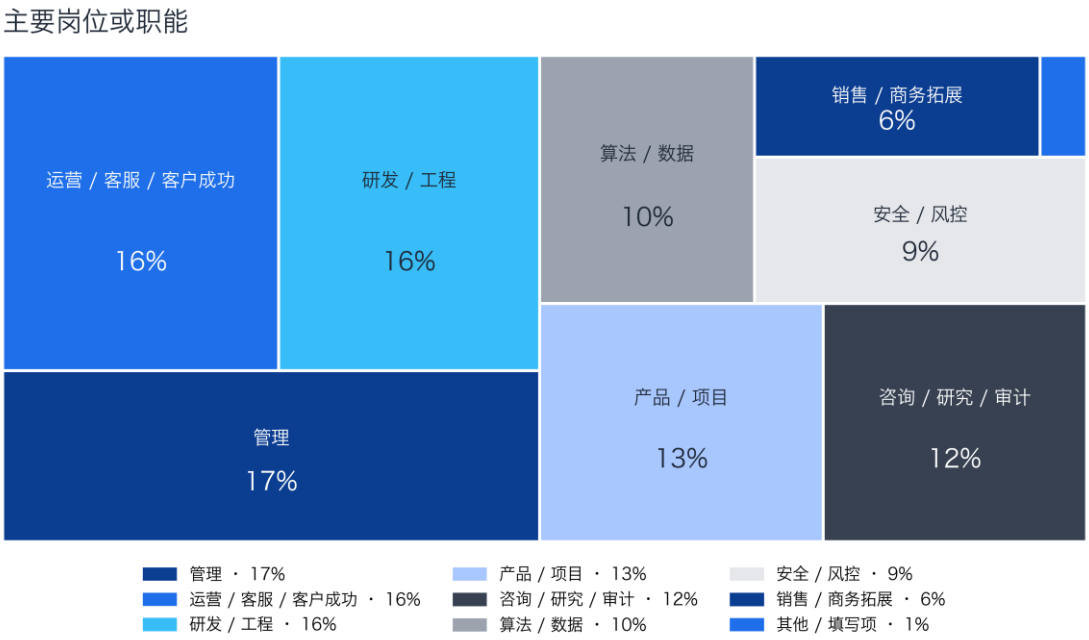


图 3. 最高项为“管理”，占 17.4%。

解读：受访者角色覆盖管理（17%）、运营与客服（17%）、研发（16%）、产品（13%）、咨询研究（12%）、算法（10%）和安全风控（9%），没有单一职能占主导。这一分布表明智能体信任问题跨越建设、运营与治理三条职能线，任何单一部门的视角都不足以完整描述风险。后续章节中问责机制、授权边界等议题的高关注度，正是在这种跨职能背景下出现的，说明相关痛点具有组织普遍性，而非个别岗位的局部体验。从主报告 3.3.7 节分层责任的设计看，智能体事故的责任认定本就涉及开发者、部署方、运营方与监督者多个角色，需要证据链与先行赔付等机制在角色之间分配损失；本样本的职能构成恰好覆盖了这条责任链上的主要节点，使关于问责与授权的回答能够反映责任分配的真实张力，而非单一立场的诉求。

Q4. 与 AI 智能体联系

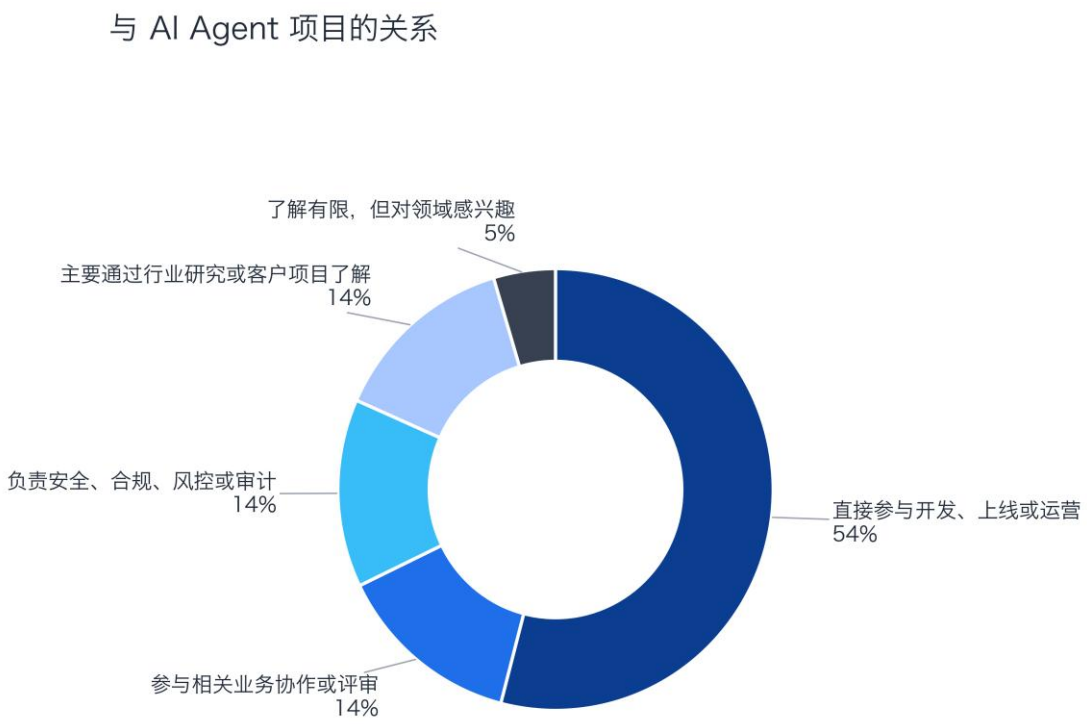
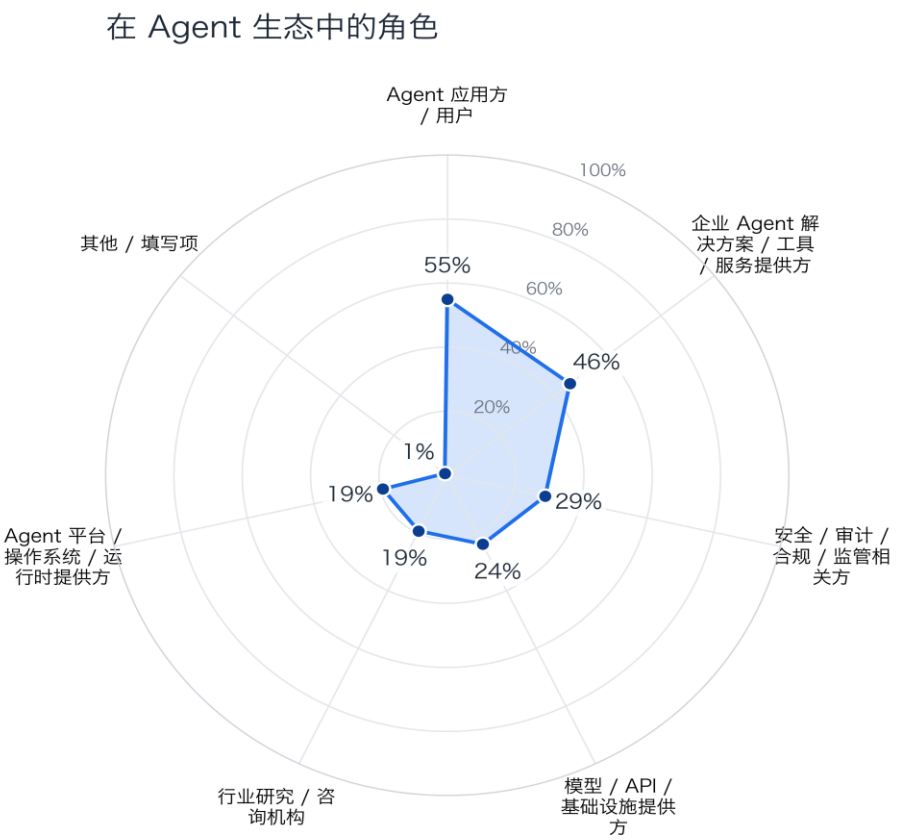


图 4. 最高项为“直接参与开发、上线或运营”，占 54.1%。

解读：54%的受访者直接参与智能体的开发、上线或运营，其余主要分布在业务协作评审、安全合规风控和行业研究等相邻职能，仅 5%表示了解有限。样本以实践者为主，数据反映的是生产环境中已经发生的信任摩擦，而非基于公开报道的外围判断。这为第一章关于信任缺位代价的论述提供了来自一线的经验证据：主报告 1.2.2 节描述的“个体—企业—市场”代价传导链条，其起点正是这些一线岗位日常处理的异常事件、客户投诉与人工干预。实践者占比高还意味着问卷捕捉到的问题更接近“需要被解决”的类型，而非“可能被想象”的类型，后续各章节的痛点排序因此具有较高的行动参考价值。同时也应保持方法论上的克制：按照主报告 5.1.4 节的原则，自评数据反映的是受访者愿意报告的状态，实际运行中的失败率与越权事件通常高于自评水平，结论应与独立测评交叉验证。

Q5. 在智能体生态中的角色

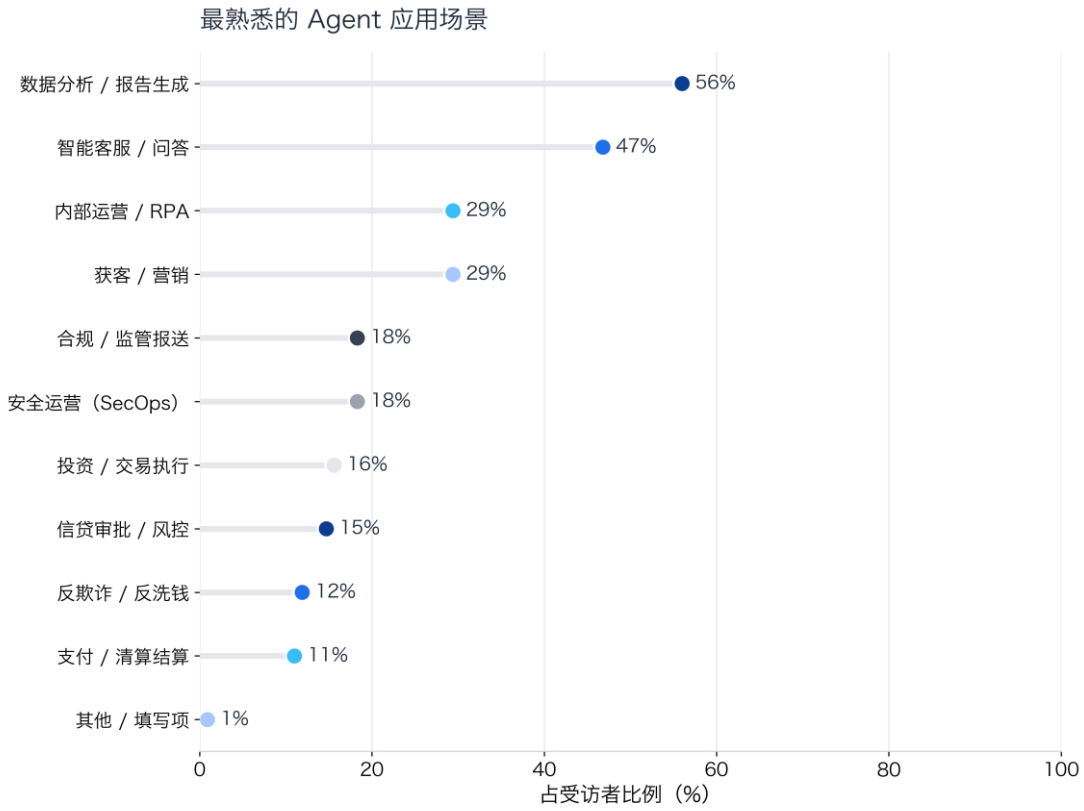


本题为多选；比例以给出有效回答的受访者为分母。

图 5. 多选 最高项：“智能体 应用方/用户”55.0%；“企业智能体解决方案/工具/服务提供方”45.9%；“安全/审计/合规/监管相关方”29.4%。多选比例合计可超过 100%。

解读：55%的受访者以应用方/用户身份参与智能体生态，46%同时是解决方案或工具提供商，29%承担安全、审计或合规职责，多重身份重叠明显。供需角色的交叉说明行业边界尚未固化，大量组织既是智能体的使用者也是建设者。对 KYA 的含义是，信用体系需要同时服务“自建的智能体”与“外部引入的智能体”，身份归属与授权关系的表达不能假设单一角色立场。具体而言，主报告 3.1.2 节提出的身份三问——它是谁、代表谁、现在还是它吗——在供需交叉的组织中会同时以两种视角被追问：作为建设方要回答“我的智能体是谁、能做什么”，作为使用方要核验“对方的智能体代表谁、出了问题找谁”。这种双重身份还意味着，未来信用记录的流通将沿既有供需关系网络展开，具备形成网络效应的现实基础：同一批组织既是信用数据的生产者，也是消费者。

Q6. 最熟悉的智能体应用场景



本题为多选；比例以给出有效回答的受访者为分母。

图 6. 多选 最高项：“数据分析/报告生成”56.0%；“智能客服/问答”46.8%；“内部运营/RPA”29.4%。多选比例合计可超过 100%。

解读：受访者最熟悉的场景是数据分析与报告生成（56%）和智能客服（47%），属于低风险、易回滚的任务类型；信贷审批（15%）、反欺诈（12%）、支付清算（11%）等资金相关场景的熟悉度明显偏低。这一梯度与第二章的成熟度逻辑一致：低风险场景先行积累履约记录，高价值场景因信任不足而渗透缓慢。信任机制的市场扩张效应在此有直接体现——可信证据不足时，交易范围会向低风险任务收缩，高价值场景的市场规模因此被抑制。反过来看，这一分布也提示了冷启动的现实路径：按照主报告 3.2.5 节的四条路径，沙盒任务与低风险准入场景是积累首批履约样本的自然载体；数据分析、客服等场景的高熟悉度意味着这些场景已具备规模化产生履约数据的条件，缺的只是将记录转化为可验证信用证据的标准化机制。资金类场景的低渗透则提示，进入这些场景前需要先完成信用证据的跨场景携带与核验能力建设。

3.2 应用现状与信任挑战

本章回答：信任问题是否已经影响落地、影响了哪些环节，以及业务/合规/技术三条线的核心痛点。

Q7. 信任/安全/问责/合规顾虑的发生频率

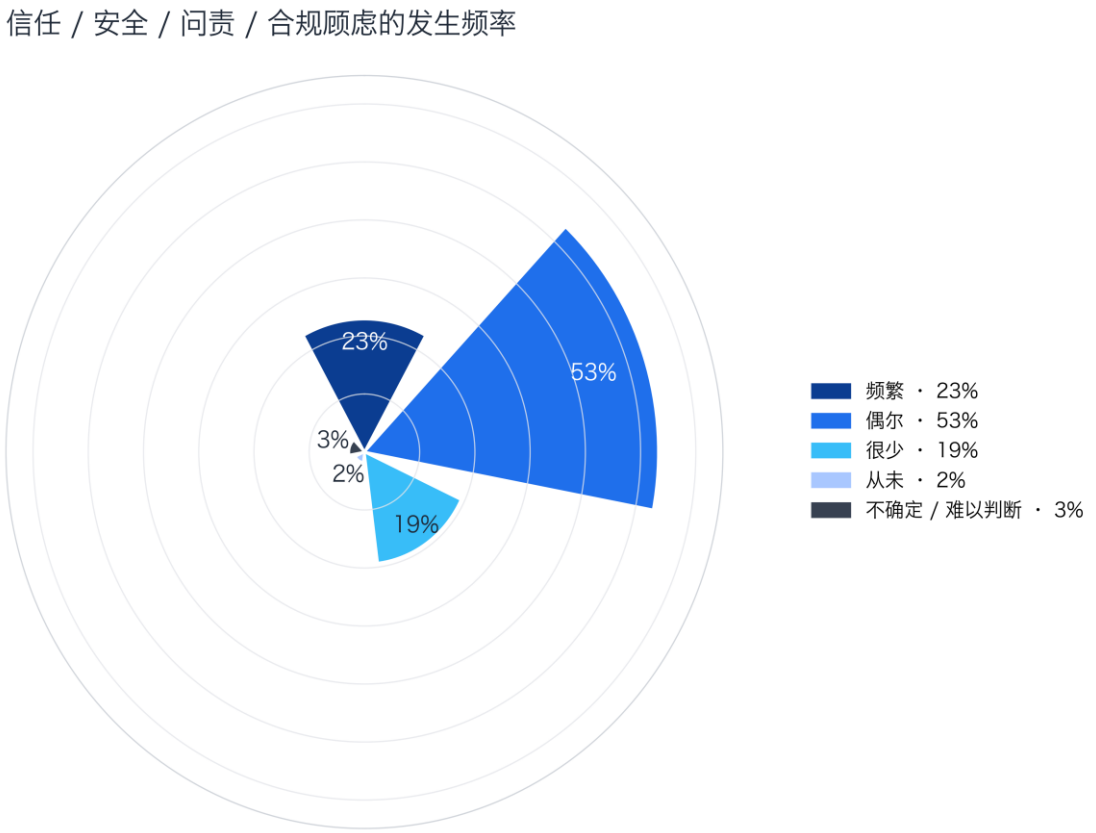


图 7. 最高项为“偶尔发生”，占 53.2%。

解读：76%的受访者在过去一年频繁（23%）或偶尔（53%）遇到信任、安全、问责或合规方面的顾虑，仅 2%从未遇到，信任摩擦已经常态化。结合主报告 1.2 节的分析，智能体错误具有速度、规模和可复制性特征，单次失误可能在数秒内被放大为系统性损失；当顾虑成为高频体验时，组织会本能地收缩授权范围，这正是信任真空转化为经济成本的微观机制。“偶尔遇到”的 53%同样具有指示意义：信任问题并非集中于少数管理不善的组织，而是分布在行业的普遍实践中。这与主报告 1.2.1 节的判断一致——信任缺位不是个别事故的结果，而是智能体无原生信用、“身份—行为—责任”链条不闭合的结构性后果。链条不闭合时，每一次顾虑都无法沉淀为可复用的信用信息，组织只能反复以临时性的人工审查应对，这正是后文 Q8 中各类效率代价持续累积的根源。

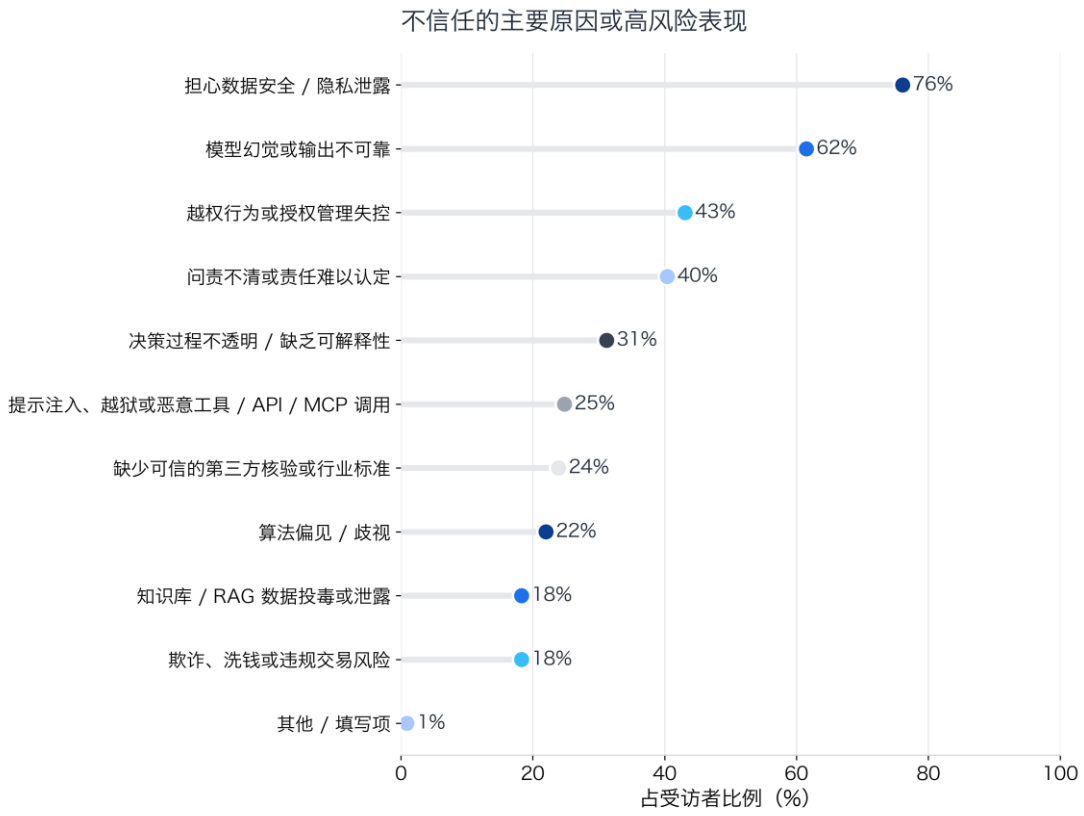
Q8. 智能体信任顾虑带来的影响



图 8. 多选 最高项：“合规、审计或监管评估压力上升”41.3%；“项目上线或扩张延迟”40.4%；“大量人工缺人，效率下降”37.6%。多选比例合计可超过 100%。

解读：信任顾虑的代价分布为：合规与审计压力上升（41%）、项目上线或扩张延迟（40%）、人工确认增加导致效率下降（38%）、数据/API/系统访问受阻（35%）、投诉与问责困难（29%），另有 23%已遭遇客户或合作伙伴拒绝使用。前三项属于隐性交易成本，与主报告 1.2.4 节 TC=f(I,R,C,E)框架中的监督执行成本和合同复杂度相对应：组织被迫用人工复核、流程加签和访问封锁来对冲信任缺口，每一项都是可直接计量的效率税。客户拒绝使用（23%）则是直接的市场收缩，对应框架中市场扩张效应的受损。两组代价的相对比例还揭示了一个时序特征：效率损失先于市场损失出现，意味着当企业感知到客户流失时，内部早已为信任缺位支付了长期的隐性成本。这一时序对建设时机的含义是：信用体系的投入应在市场损失显性化之前启动，此时效率节约即可构成可测算的投资回报。

Q9. 智能体使用中不信任的主要归因或高风险表现



本题为多选；比例以给出有效回答的受访者为分母。

图 9. 多选 最高项：“担心数据安全/隐私泄露”76.1%；“模型幻觉或输出不可靠”61.5%；“越权行为或授权管理失控”43.1%。多选比例合计可超过 100%。

解读：不信任的主要来源为数据安全与隐私泄露（76%）、幻觉或输出不可靠（62%）、越权与授权失控（43%）、问责不清（40%）、决策不透明（31%）。前两项指向能力层面的可靠性，中间两项指向制度层面的可控性，与第三章“身份—信用—权限”的三维划分基本对应。这一分布支持主报告 3.2.1 节的核心判断：能力不等于可信。幻觉与隐私问题即便通过模型改进得到缓解，越权与问责问题仍需要独立的制度安排来解决——这正是信用维度在三层评分中权重最高（45%）、且金融信贷场景进一步上调至 45%的原因。提示注入、数据投毒等对抗性风险的选择率（25%、18%）相对靠后，说明当前阶段的信任瓶颈仍以基础可靠性为主，高级攻击面尚未成为主要矛盾；但随着智能体权限扩大、暴露面增加，对抗性风险的权重应在评估框架中动态上调，而非固定不变。

Q10. 推进智能体应用的商业痛点

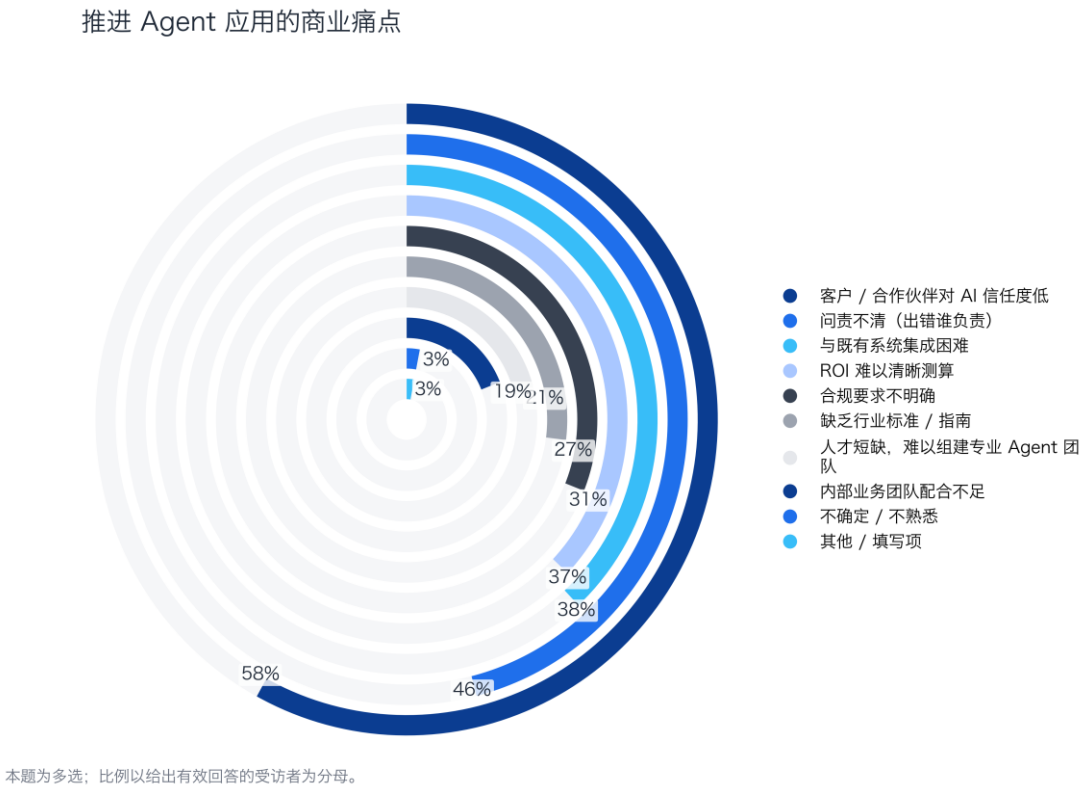
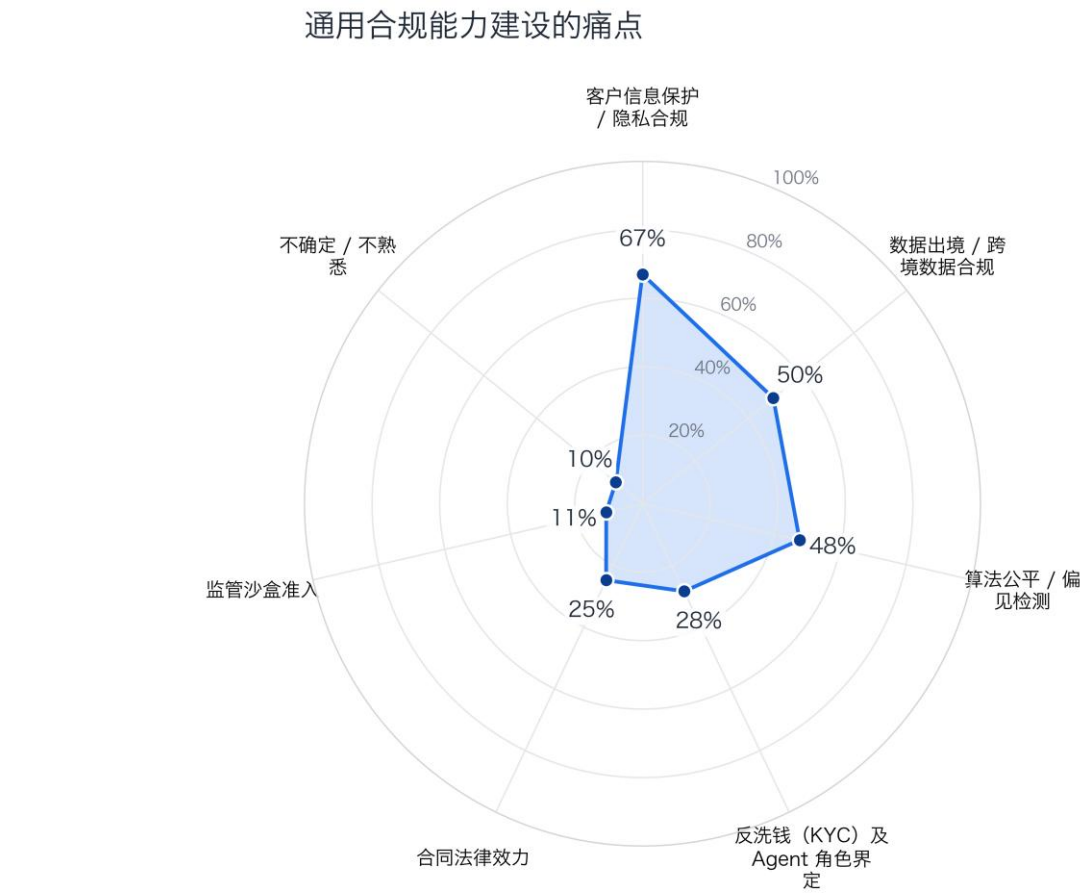


图 10. 多选 最高项：“客户/合作伙伴对 AI 信任度低下”53.2%；“问责不清”42.2%；“与既有系统集成困难”34.9%。多选比例合计可超过 100%。

解读：商业层面的首要痛点是客户与合作伙伴对 AI 的信任度低（53%），其次是出错后的问责不清（42%），高于 legacy 系统集成（35%）和 ROI 测算困难（34%）。商业化约束已经从技术可行性转向交易对手接受度，符合主报告 1.2.4 节“商业价值=价值创造×信任机制”的判断：信任机制缺位时，价值创造能力难以兑现，技术领先无法自动转化为商业成功。问责不清排名靠前还说明，责任安排本身就是交易条件的一部分，而非事后管理问题——采购方在签约前就需要回答“出错后谁赔、赔多少、凭什么证据认定”。这解释了主报告 3.3.7 节为何将证据链、先行赔付与责任险列为权限体系的配套机制：没有可执行的损失承接安排，信任度低的问题无法仅靠技术演示来化解。ROI 测算困难（34%）与信任痛点并存也并非巧合：信任成本难以计量，正是 ROI 模型失真的重要来源。

Q11. 智能体通用合规能力建设的痛点



本题为多选；比例以给出有效回答的受访者为分母。

图 11. 多选 最高项：“客户信息保护/隐私合规”67.0%；“数据出境/跨境数据合规”49.5%；“算法公平/偏见检测”47.7%。多选比例合计可超过 100%。

解读：合规建设的痛点集中在客户信息保护与隐私（67%）、数据出境与跨境合规（50%）、算法公平与偏见检测（48%），其后是 KYC 与智能体角色界定（28%）和合同法律效力（25%）。隐私与跨境属于刚性监管约束，选择率高反映数据处理是智能体落地的先决合规环节，任何信用体系的数据采集设计都必须先回答这两项约束。公平性检测接近半数，提示评估指标体系需要覆盖输出公平性维度，不能只考核准确率与效率——主报告 5.2 节 13 个维度、65 项指标的框架中，公平性与偏见属于不可省略的评估项。KYC 与智能体角色界定占 28%，说明部分组织已在处理“智能体在法律关系中处于什么位置”的问题，与主报告 3.1 节身份可信的制度化方向直接相关。仅 11%选择监管沙盒准入，反映多数组织尚未进入需要沙盒的阶段，合规重心仍在基础数据处理环节；该比例可作为行业合规成熟度的参照指标长期跟踪。

Q12. 智能体技术安全痛点

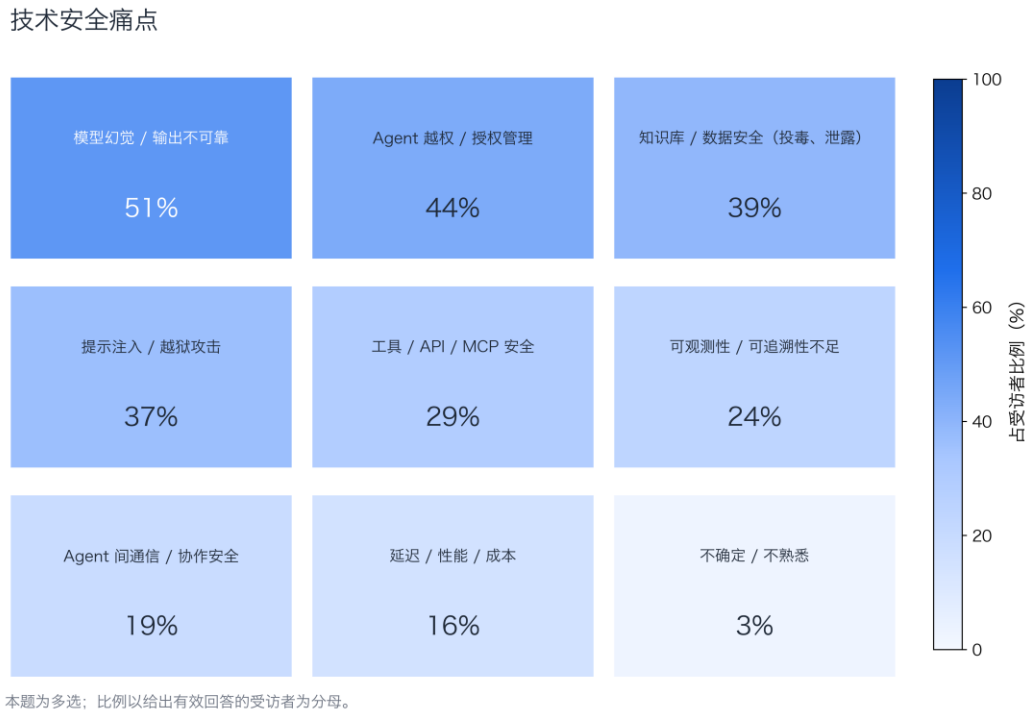


图 12. 多选 最高项：“模型幻觉/输出不可靠”51.4%；“智能体越权/授权管理”44.0%；“知识库/数据安全（投毒，泄露）”39.4%。多选比例合计可超过 100%。

解读：技术安全痛点的前三位是幻觉与不可靠输出（51%）、越权与授权管理（44%）、知识库数据投毒或泄露（39%），其后是提示注入（37%）和工具/API/MCP 安全（29%）。可观测性与追溯不足占 24%，比例虽不突出，但直接影响主报告 3.2.3 节最小事件集的采集质量：没有完整的调用链与输入输出记录，履约事件、异常事件和用户纠正都无法形成可验证的信用证据，信用体系将因缺乏数据源而空转，可观测性投入应被视为信用基础设施而非运维附属。工具/API/MCP 安全占 29%同样值得重视：主报告 5.2.4 节强调评估应检查实际执行、工具组合与委托链，智能体的风险不仅来自自身模型，还来自其可调用的工具网络和下游委托对象；越权问题（44%）的高发，正是工具权限缺乏细粒度约束、授权边界未落到每一次调用上的直接表现。

Q13. 建立智能体信用体系的重要性

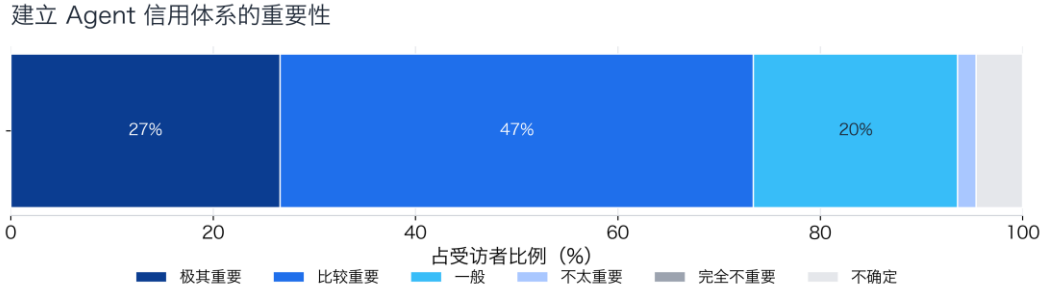


图 13. 最高项为“比较重要”，占 46.8%。

解读：73%的受访者认为建立智能体信用体系极其重要（27%）或比较重要（47%），仅 2%认为不重要，需求共识已经形成，且主要动机是提升客户与合作伙伴信任，而非单纯满足监管要求。这一结果支持主报告 1.2.4 节的分析框架：信用体系的价值在于降低交易成本并扩大交易可行范围，其回报通过效率效应与市场扩张效应兑现，而非仅体现为合规成本的下降。需求动机的结构对体系设计有实际含义：由市场信任驱动的建设，更可能接受“信用记录可携带、可跨场景展示”的设计，因为这直接服务于获客与签约；若主要由监管驱动，设计重心则会偏向报告合规而非证据流通。需要注意的是，仍有 20%选择“中等重要”，这部分组织可能尚未遇到信任约束；从 Q14 的成熟度分布看，他们多处于 L1－L2 阶段，其态度会随智能体权限扩大而变化，是未来两到三年需求增量的主要来源。

Q14. 当下 AI 智能体成熟度等级

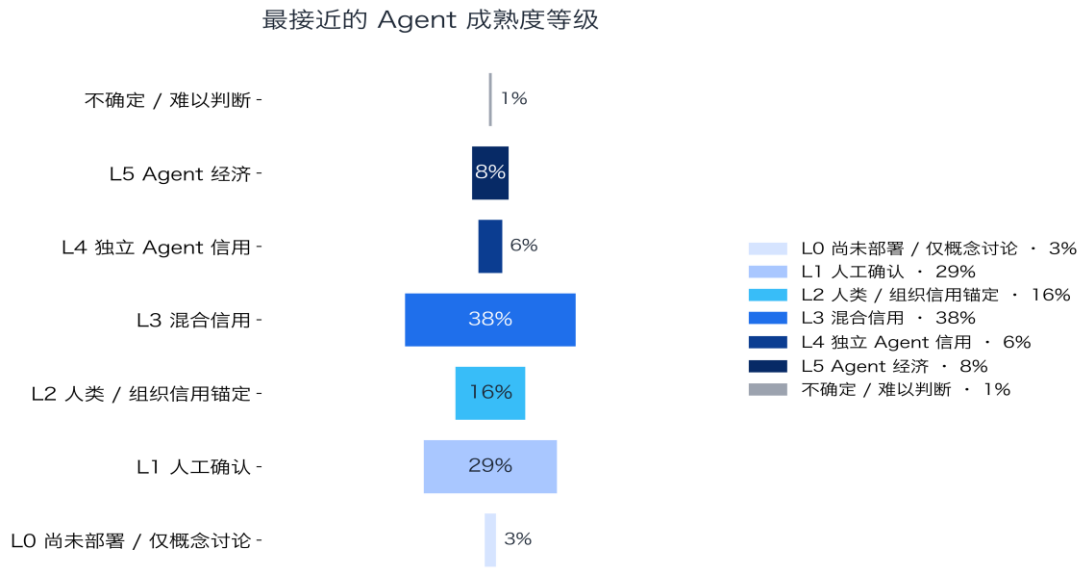


图 14. 最高项为最高项为“L3 混合信用:智能体可以在人为定义下的边界自动化运行”，占 37.6%。

解读：成熟度分布呈纺锤形：L3 混合信用占 38%，L1（29%）与 L2（16%）合计 45%，L4（6%）与 L5（8%）合计不足 15%。行业整体处于人类锚定信用向任务自治过渡的区间。对照主报告 2.1.2 节“能力证明—边界定义—控制建立—受限授权—履约记录—信用形成—扩大授权”的升级闭环，多数组织尚未完成从“有人兜底”到“凭记录授权”的转换：L3 的特征仍是人类作为后援接管者，信用锚定在操作者而非智能体自身。L4 以上占比偏低与两组因素相互印证：一是 Q41 的性能数据显示准确率与人工干预率尚不足以支撑独立信用；二是主报告 2.6.1 节提出的 L4 四支柱——持续身份、经济后果承接、可携带声誉、可执行违约机制——在当前基础设施下均不完备。纺锤形分布意味着未来三年的主要迁移将发生在 L2→L3→L4 区间，评估与认证服务的首批需求也将集中在这一迁移带上。

Q15. 智能体迈向更高成熟度的主要阻碍

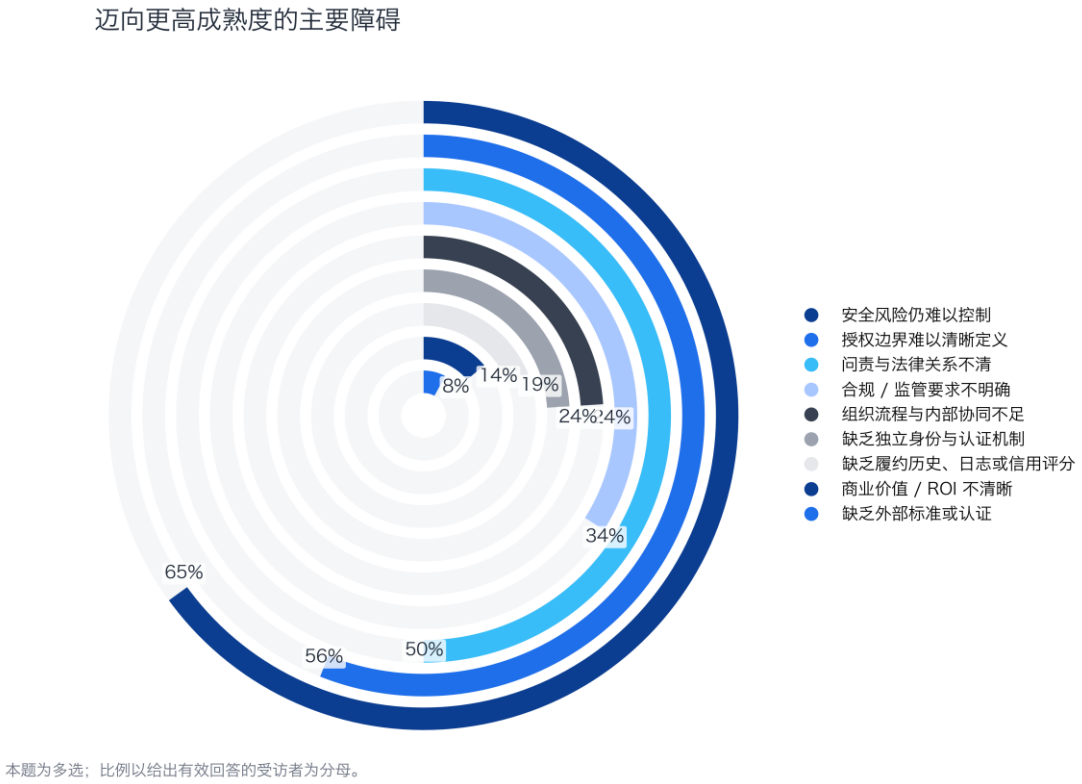


图 15. 多选 最高项：“安全风险仍难以控制”59.6%；“授权边界难以清晰界定”51.4%；“问责与法律关系不清”45.9%。多选比例合计可超过 100%。

解读：升级障碍的前三位是安全风险难以控制（60%）、授权边界难以定义（51%）、问责与法律关系不清（46%），均属于权限可信与问责机制范畴；缺乏外部标准或认证仅占 7%。这一结构表明，成熟度跃迁的主要约束在组织内部的工程与治理能力，而非外部标准缺位——行业不能把停滞归因于“没有标准可遵循”。从主报告 2.8.2 节信用飞轮的角度看，安全与授权环节不通，履约记录就无法转化为更大授权，飞轮在“控制建立”环节停滞。具体而言，授权边界难定义（51%）意味着组织无法将任务风险映射为主报告 3.3.3 节五级授权强度中的确定等级；问责不清（46%）意味着即使记录了履约数据，也没有责任主体愿意基于这些数据签署更大授权。两者共同解释了 Q14 中 L3 向 L4 迁移的阻塞点，也指明了干预顺序：先解决授权边界的可操作定义与问责的制度化，信用积累才有启动条件。

3.3 KYA 能力建设现状

本章将身份可信、履约/信用可信、授权可信拆开看，识别哪些能力已实施、哪些仍停留在计划或缺失状态。

Q16. 智能体身份可信能力

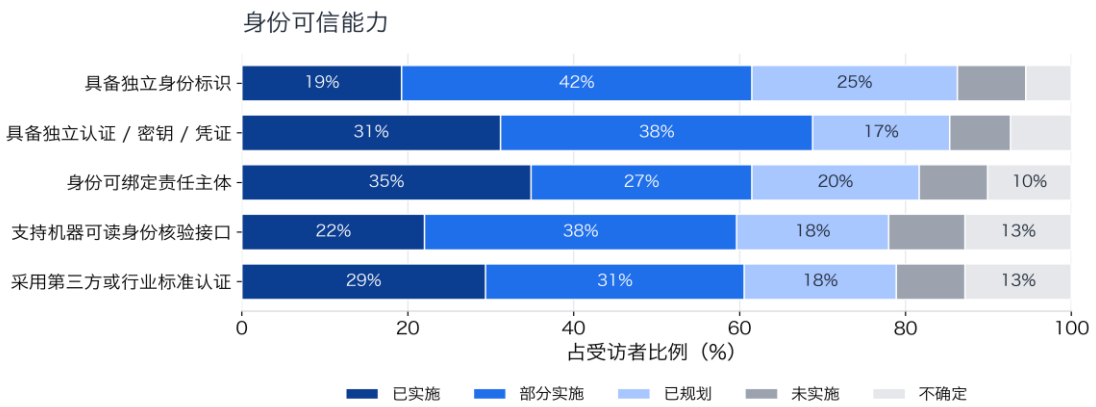


图 16. “已实施+部分实施”比例最高的条目是“智能体目前具备独立认证/密钥/凭证能力”（68.8%）。

解读：身份可信五个要素的已实施或部分实施比例处于 60% - 69%的窄区间：独立认证与密钥凭证最高（69%），机器可读验证接口最低（60%）。身份维度的建设起步相对均衡，说明“它是谁、代表谁”这两个问题已有基础解法；但机器可读接口这一短板值得关注：它是跨组织核验和主报告 3.4.5 节协议化的前提，接口不足意味着身份验证仍依赖平台内部闭环，难以支持 3.1.9 节讨论的可携带身份——智能体在一个平台积累的身份与记录，无法被另一个平台低成本核验。身份是信用的载体，归因能力不完整会向上传导至信用与权限两维：无法稳定回答“现在还是它吗”，履约记录就可能挂错主体，授权决策也失去锚点。60% - 69%的区间还意味着三成以上的组织连基础身份要素都未完整实施，身份维度的普及仍是三维建设的前置任务。

Q17. 智能体信用的可信能力矩阵

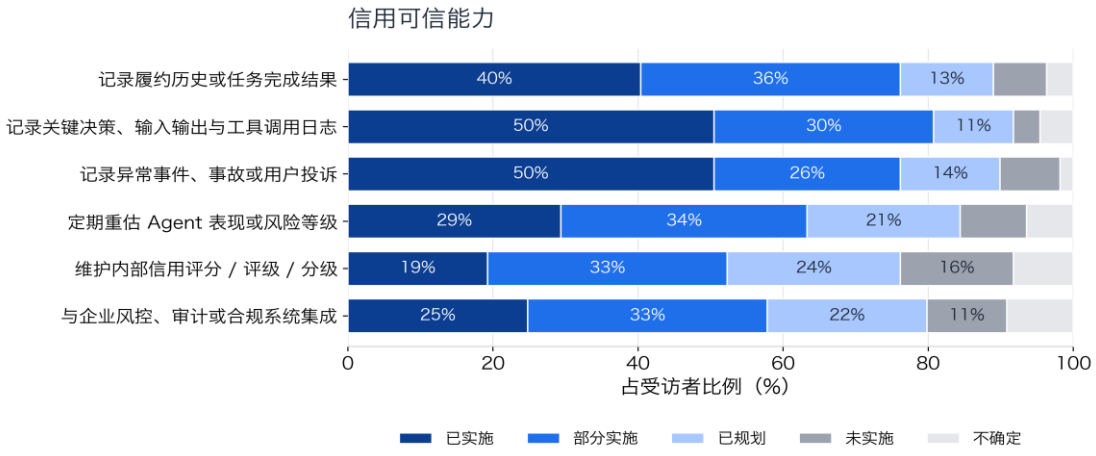


图 17. “已实施+部分实施”比例最高的条目是“记录关键决策、输入输出与工具调用日志”（80.7%）。

解读：信用可信维度内部明显分化：关键决策日志（81%）、履约历史（76%）、异常事件记录（76%）等记录类能力普及率较高，而内部信用评分（52%）和与风控、审计、合规系统的集成（58%）明显偏低。行业整体停留在数据采集层，尚未进入评估判断层，与主报告 3.4.5 节“记账层与判断层分离”的架构判断一致。记录是信用的原料而非信用本身；缺少评分与集成，已有记录无法参与授权决策，证据资产处于闲置状态——这直接限制了 3.2.4 节 α + β 混合信用的实现：α（主体信用）与 β（任务履约）的合成需要评分层对原始记录的结构化处理。从主报告 3.2.2 节履约五要素看，记录类能力覆盖了“发生了什么”，而评分与集成回答“值不值得信、敢不敢放权”；后者普及率低约 20 – 29 个百分点，说明信用维度建设的下一阶段重心应从“记全”转向“用起来的判断能力”。

Q18. 智能体权限可信能力

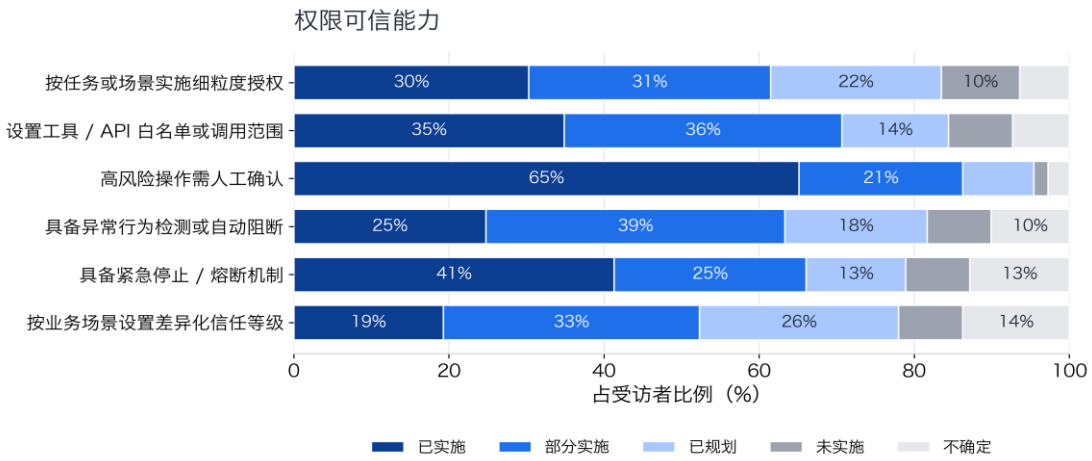


图 18. “已实施+部分实施”比例最高的条目是“高风险操作需人工确认”（86.2%）。

解读：权限可信的实施结构呈“强兜底、弱分级”：高风险操作人工确认达 86%，工具/API 白名单 71%，而按业务场景设置差异化信任等级仅 52%，为六项中最低。人工闸门的普及说明底线意识普遍存在，紧急停止（66%）与异常阻断（63%）也覆盖了主报告 3.3.6 节应急机制的一部分；但场景化分级的缺失意味着授权强度无法随任务风险动态调整，与主报告 3.3.3 节风险分级矩阵的要求存在明显差距——五级授权强度的设计前提，正是组织能够区分任务风险并匹配相应授权等级。长期依赖人工确认还会带来确认疲劳与监督成本的线性上升，分级授权是把人力从低价值审批中释放出来的必要路径。结合 Q22 中 39%受访者认为权限是最薄弱环节的主观判断，权限维度的建设方向已经明确：从“有无闸门”转向“闸门是否分级、分级是否依据信用数据”。

Q19. AI 智能体上线前的安全/合规审查

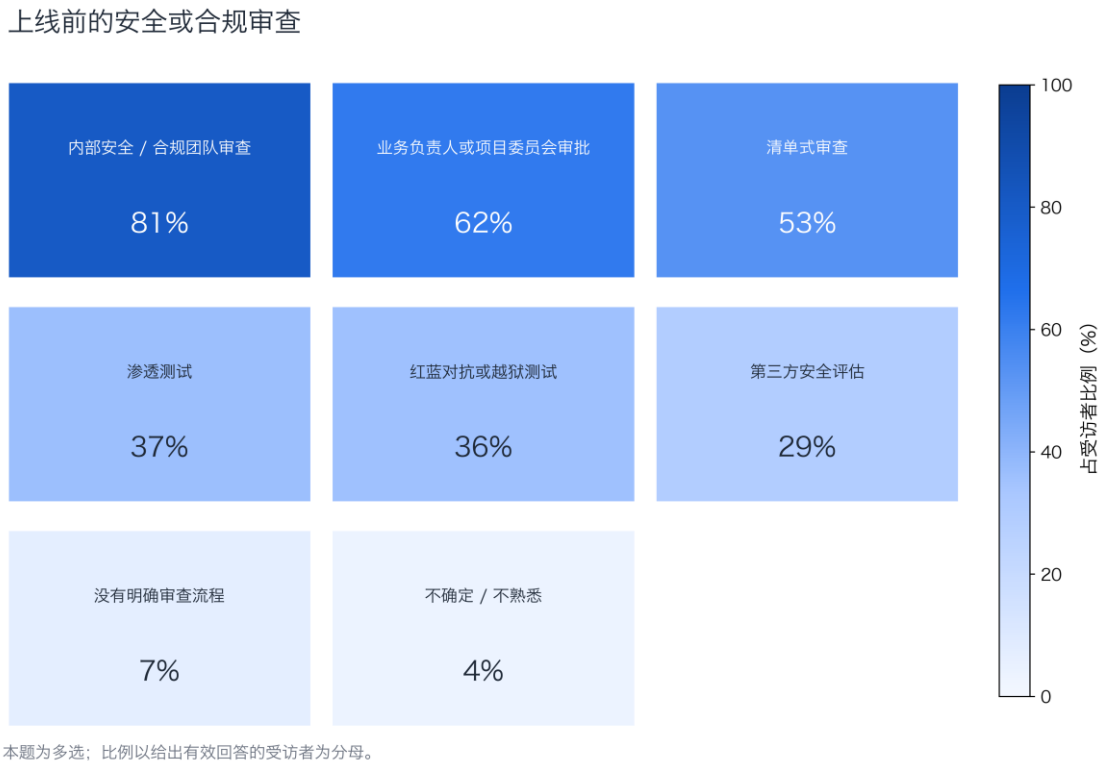
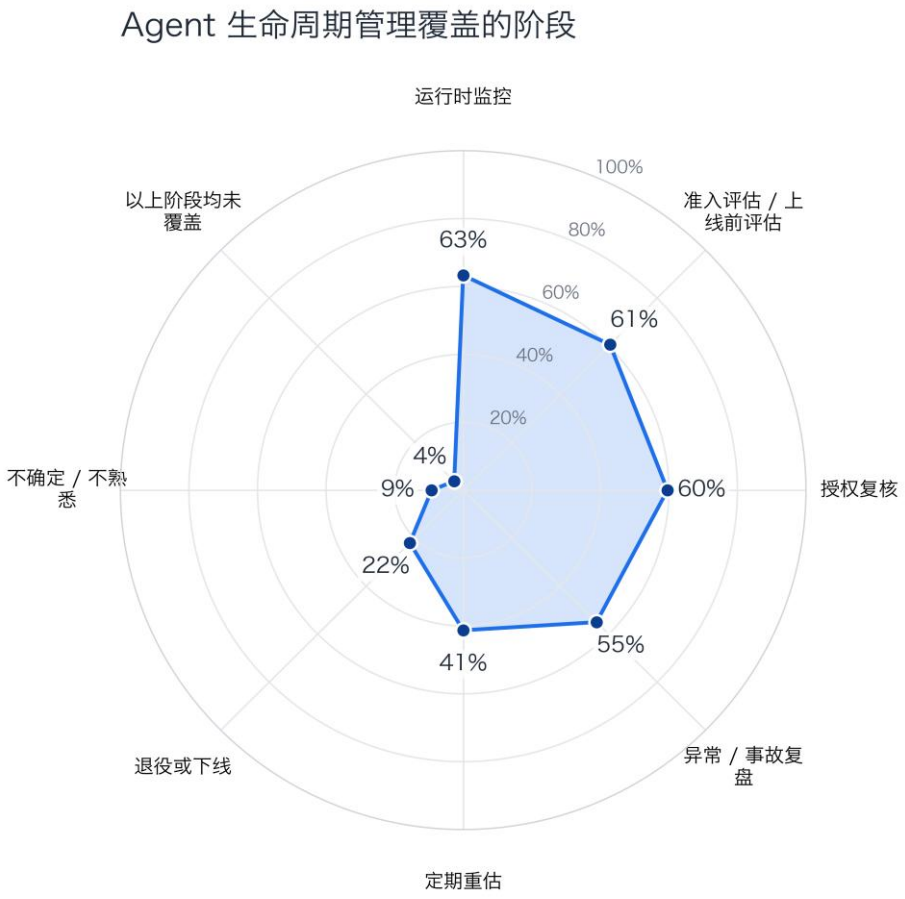


图 19. 多选 最高项：“内部安全/合规团队审查” 80.7%；“业务负责人或项目委员会审批”62.4%；“清单式审查”53.2%。多选比例合计可超过 100%。

解读：上线前审查以内部安全合规团队审查和业务负责人审批为主，第三方安全评估、红蓝对抗与渗透测试的采用率明显较低，另有约 4%没有明确审查流程。审查体系以内部把关为主、外部验证为辅。内部审查熟悉业务上下文、成本可控，但缺乏独立性和对抗性视角，容易系统性低估自身设计盲区；主报告 3.1.8 节讨论的第三方可验真机制目前渗透有限，意味着“自证可信”仍是主流，而交易对手真正需要的是“他证可信”。随着智能体权限扩大、跨组织协作增多，外部评估与对抗测试的缺口会成为高风险场景准入的硬约束：按照主报告 5.2.5 节关键门槛的逻辑，涉及资金与重要数据的场景应将独立安全评估设为准入条件而非加分项。4%的无流程比例虽小，但在高风险场景中不可接受，应作为行业底线清零。

Q20. 智能体生命周期管理



本题为多选；比例以给出有效回答的受访者为分母。

图 20. 多选 最高项：“运行时监控”63.3%；“准入评估/上线前评估”60.6%；“授权复核”59.6%。多选比例合计可超过 100%。

解读：生命周期管理覆盖集中在准入评估与运行时监控，定期重估、授权复核和退役下线的覆盖率明显偏低，约 4%未覆盖任何阶段，管理重心偏向“上线前”。这一缺口直接影响两类机制的有效性：一是主报告 5.1 节的动态评估，信用信息无法随行为持续更新，评估结果会迅速过时；二是 2.8.3 节的动态降级三原则，模型变更、权限调整或异常事件发生后缺少触发复核的制度通道，降级无从谈起。退役下线环节的低覆盖还带来一类隐性风险：不再使用的智能体若保留有效凭证与授权，会成为无人负责的攻击面。信用体系的有效性取决于运营阶段的数据流，而非一次性准入——准入评估回答的是“上线时可不可信”，而市场真正需要持续回答的是“现在还可不可信”，后者只能靠重估、复核与退役机制来保证。

Q21. AI 智能体出错、越权或造成损失后的问责机制

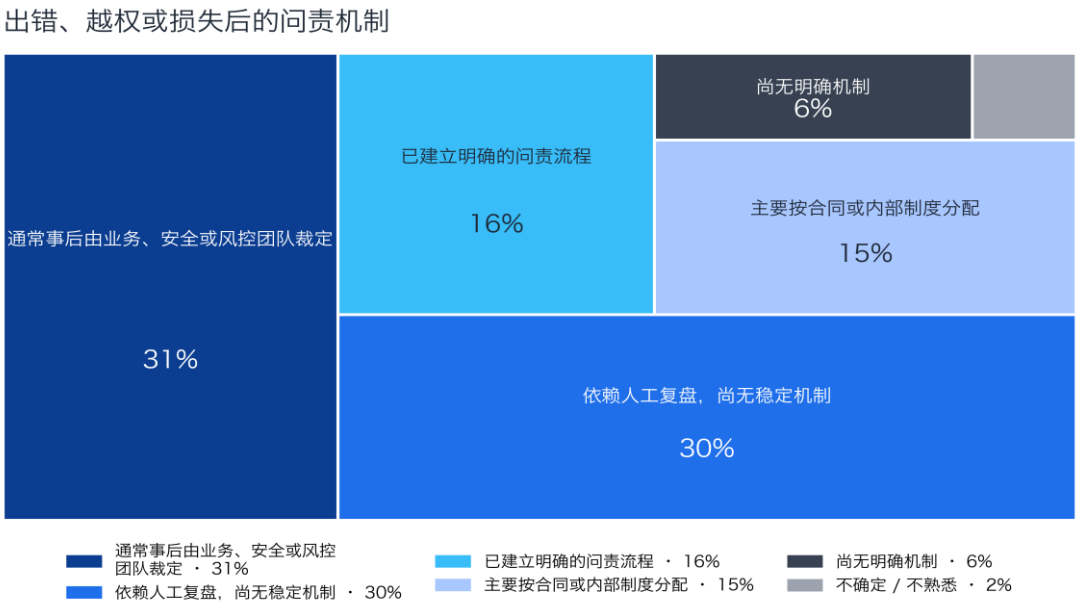


图 21. 最高项为“事后由业务、安全或风控团队裁定”，占 31.2%。

解读：问责机制的现状偏弱：仅 17%已建立明确流程，31%依赖事后由业务、安全或风控团队裁定，30%依靠人工复盘且无稳定机制，6%完全没有机制。按照主报告 2.1.2 节的框架，问责机制与运行边界（ODD）是成熟度判定的两个直接维度；问责的制度真空意味着多数组织在“谁授权、谁监督、损失如何承接”三个问题上没有稳定答案。这不仅放大事故后的处置成本——每次事故都要临时组建裁定班子、重新争论责任边界——也使信用记录缺乏责任锚点：无法归因到主体的履约数据，不能计入任何一方的信用。对照主报告 3.3.7 节的设计，完整的问责安排应包括分层责任、证据链、先行赔付与责任险四个组件；当前 17%的明确流程比例说明多数组织连第一个组件都未完成。问责是三维协同中最需要优先制度化的环节，也是评估框架中一票否决项的合理候选。

Q22. KYA 三维框架目前的局限

KYA 三维中当前最突出的短板

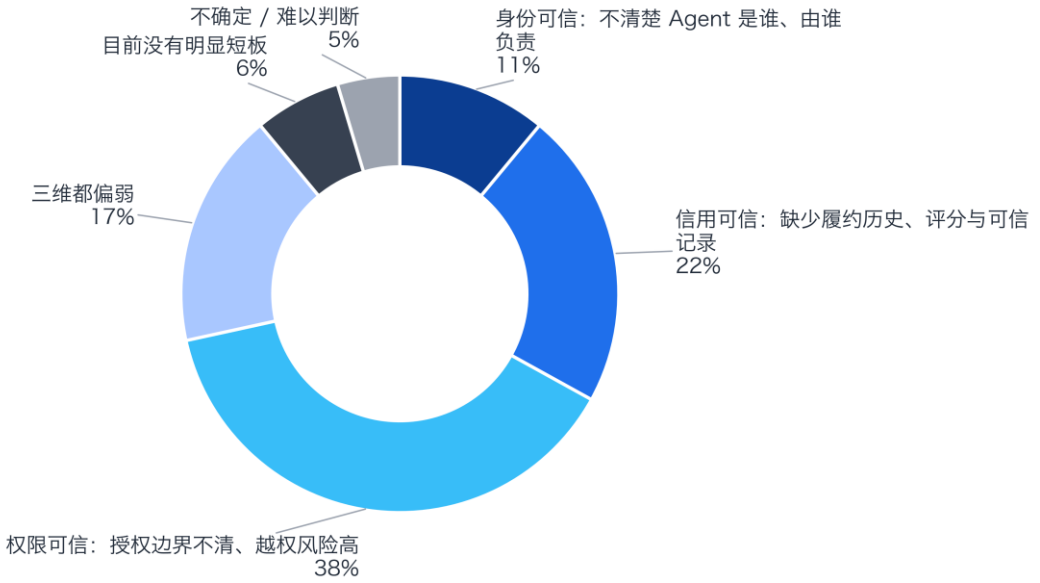


图 22. 最高项为“授权可信： 授权边界不清、越权风险高”，占 38.5%。

解读：关于三维中的最薄弱环节，39%选择权限可信，22%选择信用可信，17%认为三维都偏弱，11%选择身份可信。主观排序与能力建设数据一致：Q18 中场景化信任等级的实施率最低（52%），Q15 中授权边界难定义排名障碍第二（51%），客观实施缺口与主观薄弱感知互相印证。用主报告 3.4.1 节的表述，身份回答“归因”，信用回答“值不值得信”，权限回答“敢不敢放权”；权限边界不清时，即使身份可归因、履约可记录，授权决策仍只能依赖保守的人工判断，信用飞轮因此无法启动——积累的数据换不来更大的授权，记录行为本身就失去激励。值得注意的是17%的“三维都弱”群体：对他们而言，单点突破不足以解决问题，需要的是主报告 5.3.4 节式的整体起步清单。身份可信选择率最低（11%），与 Q16 中各要素 60%以上的实施率一致，身份是当前三维中基础最好的维度。

3.4 信用体系需求

本章聚焦行业是否需要统一标准、框架应优先包含哪些内容，以及组织面对成熟评估/认证服务时的可能态度。

Q23. 对智能体统一可信评估/认证标准的需求

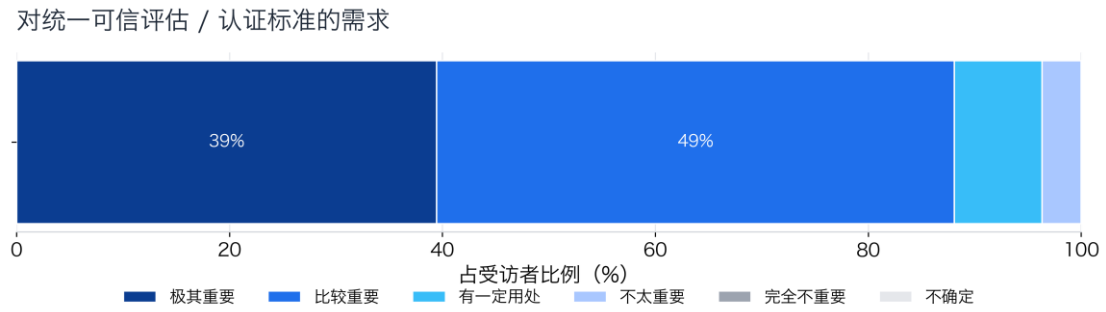
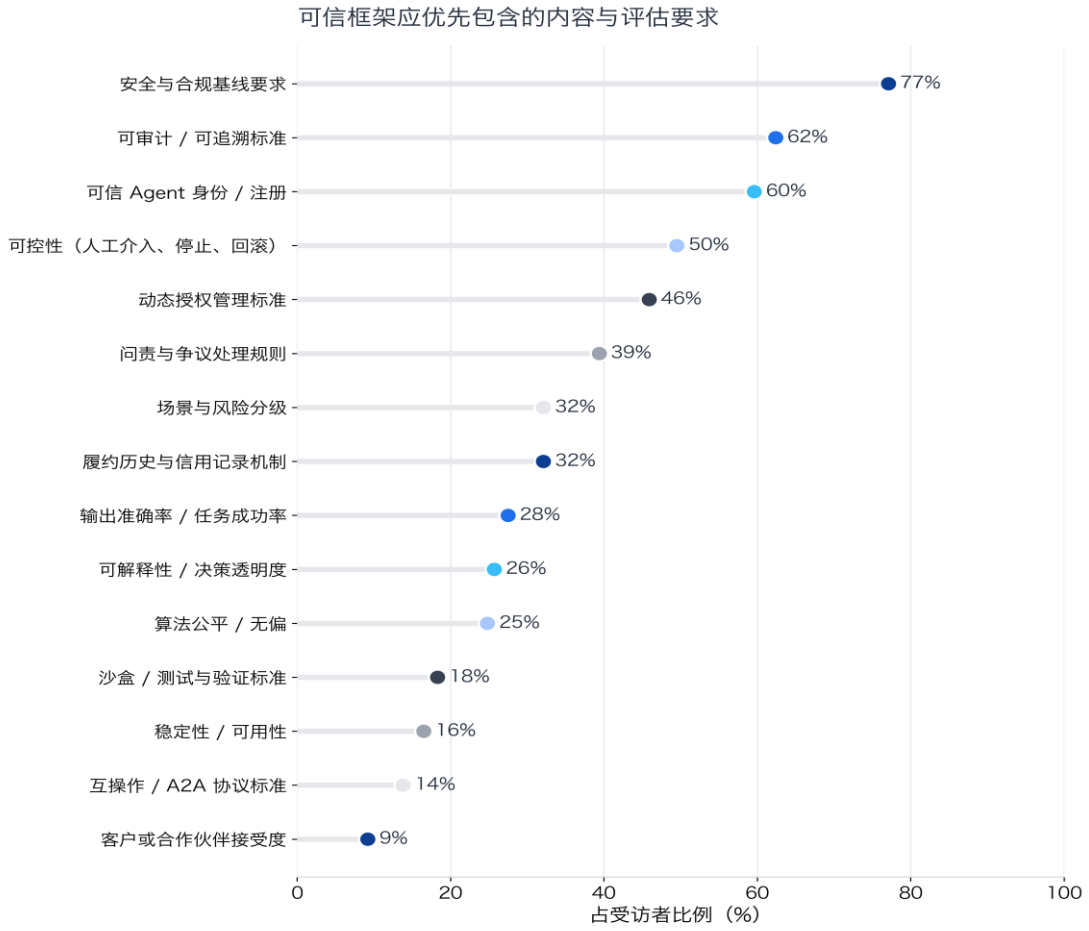


图 23. 最高项为“比较重要；可以从行业发展中收益”，占 48.6%。

解读：88%的受访者认为行业需要统一的智能体可信评估或认证标准，其中39%认为极其需要、应当紧急建立，仅 4%认为不太需要，标准需求已接近共识水平。从信任经济学的角度看，统一标准的价值在于降低跨组织核验成本：没有公认标准时，每个交易对手都需要重复尽调，信用无法跨场景流通，主报告 1.2.4 节所述的市场扩张效应会被验证成本抵消。这一结果为第五章评估框架提供了直接的需求侧证据。结合 Q15 中“缺乏外部标准”仅占 7%的障碍排序看，一个看似矛盾的现象需要解释：组织并不认为标准缺位阻碍了自己的升级，却强烈希望行业有统一标准。合理的解读是，标准的主要价值不在指导自身建设，而在降低与外部交易对手的互信成本——标准首先是一种市场基础设施，其次才是建设指南。这决定了标准设计应优先统一可验证的证据格式，而非规定内部实现路径。

Q24. AI 智能体的可信框架应优先包含的内容与评估要求



本题为多选；比例以给出有效回答的受访者为分母。

图 24. 多选 最高项：“安全与合规基线要求”77.1%；“可审计/可追溯标准”62.4%；“可信智能体身份/注册”59.6%。多选比例合计可超 100%。

解读：框架内容的优先级排序清晰：安全合规基线（77%）、可审计与可追溯标准（62%）、可信身份注册（60%）居前三位；互操作与 A2A 协议（14%）、客户或伙伴接受度（9%）排在末位。市场当前的核心诉求是先建立底线可信，再解决互联互通。这一顺序支持渐进式标准化路线：先统一身份、授权、事件记录的字段与验证方式，使信用证据可生产、可核验；跨平台互操作在证据标准稳定之后再推进，成本更低——没有共同证据格式的互操作，只是把不可比的数据连起来。可审计与可追溯（62%）的高优先级与主报告 3.2.3 节最小事件集的设计直接对应：审计的前提是事件被完整记录，追溯的前提是记录可跨环节拼接。客户接受度仅 9%排在末位，并不说明需求侧不重要，而是反映供给侧的普遍判断：在底线能力建立之前，面向客户的信任沟通缺乏可依托的事实基础。

Q25. AI 智能体信用评估框架设计原则的重要性

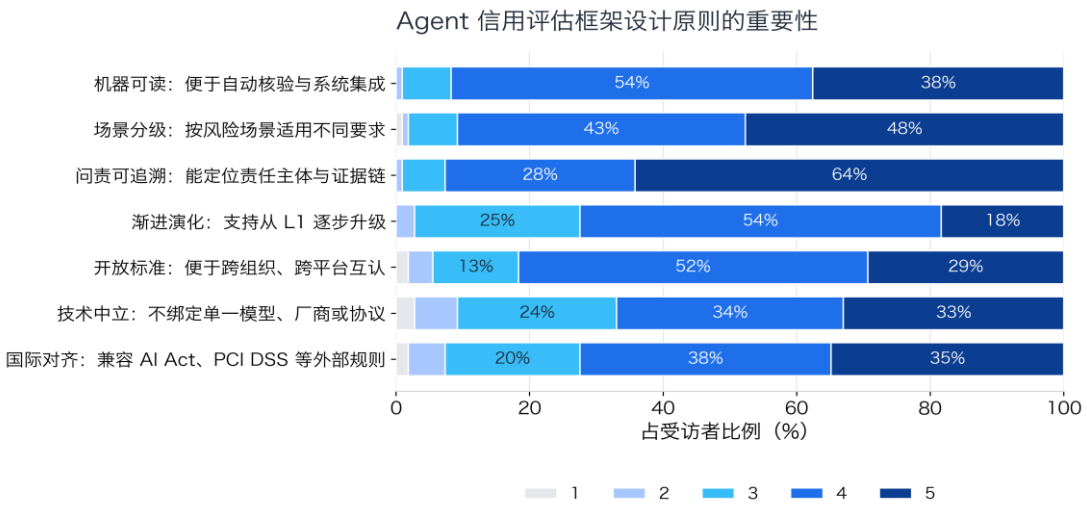


图 25. “5 分（非常重要）”占比最高的是“问责可追溯：能定位责任主体与证据链”（64.2%）。

解读：七项设计原则中，问责可追溯的高分率（4－5 分）达 93%（均分 4.56），机器可读 92%，场景分级 91%，开放标准 82%，渐进演化与国际对齐各 73%，技术中立 67%相对靠后。受访者最重视“关键行为能找到授权来源与责任主体”以及“指标要求与动作风险相适应”，与主报告 5.1.2 节确立的五项原则一致。前三项恰好构成一条完整的证据链：场景分级决定记录什么，机器可读决定记录如何被使用，问责可追溯决定记录最终指向谁——三者同时高分，说明市场对评估体系的核心期待是“证据可用”，而非“评分好看”。技术中立的优先级偏低并不意味着不重要，而是反映当前阶段的首要矛盾是责任与证据，而非技术路线的选择权；国际对齐（73%）居中，与 Q11 中跨境合规痛点（50%）相互呼应，提示标准设计需预留跨境互认的接口。

Q26. 面对成熟可信评估/认证服务的可能态度

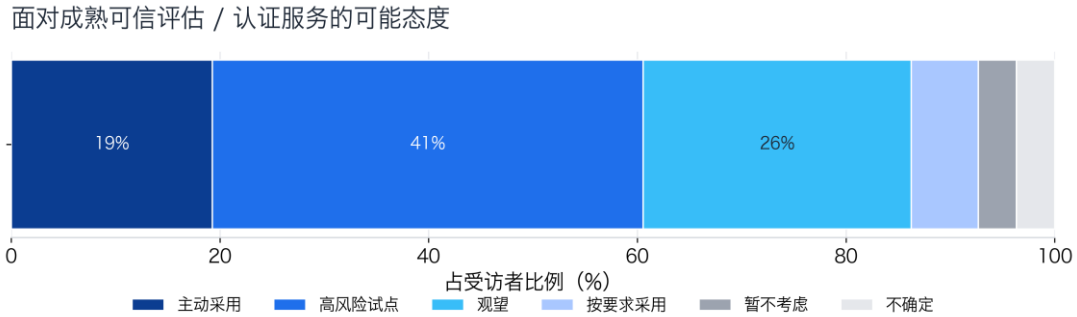
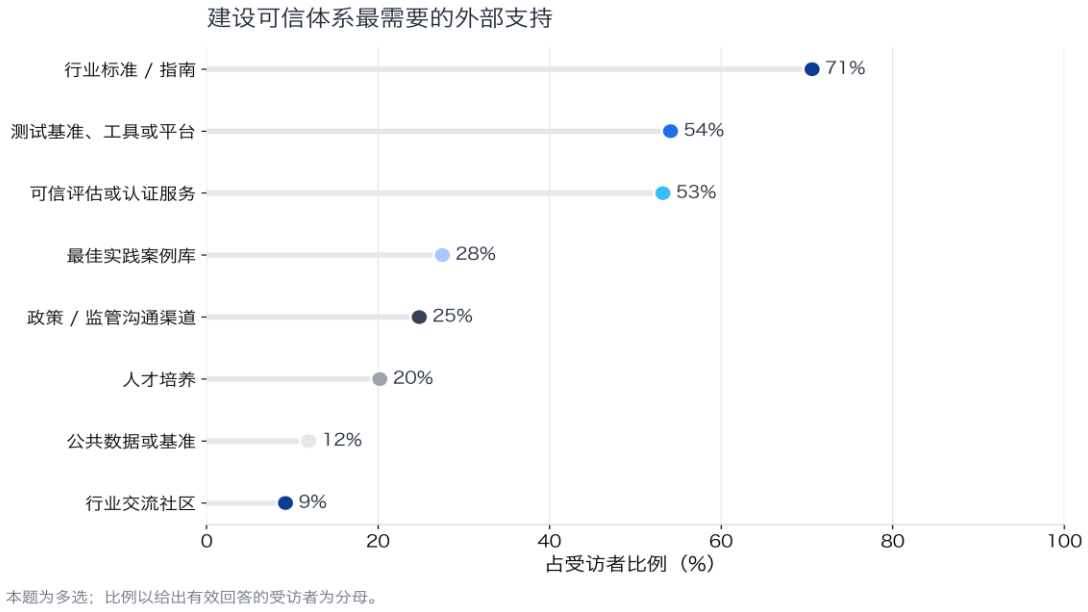


图 26. 最高项为“高风险试点”，占 41.3%。

解读：面对成熟的评估认证服务，41%愿意先在高风险或关键场景试点，19%会主动采用并设为上线或合作的前置条件，积极态度合计 60%；26%选择观望，等

待行业共识或监管要求。试点先行的路径与主报告 5.3 节的场景化设计一致：评估参数需要试点数据标定，高风险场景对信任证据的需求最强、付费意愿也最明确，从高风险场景切入可以同时获得数据与收入。19%的“前置条件”群体尤其值得关注：他们已经开始把第三方评估嵌入自身的采购与上线流程，这类组织的存在使认证服务具备了网络效应的起点——每多一个要求认证的交易对手，被认证的价值就上升一分。观望群体约占四分之一，其转化取决于首批试点的可验证成效；按照主报告 5.3.3 节，试点设计应包含明确的阈值标定与降级恢复通道，使观望者能够看到可复制的完整闭环，而非一次性的演示。

Q27. 建立可信体系最需要的外部支持

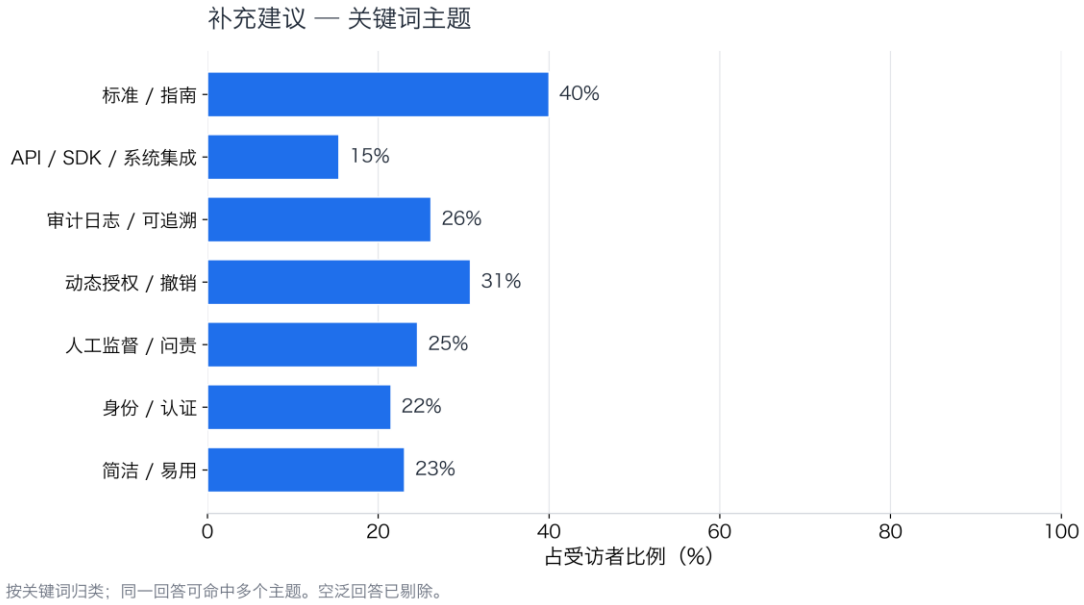


本题为多选；比例以给出有效回答的受访者为分母。

图 27. 多选最高项：“行业标准/指南”70.6%；“测试基准、工具或平台”54.1%；“可信评估或认证服务”53.2%。多选比例合计可超过 100%。

解读：最需要的外部支持依次为行业标准与指南（71%）、测试基准与工具平台（54%）、评估认证服务（53%），其后是最佳实践案例库（28%）和监管沟通渠道（25%）。标准、工具、服务三类支持的选择率均超过半数，说明仅有文本标准不足以落地：组织同时需要可执行的测试手段来验证合规，以及第三方服务来降低自建成本。这与主报告 5.2.5 节关键门槛的设计相呼应——门槛的有效性依赖独立的测试与核验能力，没有工具平台支撑的门槛会退化为自我声明。三类高票支持之间还存在依赖关系：标准定义“测什么”，工具平台解决“怎么测”，认证服务回答“谁的结果可信”，三者构成完整的市场基础设施链条，任一环节缺位都会抬高整体采用成本。监管沟通渠道（25%）选择率偏低，与 Q26 中“等待监管要求”的观望群体规模相当，反映多数组织当前更关注市场层面的互信，而非监管审批。

Q28. 补充建议-关键词



按关键词归类；同一回答可命中多个主题。空泛回答已剔除。

图 28. 最高频主题为“标准/指南”（40.0%）。

解读：开放建议的关键词主题集中在标准与指南、API/SDK 集成、审计日志与可追溯、动态授权与撤销、人工监督与问责五个方向，与前述量化结果一致。建议内容以工程化诉求为主，较少涉及概念讨论，反映受访者已经把 KYA 理解为需要落地的系统而非原则声明。主题分布还显示，授权的生命周期管理（授予、调整、撤销）受到持续关注，与主报告 3.3.2 节“授权即数据”的方向相符：授权需要成为可传递、可验证、可撤销的对象，而非一次性的配置项。五个主题与 Q24 的框架优先级、Q27 的外部支持需求形成三角互证：标准解决“依据什么建”，API/SDK 解决“如何低成本接入”，日志与追溯解决“证据从哪来”，动态授权解决“权限如何随信用调整”，人工监督解决“例外由谁处置”。开放题与选择题结论一致，说明量化结果并非选项引导的产物，需求结构具有稳健性。

3.5 可选模块：BNPL/消费金融

该模块仅由熟悉相关业务的受访者填写；解读时应注意样本量小于全样本。

Q29. BNPL/消费金融中智能体的使用程度

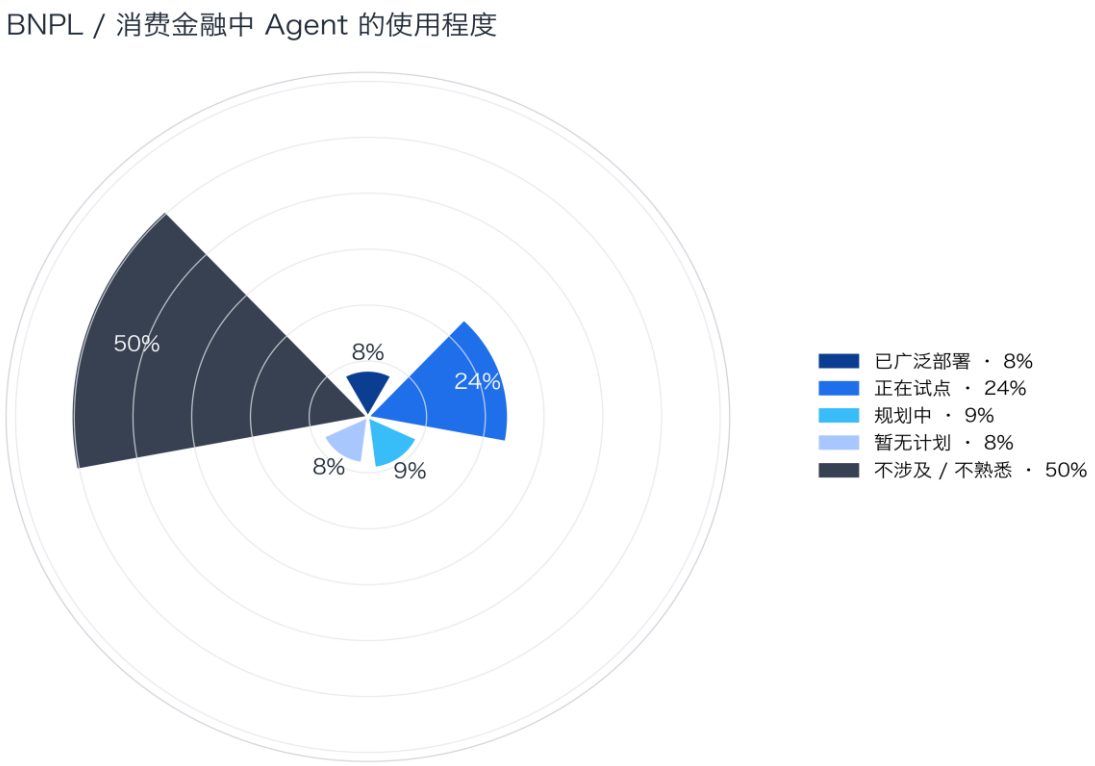
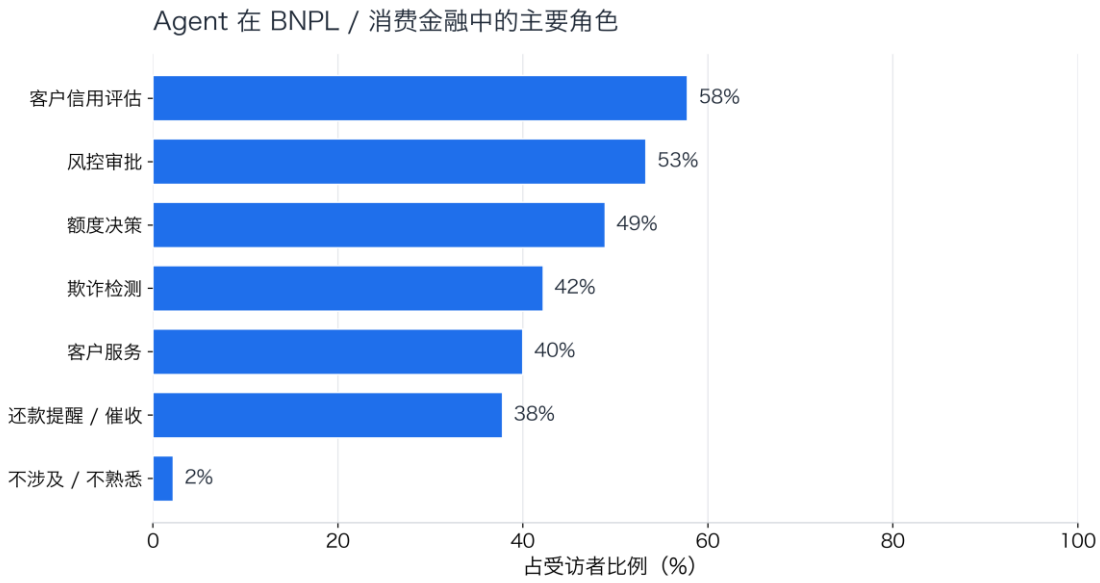


图 29. 最高项为“不涉及/不熟悉”，占 50.5%。

解读：在填写 BNPL/消费金融模块的受访者中，处于试点阶段的占 24%，已广泛部署的占 8%，另有 9%在规划中、8%暂无计划，整体处于试点为主、规模化部署尚少的阶段。消费金融是资金直接相关场景，对信任证据的要求高于一般任务。当前渗透水平与主报告 5.3.2 节对先用后付场景的分析一致：该场景的评估设计需要同时覆盖效率指标与信任指标，单一维度的优化不足以支撑规模化。试点占主导的结构还说明，行业正处在主报告 5.3.3 节所述的参数标定阶段：通过率、额度、干预阈值的合理区间需要用试点数据确定，而非从既有信贷模型直接平移——智能体审批的错误模式与人工审批不同，历史风控参数不能直接复用。8%的广泛部署者是重要的观察样本：其阈值设定、人工复核触发条件和不良处置方式，将为后续进入者提供事实上的行业基准。

Q30. 智能体在 BNPL/消费金融中的主要角色



本题为多选；比例以给出有效回答的受访者为分母。

图 30. 多选 最高项：“客户信用评估” 57.8%；“风控审批”53.3%；“额度决策”48.9%。多选比例合计可超过 100%。

解读：智能体在 BNPL/消费金融中承担的角色已深入核心授信环节：客户信用评估（58%）、风控审批（53%）、额度决策（49%）居前三位，欺诈检测（42%）、客服（40%）、催收提醒（38%）次之。智能体的输出直接进入资产质量形成过程，其错误会沿主报告 1.2.2 节描述的传导路径转化为实际资金损失，且具备规模化复制特征——一个错误的审批策略会在短时间内被复制到成千上万笔申请上。因此该场景没有低风险缓冲区：承担信用评估、额度与审批职能的智能体应当适用最高等级的评估与授权要求。对照主报告 3.4.3 节，金融信贷场景的三层评分权重为身份 15%、能力 40%、信用 45%，信用维度权重最高；这意味着在该场景中，智能体的历史履约记录比其模型能力演示更能预测未来表现。催收提醒（38%）的渗透还提示，智能体已出现在贷后管理环节，其行为的合规性与可追溯性同样属于信用评估范围。

Q31. BNPL/消费金融中智能体与人工决策对比

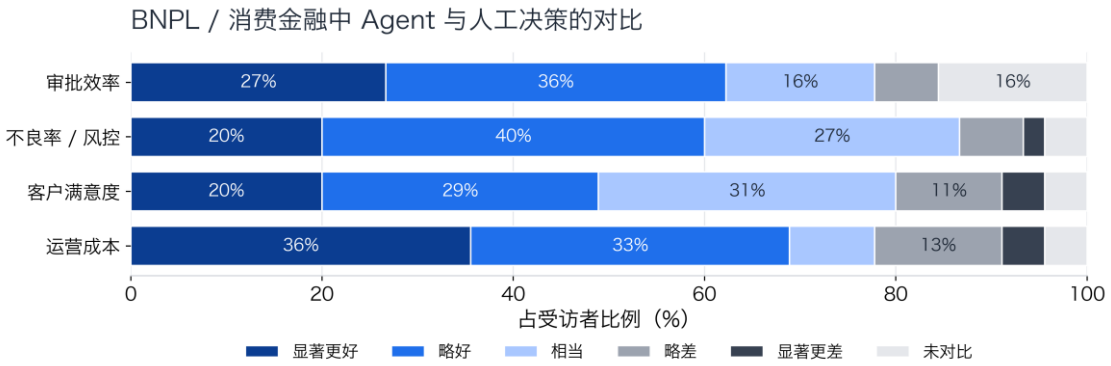


图 31. 按“显著更好+略好”合计占比排序，运营成本的正面评价最高。

解读：智能体与人工决策的对比中，运营成本（69%认为更好）和审批效率（62%）的优势最明确，不良率与风控（60%）同样占优；客户满意度仅 49%认为更好、16%认为更差，是四项中分歧最大的维度。效率与成本红利已经显现，体验与信任仍是短板。按照主报告 5.3.2 节的场景逻辑，先用后付的评估需要效率与信任并重：审批越快，越需要可解释的理由、可复核的通道和可追责的记录来维持消费者信任。满意度的分歧结构值得细读：16%认为更差的比例在四项中最高，说明部分组织的智能体审批已经引发了可感知的客户摩擦——拒贷理由不清、申诉无门、纠错迟缓都会直接损害信任。主报告 5.3.2 节还提醒，98%的审批完成率不等于回款概率，效率指标不能替代资产质量指标；同理，成本与速度的优势也不能替代满意度所反映的长期客户关系价值。

Q32. 消费者对智能体审批的信任/接受度

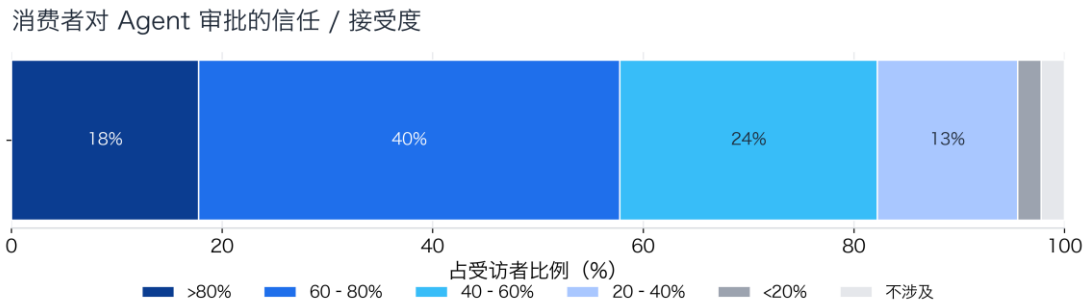


图 32. 最高项为“比较信任 60 - 80%”，占 40.0%。

解读：消费者对智能体审批的信任度：接受度 60%以上的合计 58%（非常信任 18%、比较信任 40%），中性 24%，不信任合计 16%，整体态度中性偏正面，需求侧基础存在。24%的中性群体是信任建设的主要转化对象，其态度对审批透明度、纠错效率和争议处理体验最为敏感。结合 Q33 的结果，可解释审批理由与征信体系对接是提升该群体接受度的两个直接杠杆。从主报告 1.2.4 节的框架看，消费者信任度是先用后付场景市场扩张效应的直接决定变量：信任度每上移一个档位，可服务的客群边界就外扩一圈，58%的接受度意味着当前智能体审批的可

行市场仍受限于信任边界而非技术能力。16%的不信任群体也不应被忽视：在社交媒体放大效应下，少数负面体验的传播成本极低，其声誉影响可能超过其人数占比，这为主报告 5.2.3 节强调的声誉抗操纵与舆情监测提供了场景化的注脚。

Q33. BNPL/消费金融中智能体可信的关键要素

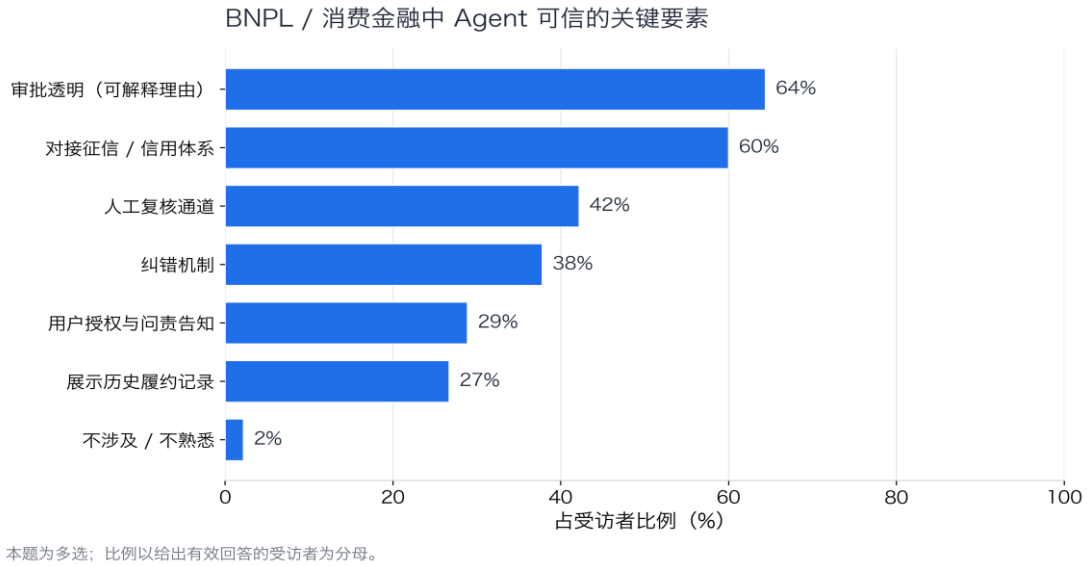


图 33. 多选 最高项：“审批透明（可解释理由）”64.4%；“对接征信/信用体系”60.0%；“人工复核通道”42.2%。多选比例合计可超过 100%。

解读：建立智能体可信度的关键要素中，审批透明与可解释理由（64%）和对接征信/信用体系（60%）居前，其后是人工复核通道（42%）、纠错机制（38%）、授权与问责告知（29%）、履约记录展示（27%）。前两项分别对应“决策可理解”和“信用有外部锚点”，构成消费金融 KYA 的两个支柱：可解释性解决单笔决策的信任问题，征信对接解决智能体作为决策主体的信用来源问题。人工复核与纠错的选择率虽然居中，但二者属于主报告 5.2.5 节关键门槛性质的要求：不是加分项，而是准入条件——没有复核通道的审批系统，其透明度承诺无法兑现。授权与问责告知（29%）排名靠后，反映当前实践更关注决策结果而非决策过程的告知；但按照主报告 3.3.2 节“授权即数据”的逻辑，告知记录本身就是争议发生时的关键证据，其建设成本远低于事后举证的成本。

3.6 可选模块：中小企业/小微客户

该模块用于观察 SME/小微客户场景下的采用、顾虑与简化方向。

Q34. SME/小微客户服务占比

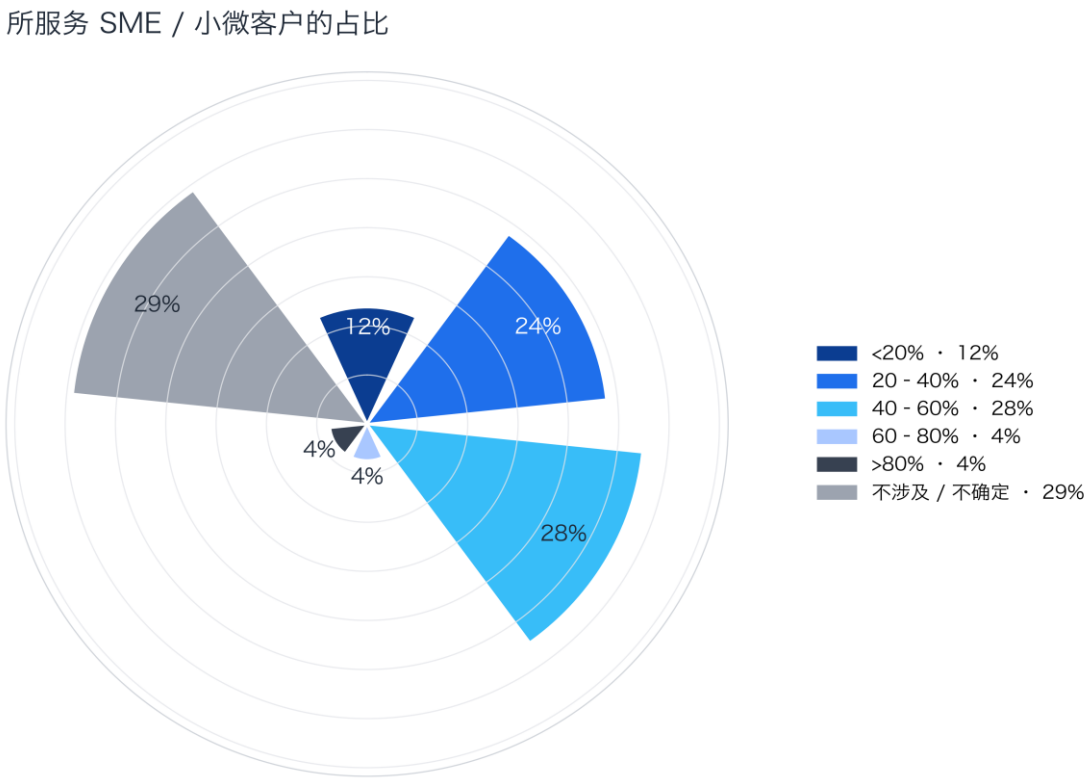


图 34. 最高项为“不涉及/不确定”，占 29.4%。

解读：受访者所在组织服务的 SME/小微客户占比分布较为均匀：40 - 60%区间占 28%，20 - 40%占 24%，20%以下占 12%，超过 60%的合计仅 7%。SME 是普遍存在但通常不占主导的客户群体。这一结构意味着 SME 信用需求容易被以大型客户为中心的标准化产品忽略；而 SME 正是信用信息不对称最严重的群体——财务报表不规范、抵押物不足、信用历史短，传统征信覆盖有限，其智能体使用场景的信用服务存在明确的需求缺口。从主报告 3.2.4 节 α + β 混合信用的视角看，SME 恰恰是混合信用最能发挥价值的客群：主体信用（α）薄弱时，可以依靠具体任务的履约记录（β）逐步积累信用，而智能体化服务天然产生结构化的履约数据。占比分布的均匀性还说明，SME 信用方案不是少数机构的边缘业务，而是多数服务机构都需要面对的通用能力。

Q35. SME/小微客户使用智能体的大致比例

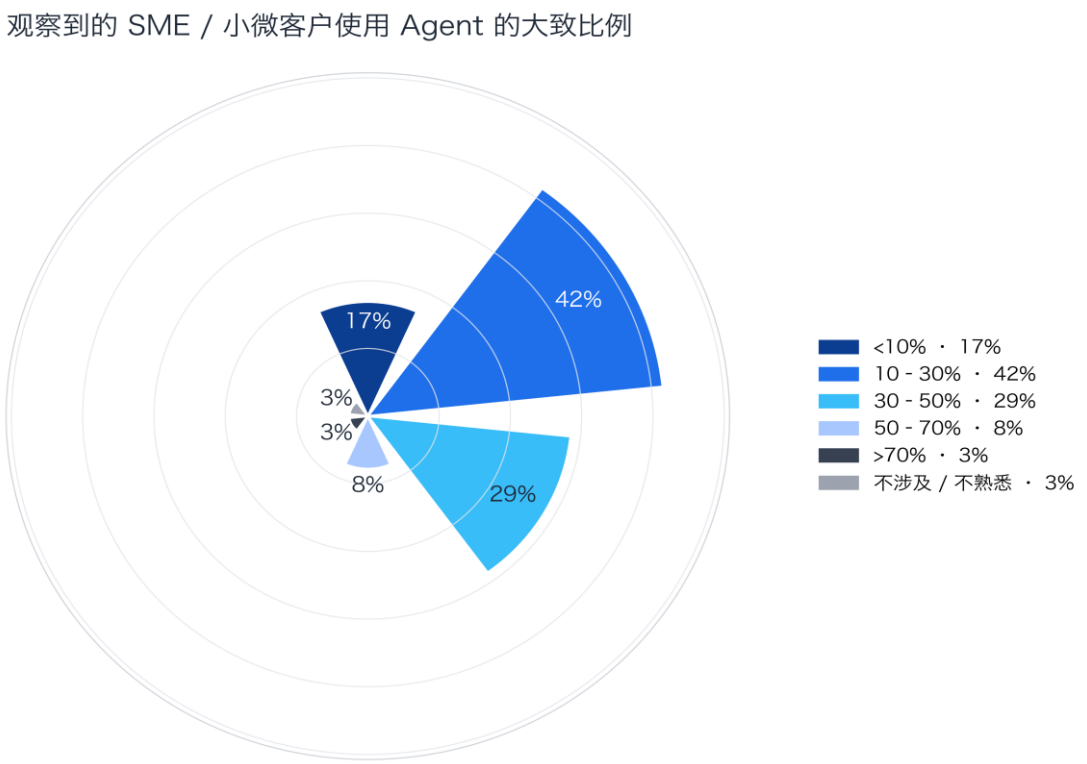
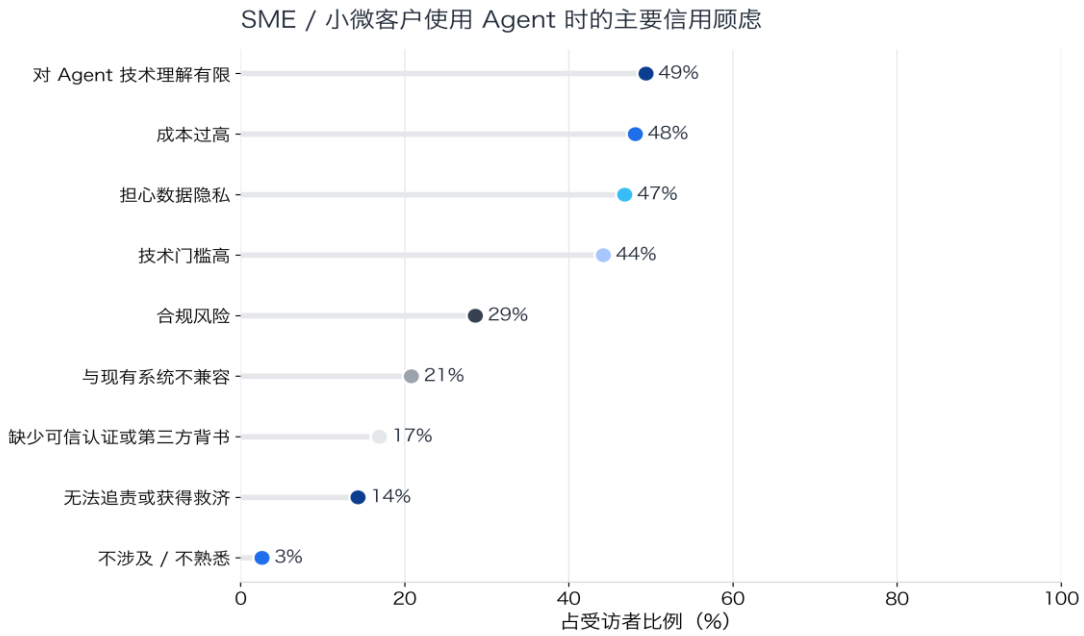


图 35. 最高项为“10 - 30%”，占 41.6%。

解读：在涉及 SME 客户的受访者中，小微客户的智能体使用率集中在 10 - 50%区间（10 - 30%占 42%，30 - 50%占 29%），超过 50%的仅 10%，渗透仍处于早期阶段。渗透率低与 Q36 的成本和认知门槛直接相关：当理解成本与实施成本都高时，小微企业会推迟采用，形成“低使用—低数据—低信用”的循环。从冷启动的角度看，这一阶段更需要平台侧提供可复用的身份、授权与日志能力，避免每家小企业重复建设——这正是主报告 5.3.4 节起步路径的核心主张：先落实身份归属、真实履约、授权执行、异常处置四项最低要求，让关键事实从第一天起就被记录。10 - 30%是最集中的区间，说明多数 SME 的智能体使用刚越过尝鲜阶段；这一时期形成的记录质量，将决定其未来申请信用服务时的起点，平台在这一阶段介入记录基础设施的价值最大。

Q36. SME/小微客户使用智能体时的主要信用顾虑



本题为多选；比例以给出有效回答的受访者为分母。

图 36. 多选 最高项：“对智能体技术理解有限”49.4%；“使用成本过高”48.1%；“担心数据隐私”46.8%。多选比例合计可超过 100%。

解读：SME 使用智能体的信用顾虑前四位是技术理解有限（49%）、成本过高（48%）、数据隐私担忧（47%）和技术门槛（44%），合规风险（29%）与系统不兼容（21%）次之。认知与成本构成双重门槛，且高于制度性顾虑——这与大型企业样本中问责与合规痛点居前的结构明显不同，说明 SME 市场的主要矛盾还不是治理精细度，而是可及性。这一排序对产品设计有直接含义：SME 版本的信用方案应先做减法，通过模板化和平台化降低理解与实施成本；需要强调的是，简化的是流程而非门槛，已开放动作的控制要求不能豁免——主报告 5.3.4 节的起步清单正是按此逻辑设计：身份归属、真实履约、授权执行、异常处置四项要求不可裁减，但每一项都可以用平台默认配置实现。隐私担忧（47%）接近半数，提示平台化方案必须同时解决数据托管的信任问题，否则成本下降也无法换来采用。

Q37. 面向年收入 15 万美元以下小微企业的简化信用方案供给

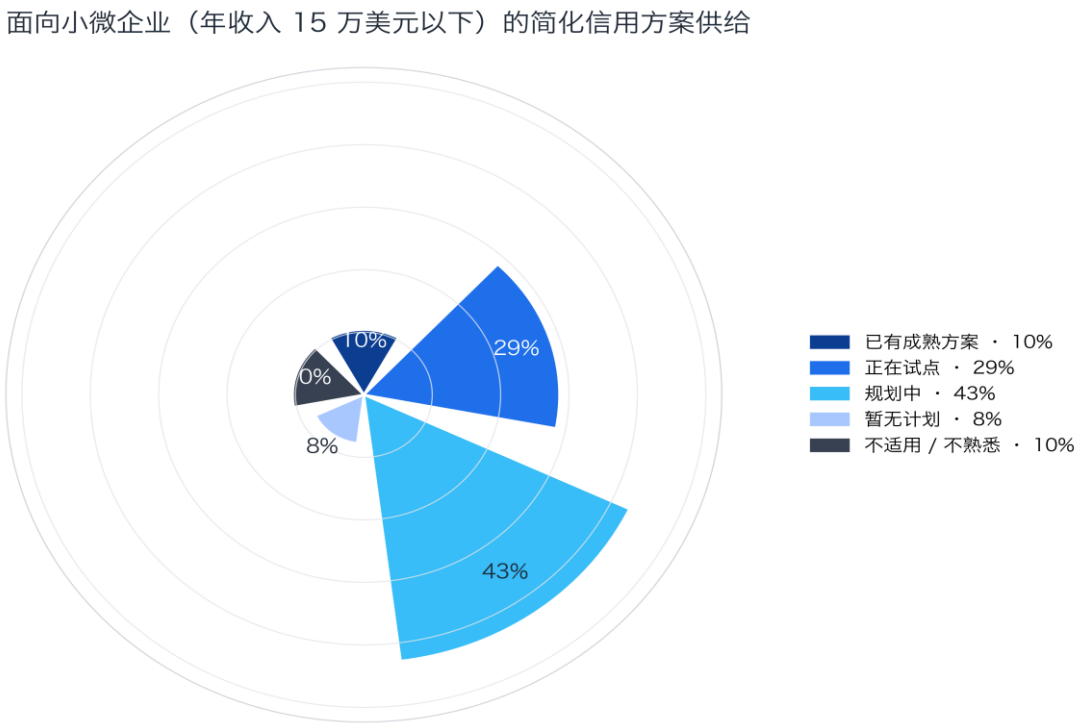
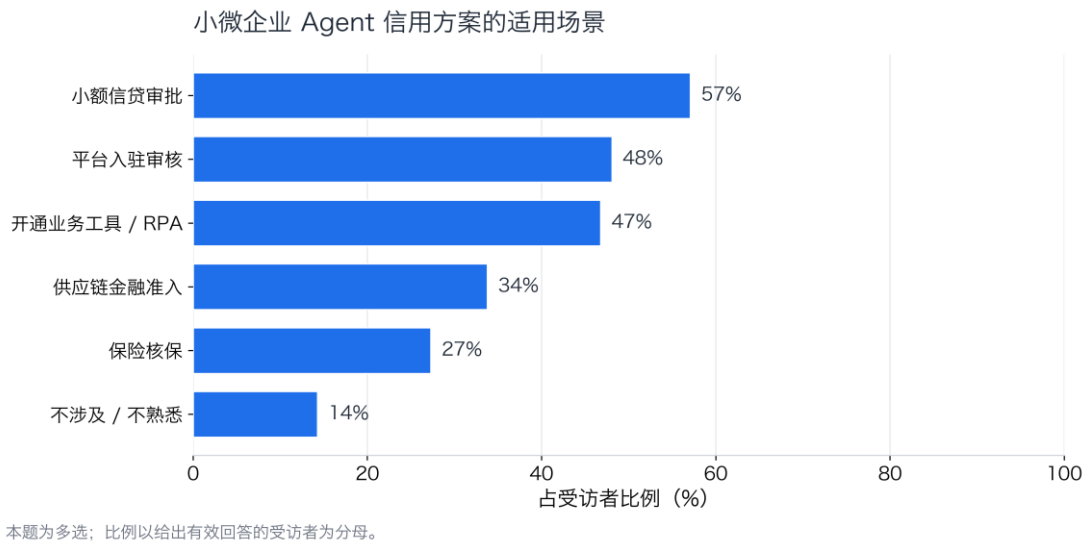


图 37. 最高项为“规划中”，占 42.9%。

解读：面向年收入 15 万美元以下小微企业的简化信用方案中，处于规划中的占 43%，试点中占 29%，已有成熟方案的仅 10%，另有 10%不适用或不熟悉，供给侧明显不足。小微企业的信用需求长期存在但服务成本高：单笔金额小、尽调成本刚性、违约处置不经济，传统模式难以覆盖；智能体化服务若能复用平台既有记录，有机会把边际服务成本降到可商业化的区间。当前 10%的成熟供给率说明该市场仍处于早期，先入者可以参与定义事实标准。从主报告 3.2.5 节冷启动四路径看，担保额度与主体锚定是该客群的现实起点：平台以既有交易数据为小微主体提供初始信用锚，智能体的任务履约记录再逐步积累为独立信用。43%的规划占比预示着未来一到两年供给将快速增加；规划期正是嵌入记录标准与授权模板的最佳窗口，一旦方案上线再补记录机制，改造成本会显著上升。

Q38. 小微企业智能体信用方案的适用场景

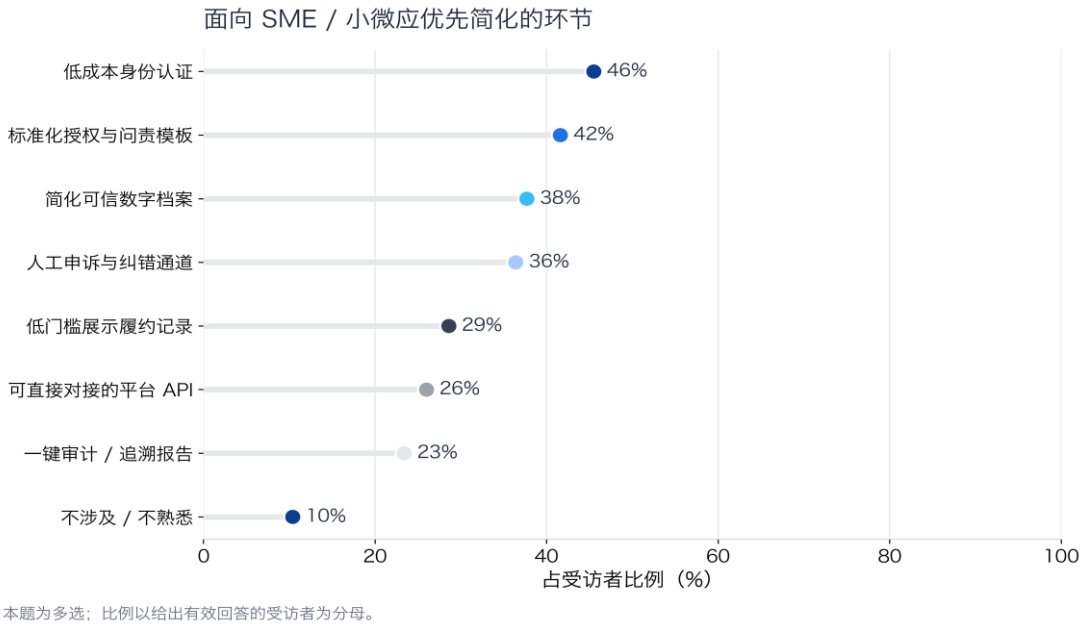


本题为多选；比例以给出有效回答的受访者为分母。

图 38. 多选 最高项：“小额信贷审批”57.1%；“平台入驻审核”48.1%；“开通业务工具/RPA”46.8%。多选比例合计可超过 100%。

解读：小微智能体信用方案的适用场景中，小额信贷审批（57%）、平台入驻审核（48%）、商业工具与 RPA 开通（47%）居前三位，供应链金融准入（34%）和保险核保（27%）次之。前三位均属于准入类判断：任务边界清晰、结果可验证、单笔风险可控，符合主报告 3.2.5 节沙盒任务的特征，适合作为信用冷启动的首批场景。从准入场景积累履约样本，再向资金类场景扩展，是风险可控的推进顺序。值得注意的是，小额信贷审批虽然涉及资金，但“小额”属性使其单笔损失有界，配合额度控制可以满足主报告 3.3.3 节低风险授权的条件，因此排在首位并不与风险分级原则冲突。供应链金融（34%）与保险核保（27%）相对靠后，主要因为这两类场景涉及多方数据核验与跨组织委托链，按照主报告 5.2.4 节对委托链的检查要求，其信用评估的复杂度更高，适合在单点场景验证之后再行扩展。

Q39. 面向 SME/小微企业应优先简化的环节



本题为多选；比例以给出有效回答的受访者为分母。

图 39. 多选最高项：“低成本身份认证”45.5%；“标准化授权与问责模版”41.6%；“简化可信数字档案”37.7%。多选比例合计可超过 100%。

解读：最应优先简化的环节为低成本身份认证（46%）、标准化授权与问责模板（42%）、简化可信数字档案（38%）、人工申诉与纠错通道（36%）。前三项分别对应 KYA 三维的轻量版本：可验证的身份归属、明确的授权边界、可积累的履约记录，与主报告第三章的三维框架逐项对应，说明 SME 场景需要的不是另一套框架，而是同一框架的低成本实现。与 Q36 的成本和认知门槛结合看，SME 破局的关键不是降低评估标准，而是把三维能力封装为开箱即用的模板，让关键事实从第一天起就被记录——主报告 5.3.4 节的起步清单（哪个智能体、什么任务、何种授权、结果如何、谁干预）正是这种模板的最小形态。申诉与纠错通道（36%）进入前四值得注意：它说明 SME 对“出错后怎么办”的关注不亚于“如何不出错”，纠错通道的存在本身就是信任来源，其建设应与其他三项同步而非滞后。

3.7 规模与性能指标

本章用于形成报告第六章发展指数所需的区间化规模与性能画像；区间为问卷原始设计，不要求精确数值。

Q40. 智能体部署规模

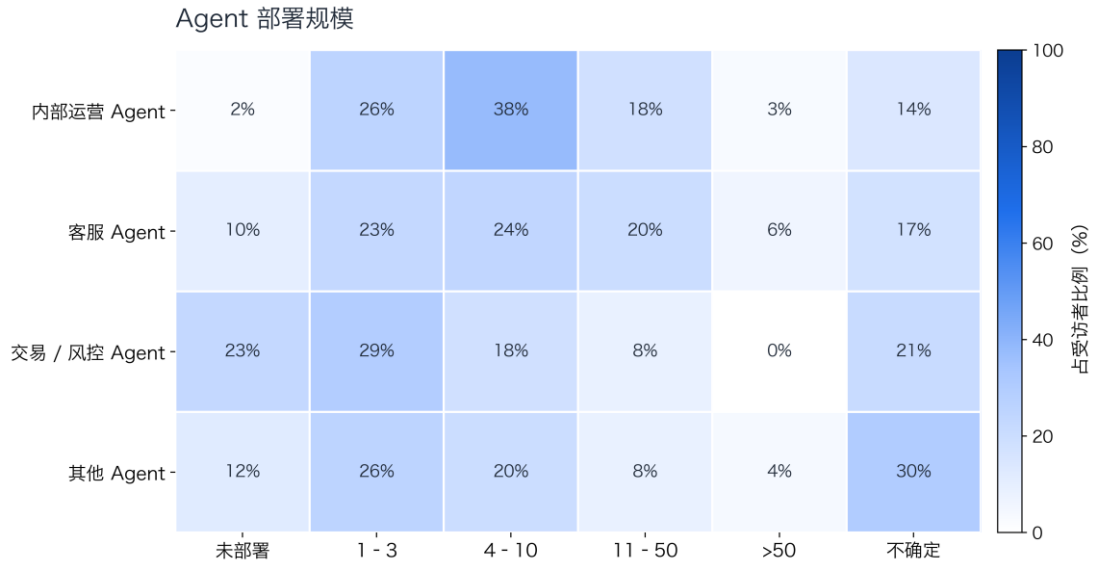


图 40. 颜色越深表示该部署规模区间占比越高；四类智能体的部署规模分布差异明显。

解读：部署规模按智能体类型分化：内部运营智能体部署最广，4 - 10 个区间占 38%，未部署仅 2%；客服智能体次之，11 个以上的占 26%；交易与风控智能体最保守，23%未部署，11 个以上的仅 8%。部署规模随业务风险递增而递减，梯度清晰，与主报告 2.8.3 节动态降级原则背后的风险逻辑一致：组织在低风险场景愿意扩大授权，在资金相关场景保持谨慎。这一分化还印证了主报告 3.3.3 节风险分级矩阵的现实基础——组织事实上已经在按场景风险调整智能体的部署密度，缺的只是将这种隐性实践显性化为可核验的授权记录。从信用体系建设角度看，内部运营与客服智能体的高密度部署意味着大量履约数据正在产生但未被信用化；交易风控场景的低渗透则预示着，一旦信任证据问题得到解决，该场景存在最大的授权扩张空间。规模数据可作为后续发展指数中“应用广度”维度的区间化输入。

Q41.1 智能体性能指标-业务自动化程度

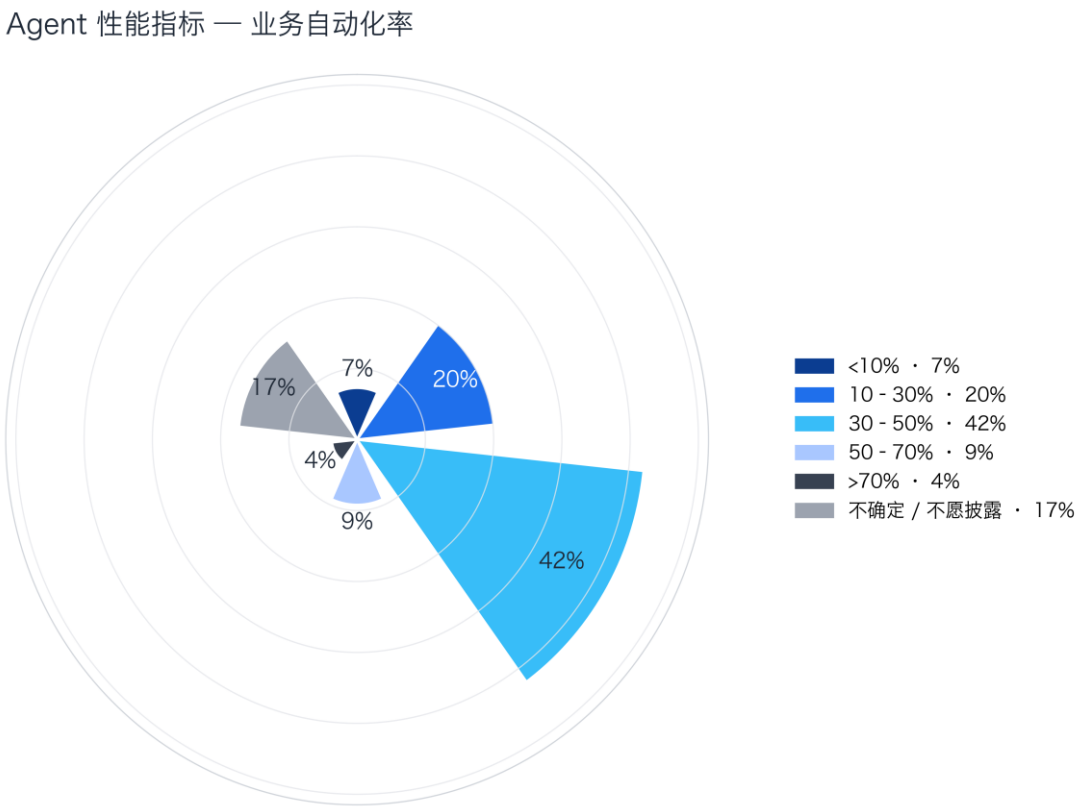


图 41. 最高项为“30 - 50%”，占 42.2%。

解读：业务自动化率集中在 30 - 50%区间（42%），10 - 30%占 20%，超过 50%的合计仅 13%。智能体已承担实质性工作量，但多数组织尚未跨过“半数业务自动化”的临界点。自动化深度决定信用需求的强度：当智能体仅处理边缘任务时，信任问题可以被人工流程吸收；当自动化率超过半数，人工兜底在成本上不再可行，信用体系就从可选投入变为运营必需品。当前分布说明多数组织正处在这一转换的前夕。结合 Q14 的成熟度分布看，30 - 50%的自动化率与 L3 混合信用的占比峰值（38%）在结构上相互吻合：人类仍作为后援接管者深度参与，自动化尚未转化为自治。按照主报告 2.1.2 节的升级闭环，从这一区间继续上移的前提不是更高的模型能力，而是履约记录的积累与授权机制的分级——这正是信用体系介入的位置。

Q41.2 智能体性能指标-决策准确率

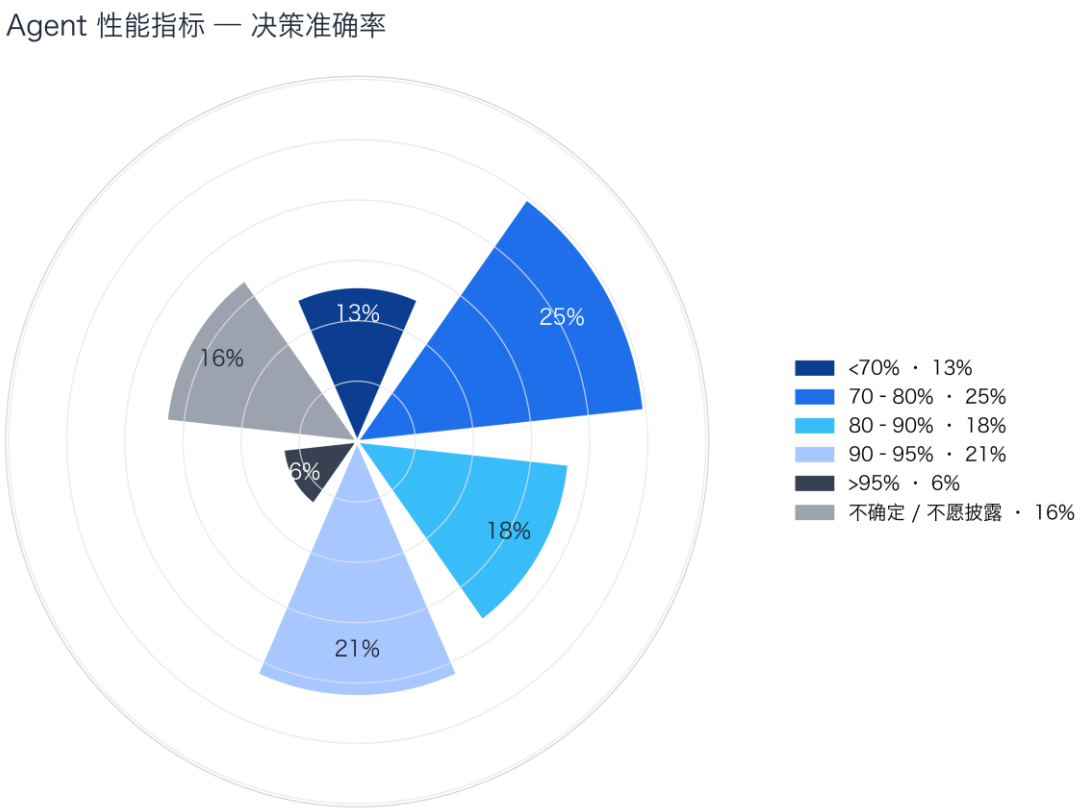


图 42. 最高项为“70 - 80%”，占 24.8%。

解读：智能体决策准确率的分布分散：70 - 80%占 25%，90 - 95%占 21%，80 - 90%占 18%，95%以上仅 6%，另有 13%低于 70%。以资金和高风险场景的要求衡量，当前准确率水平不足以支撑高等级自治，这与 Q14 中 L4 以上占比不足 15%直接相关。需要区分的是，准确率是结果类指标；按照主报告 5.1.2 节的框架，结果指标必须与控制类指标结合观察，短期顺利运行不能替代对拦截与纠错机制的主动验证——一个 95%准确但没有越权拦截能力的智能体，其风险高于 90%准确但控制完备的智能体。分布的分散性本身也是信息：准确率从 70%以下到 95%以上连续分布，说明任务类型、数据质量与工程投入的差异正在拉开组织间的差距，统一评估标准的价值之一就是让这种差距变得可比较、可定价。13%的低准确率群体应被视为高风险场景准入的排除对象。

Q41.3 智能体性能指标-人工干预率

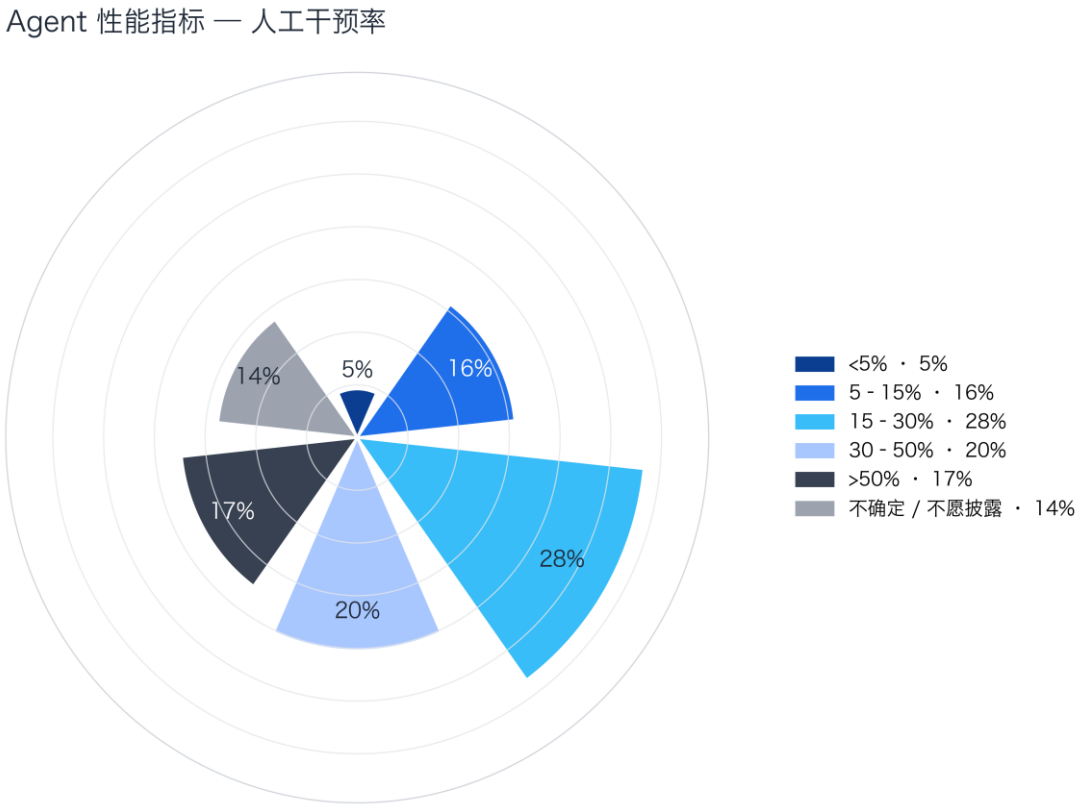


图 43. 最高项为“15 - 30%”，占 28.4%。

解读：人工干预率以 15 - 30%（28%）和 30 - 50%（20%）为主，超过 50%的占 17%，低于 5%的仅 5%。多数智能体实质运行在需要频繁人工介入的状态，与 L1 - L2 等级的特征相符，“名义自治、实际人盯”的现象普遍。人工介入本身不是负面信号——主报告 3.2.3 节把人工介入事件列为最小事件集的一部分，介入的时间、原因与处理结果都是校准授权的有效数据；关键在于介入事件是否被记录并用于校准授权。高干预率叠加低记录率，才是信用建设的真实缺口：组织支付了人工监督的成本，却没有积累下可用于降低未来监督强度的证据。从成本结构看，30%以上的干预率意味着每三笔业务就有一笔需要人工处理，自动化节约的人力被监督人力大量回补，这与 Q8 中“人工确认增加导致效率下降”（38%）的痛点直接对应。把介入事件转化为信用数据，是打破这一成本循环的关键。

Q41.4 智能体性能指标-平均响应延迟

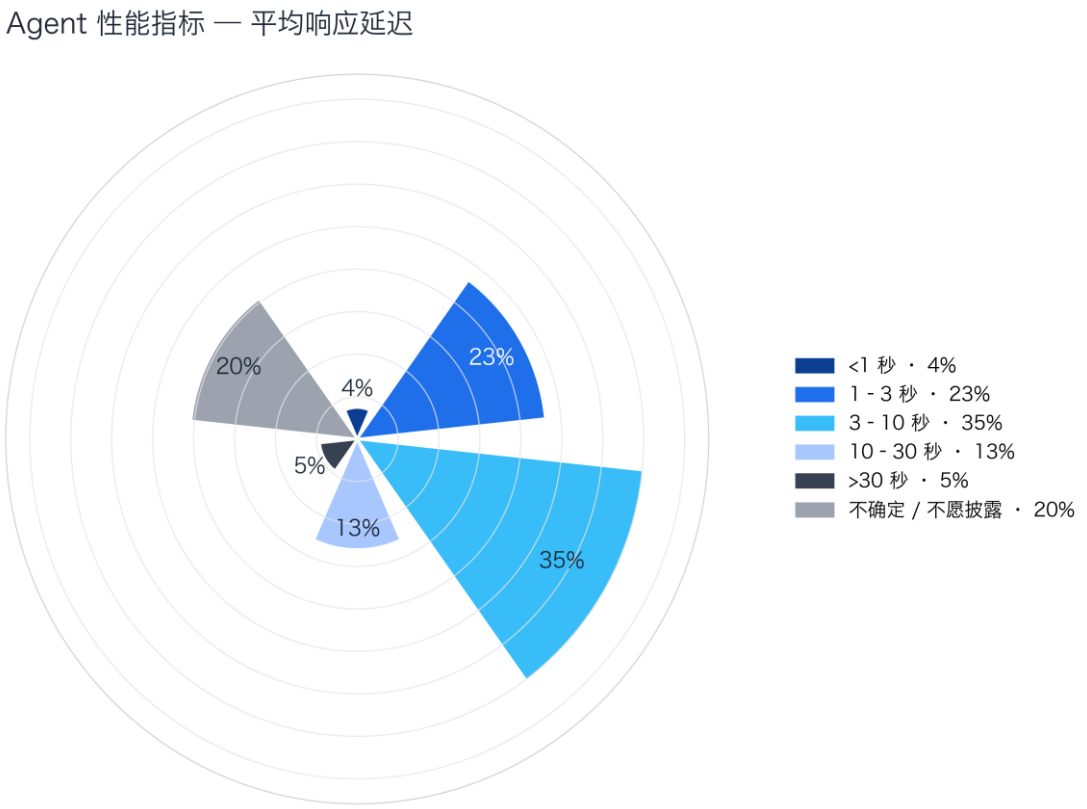


图 44. 最高项为“3 – 10seconds”，占 35.2%。

解读：响应延迟以 3 – 10 秒（35%）和 1 – 3 秒（23%）为主，10 秒以上合计 18%，延迟整体处于可接受区间，性能不是当前信任瓶颈的主要来源。结合 Q41.2 和 Q41.3 的结果看，资源投入的优先级应当是准确率和干预机制，而非响应速度。对评估体系的含义是，性能指标可以维持基础门槛要求，而决策质量、越权拦截和异常处置应占据更高权重——这与主报告 5.1.2 节“结果类与控制类指标结合”的原则一致：延迟属于体验类结果指标，而拦截率、复核触发率属于安全类控制指标，后者对信任的边际贡献更大。3 – 10 秒的主流区间还有一个实务含义：它为人工介入预留了时间窗口。按照主报告 3.3.6 节应急机制的设计，紧急停止与熔断需要在智能体完成动作之前生效，秒级的响应延迟意味着实时拦截在工程上可行，延迟数据因此可以作为授权等级评定时判断“可干预性”的辅助输入。

3.8 后续访谈意愿（匿名统计）

仅展示是否愿意作为案例的匿名分布，不展示任何姓名、组织、邮箱或电话。

后续跟进：是否愿意作为研究报告案例

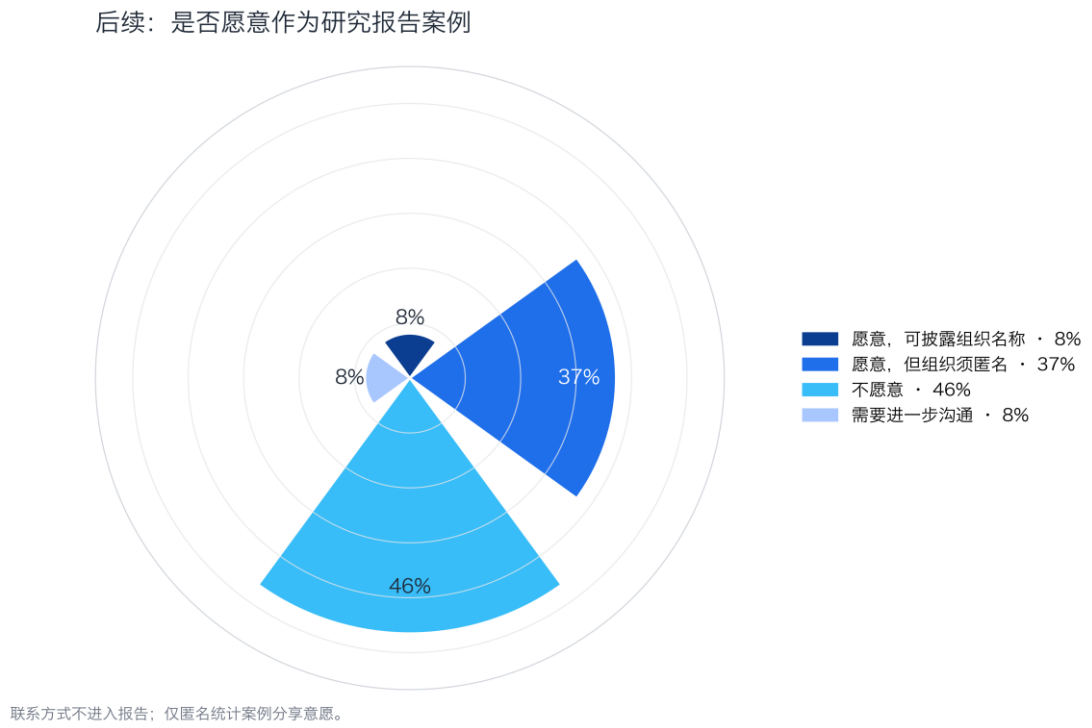


图 45. 最高项为“不愿意”，占 46.5%。

解读：案例分享意愿：46%不愿意，37%愿意但要求匿名，可具名公开的仅 8%，另有 8%需要进一步沟通，企业对智能体实践的公开态度整体谨慎。这一结果从侧面说明，智能体能力目前更多被视为内部运营信息，而非可对外展示的信用资产。信用体系建设的目标之一是改变这一状态：当可信记录可验证、可携带时，主动展示履约能力才会成为企业的竞争策略——正如主报告 2.8.2 节信用飞轮所描述的，可展示的信用记录会吸引更大授权与更多交易，形成正向循环。37%的匿名分享意愿是一个积极的中间信号：企业并不排斥让实践进入行业知识库，排斥的是不可控的声誉暴露；这为行业案例库的建设提供了可行路径——以匿名化、区间化的方式汇集履约与事故数据，先解决“有没有数据”，再随信任环境改善逐步提高署名比例。本报告附录的访谈邀请设计，正是基于这一偏好结构。

4 综合性总结

4.1 对第一章“信任新挑战”的实证支持

- 信任真空的存在性获得量化支持：76%的受访者在过去一年遭遇信任、安全、问责或合规顾虑，仅 2%从未遇到。顾虑的代价首先表现为合规审计压力上升（41%）、项目上线或扩张延迟（40%）和人工确认导致的效率损耗（38%），与 1.2.4 节交易成本框架中的监督执行成本相对应；23%的受访者已遭遇客户或合作伙伴拒绝使用，表明信任缺位已开始产生直接的市场收缩效应。

- 不信任的结构呈“能力—制度”双缺口：数据安全与隐私（76%）、幻觉与不可靠输出（62%）属于能力层面；越权失控（43%）与问责不清（40%）属于制度层面。制度性缺口正是 KYA 相对传统 KYC 需要覆盖的增量：KYC 回答“客户是谁”，而智能体场景必须回答“谁在代表谁行动、行动是否越权、出错由谁承担”。

- 商业化约束已经完成从能力到信任的切换：客户与伙伴信任度低（53%）超过 ROI 测算（34%）与系统集成（35%）成为首要商业痛点。价值创造能力不再是主要约束，信任机制才是。

4.2 对第二章“成熟度模型”的实证支持

- 行业成熟度呈纺锤形分布：L3 占 38%，L1 与 L2 合计 45%，L4 以上不足 15%。多数组织处于人类锚定信用向任务自治的过渡区间，尚未完成从“有人兜底”到“凭记录授权”的转换。

- 性能数据解释了成熟度天花板：决策准确率 95%以上的仅占 6%，人工干预率超过 30%的占 37%。在准确率与干预机制改善之前，L3 向 L4 的跃迁缺乏工程基础，独立信用仍将保持稀缺。

- 升级障碍集中在权限与问责环节：安全风险（60%）、授权边界（51%）、问责不清（46%）居前三位，缺乏外部标准仅占 7%。信用飞轮停滞在“控制建立”环节，约束在内部工程能力而非外部标准供给；率先建成内部授权与问责能力的组织将获得先发优势。

4.3 对第三章“KYA 三维能力”的实证支持

- 三维能力建设不均衡：身份可信各要素实施率处于 60%–69%的窄区间，起步相对均衡；信用可信呈“强记录、弱评分”，日志记录 81%而内部信用评分仅 52%；权限可信呈“强兜底、弱分级”，人工确认 86%而场景化信任等级仅 52%。

- 权限可信被 39%的受访者列为最薄弱环节，居三维之首。能力数据与主观判断一致：行业普遍建立了点状的人工闸门，但缺乏体系化的风险分级授权。权限边界不清时，身份可归因、履约可记录也无法转化为更大授权。

- 问责机制存在制度真空：仅 17%有明确流程，30%无稳定机制。生命周期管理“重上线、轻运营”，定期重估与退役覆盖不足。问责是信用记录的责任锚点，也是动态评估与动态降级的前提，应作为三维协同中优先制度化的环节。

4.4 对第五章“信用评估框架”的需求侧证据

- 标准需求接近共识：88%认为需要统一的评估认证标准，39%认为极其需要。框架优先级为安全合规基线（77%）、可审计可追溯（62%）、可信身份注册（60%），互操作协议（14%）排在末位，先底线可信、后互联互通的渐进路线获得数据支持。

- 设计原则的需求排序与 5.1.2 节一致：问责可追溯（4 - 5 分占 93%）、机器可读（92%）、场景分级（91%）居前三位。责任与证据是当前阶段的首要矛盾，技术路线中立性的优先级相对靠后。
- 采用路径与支持需求明确：60%会主动采用或先在高风险场景试点；最需要的外部支持是行业标准（71%）、测试基准与工具（54%）、认证服务（53%）。标准、工具、服务需要同步推进，门槛式要求必须配套独立的测试与核验能力。

4.5 场景模块结论（BNPL 与 SME）

- BNPL/消费金融：智能体已进入信用评估（58%）、风控审批（53%）、额度决策（49%）等核心环节，输出直接进入资产质量形成过程。与人工相比，成本（69%）与效率（62%）优势明确，客户满意度分歧最大（49%更好、16%更差）。消费者接受度中性偏正面（58%），可解释审批与征信对接是两大信任杠杆，人工复核与纠错属于门槛性质要求。
- SME/小微：渗透率集中在 10 - 50%区间，顾虑以认知（49%）与成本（48%）门槛为主；年收入 15 万美元以下企业的成熟方案供给仅 10%。优先落点是边界清晰的准入类场景（小额信贷 57%、入驻审核 48%）；产品方向是把三维能力封装为轻量模板，即低成本身份认证（46%）、标准化授权与问责模板（42%）、简化可信数字档案（38%）。

4.6 综合判断与行动建议

- 第一优先级是权限可信的工程化：授权边界、动态限额、人工确认阈值、紧急停止与越权阻断应成为可验证、可审计的默认能力，并逐步从点状人工闸门过渡到场景化分级授权。
- 信用可信需要从记录层走向评估层：先统一关键决策、输入输出、工具调用、异常事件与人工介入的记录 schema，再建设内部评分与跨系统集成能力。
- 标准与评估设计应坚持问责可追溯、机器可读、场景分级与渐进演化；商业化从高风险场景试点切入，同步建设标准、测试基准与认证服务三类外部支持。
- 场景策略上，BNPL 优先补齐可解释审批、人工复核、纠错机制与履约记录展示；SME 优先提供低成本身份认证、简化可信数字档案与标准化授权问责模板。
- 长期看，准确率与干预机制的改善将打开 L3 向 L4 跃迁的空间；当可信记录可验证、可携带时，智能体信用将从内部管理工具转变为可对外展示的资产。当前仅 8%的组织愿意具名分享案例，这一比例的后续变化可以作为行业信任成熟度的观察指标。

附录 开放建议主题说明

Q28 为开放题。报告不逐条引用可识别个人风格的原文，仅做关键词主题统计；同一回答可命中多个主题，空泛回答（如 N/A、none）已剔除。

代表性建议（已做匿名化筛选，仅保留不指向个人身份的内容）：

-必须有多套公司规范，并明确相应责任人；

-智能体需要记录推理步骤，以便清楚看到某一动作为何被执行；

-标准应保持简单、可落地；需要基础、开放的 API，让不同平台都能低成本核验智能体，而不增加过多负担；

-应明确披露智能体的身份、数据来源、权限等信息；当行为偏离预期时，还要有清晰的权限撤销方式；

-把智能体身份当作短期令牌来管理，不要给智能体静态 API 密钥或长期凭证；

-需要人才培养，以及企业方面更多的支持；

-需要从静态 API 密钥，转向动态行为治理、密码学身份与持续审计；

-需要数字证书、集中式注册，以及令牌校验。

版权与免责声明

本报告为香港大学 KYA 研究课题组的研究成果，报告中的判断与观点仅代表课题组立场，不构成任何机构的产品承诺、商业建议或投资建议。

报告引用的产业数据与案例信息来自公开资料、企业访谈与课题组整理，均已尽可能注明出处；如有疏漏或出入，欢迎指正。

报告内容将持续修订。未经课题组许可，任何单位或个人不得以任何形式转载、摘编报告内容；引用请注明出处。