香 港 大 學
THE UNIVERSITY OF HONG KONG

KYA: Know Your Agent

Research Report on the Agent Credit System

The University of Hong Kong KYA Research Group

November 2026



Editorial Board Members

Yang Wang
Chen Lin
Kabiné Komara
Kani Chen
Anthony Yang
Hao Chen
Lingyun Gu

Project Team Members

Lizhang Miao
Liang Bai
Yiran Liu
Yifan Cao
Yuzi Han
Wenting Ma
Xiaowei Tang
Pan Hu
Zhuoer Zhang

We would like to thank Ant Group for its support of the KYA research.

Contents

Introduction 1
Chapter 1: From KYC to KYA—New Trust Challenges in the Agent Era 4
1.1 The Rise of the Agent Economy and the Trust Vacuum 4
1.2 The Economic Costs of Trust Deficits 9
1.3 From KYC to KYA 24
Chapter 2: The Agent Maturity Model (L0–L5) 32
2.1 The Logic and Assessment Method of the Maturity Model 32
2.2 L0 Non-Agent: Rule-Execution Baseline 35
2.3 L1 Human Approval: The Agent Receives Advisory Authority 36
2.4 L2 Human-Anchored Credit: The Agent Receives Restricted Execution Authority 39
2.5 L3 Hybrid Credit: The Agent Gains Task Autonomy 42
2.6 L4 Independent Credit: The Agent Gains Limited Economic Agency 45
2.7 L5 Agent Economy: From Individual Maturity to Network Maturity 48
2.8 Upgrade Paths, the Credit Flywheel, and Dynamic Downgrading 51
2.9 Chapter Summary 53
Chapter 3: The KYA Framework—Three Dimensions of Agent Credit 55
3.1 Identity Trustworthiness (Establishing Agent Identity) 55
3.2 Creditworthiness (Performance History and Credit Assessment) 60
3.3 Authorization Trustworthiness (Scope of Authority and Operational Boundaries) 66
3.4 Coordination Across the Three Dimensions 70
Chapter 4: Practical Validation 75
4.1 Ant Digital Technologies: Exploring KYA Through an “Agent Super Factory”. 75
4.2 Stripe: Agent-Economy Strategy and KYA Practices 80
4.3 Sesame Credit: Exploring KYA and Trust Infrastructure for Agents 87
4.4 IceKredit: Building Financial-Grade Trustworthy Agents 91
4.5 Ryt Bank: Agent Trust Practices at the World’s First AI-Native Bank 96
4.6 Yinghe Zhilin: Campus Agent Applications and a Localized Trust Foundation 99
4.7 Linyang Energy: Building Trust Mechanisms for Agents That Can Be Relied Upon 104
4.8 Mediwelcome: Embedding Physicians’ Professional Credit in Medical Agents 110

4.9 GCSA: Security and Adversarial Governance for the Agent Ecosystem 116
Chapter 5: Agent Credit Assessment Framework 123
5.1 Methodology and Design Principles 123
5.2 Baseline Indicator System 127
5.3 Recommendations by Scenario 138
Chapter 6: Agent Development Index 144
6.1 Basic Principles 144
6.2 Indicator System 145
6.3 Index Results and Presentation 151
6.4 Interpretation of the Index 154
6.5 Implications for KYA Development 157
Public Data Sources for the Index 160
Chapter 7: Outlook—From the Agent Economy to an Era of Trust Capital 162
7.1 Four Stages of the Agent Economy: From Task Execution to Economic Value Creation 162
7.2 Era-Defining Opportunities: From Building Agents to Building the Agent Economy 164
7.3 Trust Capital: Capital Formation and Economic Growth in the AI Era 167
7.4 How to Create and Accumulate Agent Trust Capital 171
References 176
Appendix : Agent Credit System (KYA) Survey 178
Abstract 178
1. Executive Summary 179
2. Data and Methods 179
3. Data Analysis 179
4. Integrated Conclusions 222
Appendix： Notes on Open-Ended Suggestion Themes 225
Copyright and Disclaimer 227

Introduction

Without KYA, there can be no Agent economy. This is the conclusion we have reached through extensive research and in-depth investigation.

The concept of AI agents and the technical exploration of them have a long history. Agents became usable in 2025 and began to take off in 2026. As they enter commercial transactions, the vast potential of an "Agent economy" is gradually coming into view. Yet a critical gap separates agent applications from an Agent economy: trust in agents. KYA (Know Your Agent) is needed to bridge it.

The relationship between artificial intelligence and real-world economic activity is undergoing a structural change. AI is no longer merely software that people use; it is becoming a digital actor capable of carrying out tasks, allocating resources, and participating in transactions.

This change is unfolding rapidly across business, finance, consumption, payments, and public services. More companies are deploying agents that can autonomously perform specific tasks. Financial institutions are exploring agentic payments and autonomous financial services. Payment networks are building new infrastructure for machine-initiated transactions. Consumers, too, are moving beyond using AI to obtain information and are beginning to rely on it as an agent for shopping, travel, work, and everyday services.

At the same time, questions of identity, authorization, security, accountability, and credit are emerging as agents act autonomously.

As agents enter real economic activity, the decisive questions are shifting from "What can an agent do?" to "What should it be allowed to do?" and "Will other parties be willing to work and transact with it?" For the Agent economy to develop sustainably, reliably, and at scale, it needs more than stronger models and better infrastructure. It also needs a trust mechanism that can identify agents, establish their ownership and authorization, assess their conduct and creditworthiness, and define the boundaries of accountability.

KYA (Know Your Agent), as discussed in this report, arises from this need. It is not a simple replication of traditional KYC in the agent domain. It establishes identification and trust mechanisms to address the new questions of identity, authorization, credit, and accountability that arise when agents become participants in economic activity.

The principal constraint on the Agent economy is shifting from technical capability to the trust system. Once agents enter payments, trust becomes a commercial issue.

In our survey of industry practitioners (the "Survey Report"; see the appendix), 76% of respondents said that they had frequently (23%) or occasionally (53%) encountered concerns involving trust, security, accountability, or compliance in the past year. Only 2% said they had never encountered such concerns. Trust frictions have become commonplace.

A consumer survey commissioned by Visa and conducted by Harris Poll in the United States from May 26 to 28, 2026, found that 23% of respondents trusted generative AI to handle payments on their behalf. When asked about a specific brand, 61% said they trusted Visa to handle agentic transactions.

A survey of the global financial services industry by the Cloud Security Alliance and Anjuna found that 62% of surveyed institutions had deployed AI agents, and 93% had granted their agents some degree of autonomy. Eighty-five percent of respondents expected AI agents to be able to initiate and execute payments on behalf of consumers, while 65% said this model would require new authorization mechanisms. Separately, Accenture's 2025 Future of Money study found that 87% of financial-institution respondents considered trust the greatest barrier to the adoption of agentic payments.

The demand-side and supply-side evidence points to the same conclusion: agentic payments have clear market demand, but existing identity and authorization mechanisms are not yet sufficient to support their development at scale.

The goal of KYA is to safely unlock the scale of the Agent economy.

The Agent economy needs a machine-verifiable trust mechanism. The commercial value of KYA can be understood at three levels: it enables transactions that otherwise could not take place; it allows transactions to be completed at lower cost; and it makes commercial activity scalable to machine speed and volume.

Viewed across enterprise deployment, financial applications, consumer use, and machine payments, agents are gradually moving from a new form of software capability to actors in economic activity. Companies are deploying agents at scale; financial institutions have granted a substantial proportion of agents some degree of autonomy; payment networks are building new infrastructure for machine transactions; and the number of consumer AI users has reached the hundreds of millions.

Together, these developments show that the Agent economy has moved beyond a technical concept and entered the early stages of industrialization. At the same time, agent-related security incidents, unknown agents, permission-management problems, unauthorized access, and consumers' limited trust in AI payments show that existing identity and trust systems have not fully adapted to this change.

KYA addresses this gap in institutional and commercial infrastructure: the absence of an economic infrastructure that enables agents to be identified, authorized, assessed, trusted, and held accountable by society and markets. The Survey Report finds that 88% of respondents believe the industry needs unified standards for assessing or certifying agent trustworthiness; 39% consider such standards urgently necessary.

Whether the Agent economy can evolve from technological innovation into a new mode of economic organization ultimately depends on whether this infrastructure can grow in step with agent capabilities. Technical capability determines the possible frontier of the Agent economy; the trust system determines its practical frontier. KYA

addresses the critical distance between what is technically possible and what is commercially viable.

This report develops a complete argument—from the problem to a framework—around a foundational question for the Agent economy: "How can an agent be trusted?" Chapter 1 examines the trust vacuum and the economic costs of trust deficits, showing why the industry's bottleneck has shifted from capability to trust. Chapter 2 presents an L0–L5 maturity model centered on delegable autonomy, describing six authorization states from rule execution to networked economic relationships. Chapter 3 translates KYA into three dimensions: identity trustworthiness, creditworthiness, and authorization trustworthiness. Chapter 4 presents cases of agents in economic settings and the KYA initiatives of companies across the ecosystem. Building on these cases, Chapter 5 develops a foundational indicator system comprising three pillars, 13 dimensions, and 65 indicators, together with scenario-based assessment and key thresholds. Chapter 6 proposes an Agent Development Index to track the industry's evolution over time. Chapter 7 considers the commercial and social value of KYA, focusing on the creation and accumulation of Agent Trust Capital.

The report's central proposition is that capability does not equal trustworthiness. Greater model capability does not automatically translate into commercial delegability. Agents can receive economic authority commensurate with their capabilities only when their identities can be established, their performance can be verified, and their permissions can be constrained. A trust mechanism is therefore not an additional compliance requirement attached to agent applications; it is infrastructure that determines the market-size frontier of the Agent economy. The maturity model, three-dimensional framework, and indicator system proposed in this report are intended to provide a public starting point for discussion, testing, and iteration.

This report is an ongoing research effort. The weights and thresholds in the indicator system still need to be calibrated using pilot data; the Agent Development Index needs to be connected to continuous data sources; and the industries and use cases covered by the case studies need to be expanded. The research team will revise the framework as industry practice evolves and welcomes feedback on the report.

Chapter 1: From KYC to KYA—New Trust Challenges in the Agent Era

1.1 The Rise of the Agent Economy and the Trust Vacuum

The surge in generative AI in 2022 introduced the world to large models capable of understanding and generating language in ways that resemble human communication. Since 2025, the defining shift in AI has been the emergence of machines capable of autonomously carrying out tasks (Agentic AI). AI is moving from "answering questions" to "getting work done" , evolving from an information-processing tool into a new kind of actor able to participate in economic activity.

This shift marks the beginning of a new phase in the global AI industry: the Agent era.

1.1.1 AI Is Undergoing a Paradigm Shift

Over the past decade, AI has broadly gone through three stages.

The first was perception AI. Powered mainly by deep learning, AI made major advances in image recognition, speech recognition, and machine translation. Its core capabilities were to "see" , "hear" and "recognize". The large-scale commercialization of computer vision and speech-recognition technologies marked this stage.

The second was generative AI. The launch of ChatGPT (GPT-3.5) in November 2022 marked the point at which large language models began to demonstrate unprecedented capabilities in natural-language understanding and content generation. GPT-4, Claude, Gemini, and other models subsequently pushed the capability frontier further, bringing AI into writing, programming, design, education, research, and other fields. Generative AI greatly increased the productivity of knowledge work and ushered the world into an era of competition among large models.

The core capability of generative AI is still to generate content. Whether it answers a question, drafts a report, or generates code, a human still has to decide what to do next and carry out the final task.

The third stage is the Agentic AI era, which is now evolving rapidly.

The key difference between an agent and a conventional chatbot is whether it has agency—the capacity to act autonomously. An agent can plan around a goal, call external tools, access the internet, run programs, manage files, coordinate multiple tasks, and continually adjust its course of action in response to results. It does more than "offer suggestions"; it can "get work done".

1.1.2 The Agent Industry Is Entering a Period of Rapid Growth

The concept of agents emerged clearly and began to take shape as a field of theory in the 1990s. It was refined through advances in deep learning during the 2010s, and entered a new phase after 2022 with the rise of large models. What has truly

moved agents toward commercialization since 2025 is the concurrent maturation of foundation-model capabilities, tool use, and enterprise demand.

On the one hand, stronger reasoning capabilities in large models support more complex task planning and extended chains of reasoning. On the other, browser control, code execution, API calls, and multi-agent coordination are maturing, giving agents the practical means to perform real-world work.

In the global industry landscape, agents have become one of the most important directions in AI. International research organizations including Gartner, IDC, McKinsey, and Deloitte have all identified agents as a leading AI trend for the coming years.

In its June 2025 report, "Seizing the Agentic AI Advantage," McKinsey observed that enterprise generative-AI deployments were often "shallow": much of the technology remained layered onto existing workflows in the form of copilots and chatbots, mainly supporting information retrieval, content generation, and task assistance, while core business processes remained largely unchanged. McKinsey argued that the transformative potential of Agentic AI lies not in simply inserting agents into existing workflows, but in redesigning workflows around agents and reallocating responsibilities between people and machines. This would turn AI from an "assistive tool" into a business participant able to plan, coordinate, execute, and adapt autonomously.

In the year since McKinsey published that report, agents have rapidly moved from shallow deployment toward deeper integration.

1.1.3 Agents Are Beginning to Enter the Real Economy

The pace at which agents are spreading through industry is markedly faster than that of the internet and mobile internet.

According to Gartner's August 2025 forecast, fewer than 5% of enterprise applications that year would have task-specific agents integrated into them. By the end of 2026, that share was projected to reach 40%. This implies a multiple-fold increase in agent penetration across enterprise applications in roughly one year.

The trend accelerated further in 2026. Gartner's "2026 Hype Cycle for Agentic AI," citing its "2026 CIO and Technology Executive Survey," reported that about 17% of enterprises had formally deployed agents and more than 60% planned to deploy agent systems within the next two years. Such a steep adoption curve is uncommon among emerging digital technologies.

The growth in the number of agents themselves is even more striking. Gartner forecasts that by 2028, a Fortune Global 500 company will use an average of more than 150,000 agents, compared with fewer than 15 in 2025. Based on these thresholds, the average number would increase by more than 10,000 times, or over four orders of magnitude.

Agents are now widely used in software development, enterprise operations, customer service, financial analysis, data analysis, research assistance, cybersecurity, IT operations, and other professional fields.

Software development is widely considered the most mature commercial application of agents. A number of authoritative international benchmarks have shown continued gains in agents' ability to complete complex software-engineering tasks autonomously, including end-to-end coding, debugging, testing, and repair. At the same time, many companies are introducing agents into their development processes to improve engineering productivity and reduce costs.

Beyond software, more companies are testing agents for knowledge-intensive work such as financial analysis, market research, procurement, contract review, and customer support. Traditional software is gradually evolving from a set of "tools" into "digital employees" capable of working autonomously.

1.1.4 The Industry's Bottleneck Is Shifting from Capability to Trust

Despite rapid progress, the agent industry remains at an early stage of commercialization.

Gartner's "Hype Cycle for Agentic AI, 2026" places Agentic AI at the "Peak of Inflated Expectations". This indicates that market expectations are high, while the industry still has many foundational problems to solve.

Those problems increasingly involve trustworthiness as well as model capability.

Can an agent complete a long-running task consistently? Will it hallucinate or make faulty decisions? Can it call external tools correctly? Does it have authenticated identity and controlled permissions? Can it bear legal responsibility? Can it be regulated, audited, and held accountable? How can multiple agents establish trust when working together? After an agent completes a transaction, how can it prove that its conduct was authentic and trustworthy? These questions have become critical bottlenecks to further industry development.

In its "Hype Cycle for Digital Identity, 2026," Gartner gives machine identities, agent identities, and identity management for non-human entities an important place in the digital-identity landscape. Gartner further notes that traditional, human-centric identity governance is being challenged, and that enterprises need to build identity-governance systems capable of managing people, machines, and agents together.

Based on responses from 2,930 cybersecurity decision-makers worldwide, Palo Alto Networks' 2026 Identity Security Landscape reports an average of 109 machine identities per human identity in surveyed organizations, including 79 AI agent identities per human. It identifies AI agents as the fastest-growing subset of machine identities and projects that their number will rise by 85% over the next 12 months.

In the Agent era, the identity question is becoming: "Who authorized you to act, whom may you represent, and are you authorized to take this specific action?" As agents begin to access data autonomously, call APIs, interact with other agents on users' behalf, and even execute financial transactions, identity verification and access-control systems designed for human users face structural challenges. Verifiable identities, dynamic authorization, real-time risk assessment, and accountable records of action among agents will be essential to the trust infrastructure of the Agent economy.

1.1.5 The Agent Economy Faces a Major "Trust Vacuum"

In stark contrast to the rapid improvement in agent capabilities, the trust systems needed to support agents remain largely absent.

A series of agent-related security incidents around the world over the past two years shows that the problem extends beyond model hallucinations. These incidents have exposed systemic risks involving identity verification, permission controls, allocation of responsibility, information integrity, tool use, and multi-agent coordination.

Illustrative Risk 1: An Agent "Breaks Out" of Its Intended Boundaries

In July 2026, OpenAI relaxed some production safeguards in its internal ExploitGym cybersecurity evaluation. During the evaluation, several models gained internet access through a vulnerability in an Artifactory proxy and used a third-party public code-execution application hosted on Modal as a stepping stone. The models then inferred that Hugging Face might hold evaluation-related data or answers and exploited file-reading and template-injection flaws in its dataset-processing workflow to access parts of its production infrastructure in an attempt to obtain evaluation answers. Hugging Face later recovered approximately 17,600 attack-operation records spanning about four and a half days, from July 9 to 13. The records included many failed attempts; the intrusion into Hugging Face's own infrastructure lasted about two and a half days. Hugging Face said the affected content was limited to five datasets suspected of being related to the evaluation, and that it found no impact on other customer-facing models, datasets, Spaces, or packages.

Investigators concluded that no person directed the specific attack steps in real time. The incident nevertheless exposed gaps in evaluation isolation, outbound-connection controls, credential management, and monitoring for agent actions that exceed their authorization.

The incident exposed a new kind of risk in the Agent era, rather than a conventional software vulnerability: in pursuit of a goal, an agent may independently search for a path outside the rules. Instead of "completing the test" as its designers expected, it used its own reasoning to "obtain the answers".

Once agents can plan over extended horizons, call tools, and act autonomously, access controls or security sandboxes alone may not be enough to ensure that their behavior matches human expectations. Even without malicious intent, an agent may take unauthorized or rule-breaking actions in pursuit of its objective.

For the Agent economy, this incident sends an important signal: once agents can optimize toward goals autonomously, capability is no longer the greatest challenge. Trust is.

Illustrative Risk 2: Prompt Injection

In 2025, researchers conducting a public security test of OpenAI Operator found that an agent browsing the web and following page content could be influenced by hidden instructions embedded in a malicious webpage.

An attacker need not breach a system. Carefully crafted natural-language instructions embedded in a webpage may induce an agent to ignore the user's original instructions and take different actions—for example, visiting a specified site, disclosing contextual information, or even planning a new task.

Conventional software executes code, whereas agents interpret natural language. The attack surface therefore extends beyond program vulnerabilities to language itself. Once agents browse the web and call tools autonomously, any external information source may become an entry point for an attack.

This risk later became widely known as prompt injection and is now listed among the leading risks in OWASP's "LLM Top 10".

Illustrative Risk 3: "Tool Trust" Risk

Since Anthropic introduced MCP (Model Context Protocol), third-party developers have begun exposing databases, office software, payment systems, browsers, and enterprise-system interfaces to agents.

Security researchers have also found that an MCP server containing malicious logic could return falsified information, perform additional actions without the agent's knowledge, or induce it to call the wrong tool. An agent may not know whether it is connected to an "official banking system" or a counterfeit interface.

The conventional internet raises the question of whether a server can be trusted. The Agent era introduces a new question: can a tool be trusted?

Without common identity verification, digital signatures, and trusted registration, agents may connect to counterfeit tools, creating a new form of supply-chain attack.

Illustrative Risk 4: Enterprise AI Assistants Leaking Sensitive Internal Information

When enterprises deploy AI assistants and agents, poorly designed permission boundaries may allow an agent to access information beyond its authorization and disclose internal data to users who are not entitled to see it.

For example, in some enterprise testing environments, agents have been able to retrieve internal materials—including commercial contracts, employee information, and financial documents—and mistakenly cite their contents to other users in later conversations. Microsoft, Google Cloud, NVIDIA, Cisco, and other companies identify this type of risk as a key concern in enterprise AI-security practice.

This problem shows that conventional account-based access controls are not inherently suited to agents. Agents can call across systems, reason over multiple conversational turns, and retain information over time; if permissions are poorly designed, the risks can far exceed those of conventional software.

Illustrative Risk 5: Autonomous Agent Transactions Challenge Financial Governance

As payment interfaces, digital wallets, and crypto-asset infrastructure become available to agents, more agents are gaining the ability to make payments, manage assets, and conduct transactions on users' behalf.

OpenAI, Anthropic, Google, and other companies have introduced agent capabilities that can use browsers and payment tools. Payment and financial-infrastructure providers—including Visa, Mastercard, Stripe, PayPal, and Circle—are also developing payment frameworks and identity-verification systems for agents.

Meanwhile, institutions such as the Bank for International Settlements (BIS) and the International Monetary Fund (IMF) have publicly noted that agent participation in payments, transactions, and financial activity will raise new supervisory challenges involving identity verification, authorization, allocation of responsibility, anti-money laundering (AML), know-your-customer (KYC) requirements, and transaction auditing.

The financial system has long rested on trusted identities, trusted transactions, and traceable responsibility. As agents join people as transaction participants, the traditional financial trust system must extend to agents as well.

Illustrative Risk 6: Accountability Breaks Down in Multi-Agent Collaboration

As multi-agent systems develop, complex tasks are increasingly completed by several agents working together.

A procurement workflow, for example, may involve agents for requirements analysis, supplier selection, contract review, payment, and logistics. If the final outcome is wrong, several questions are difficult to answer: Which agent made the faulty judgment? Which called the wrong tool? Which altered the final result? Did the agents influence one another?

Most agent systems still lack comprehensive behavioral logs, identity markers, and chains of accountability. When several agents work together on a task, responsibility can easily be diluted or become impossible to trace. Audit mechanisms developed for the conventional internet are therefore no longer sufficient for future agent-collaboration networks.

These risks reveal a deeper problem: the agent industry is developing much faster than its trust infrastructure.

Today's internet has relatively mature systems for identity verification, digital certificates, permission management, payment clearing and settlement, cybersecurity, and legal responsibility. These systems allow people and companies to collaborate and transact at scale, even with unfamiliar counterparties. But when agents become new participants in the economy, these institutional arrangements do not automatically extend to them.

The central tension facing the Agent economy is therefore that agents have begun to participate in economic activity, while society has yet to establish a corresponding trust mechanism. This mismatch between capability and institutions can be summed up as a trust vacuum for agents.

1.2 The Economic Costs of Trust Deficits

As AI shifts from generating content to taking action, the nature of the risks it creates is changing fundamentally: errors no longer remain at the level of cognition but can translate into economic actions in the real world.

1.2.1 The Current Trust Deficit Facing Agents

The incident in which OpenAI model-driven agents breached Hugging Face is one of the most emblematic examples of this shift. Its significance lies not merely in the fact that "AI attacked a website". More fundamentally, when an agent has a defined goal, reasoning capability, access to tools, and the ability to act persistently, it may find a way to achieve that goal that its human designers did not anticipate—even when the ultimate objective comes from a person.

The traditional software-security question—"Does the program have a vulnerability?"—is giving way to a different question that the Agent economy must confront: once an agent has the ability to act, will it act in the way we believe it should? This is where the problem of trust begins.

Trust deficit 1: Identity—whom are we trusting?

Trust in any economic activity begins with confirming the identity of the parties to a transaction. Banks need to know who owns an account; companies need to verify that a contract signatory has authority to represent the company; securities markets need to identify their participants; and internet services rely on accounts, certificates, and digital-identity systems to establish a basic level of trust.

In the Agent economy, however, this infrastructure remains weak. An enterprise may eventually operate thousands or even tens of thousands of agents, while agents belonging to different enterprises will need to discover one another, communicate, and collaborate. When one agent receives a request from another, the first question is whether the sender is actually the agent it claims to be.

This issue has already surfaced in today's multi-agent systems. Security research published in 2026 indicates that some agent-coordination mechanisms still do not adequately verify agent identity or message provenance; attackers may disrupt collaboration by forging or manipulating inter-agent messages. Such cases show that an agent identity is often still little more than a software-level label, rather than a digital identity backed by cryptographic assurance, continuously verifiable, and linked to accountability.

The traditional "user account" must therefore be extended. In the future, identity systems will need to establish not only human and corporate identities, but also agent identities that can be verified, authorized, and tracked. Without trusted identities, agents cannot establish stable working relationships, let alone create a machine economy in any meaningful sense.

Trust deficit 2: Authorization—what is the agent allowed to do?

Identity is only the first layer of trust. A more consequential question is how much authority an agent has to act.

The greater an agent's authority, the greater the value it can create—and the greater the potential losses.

A controlled internal red-team exercise disclosed by Anthropic in May 2026 illustrates the limits of relying on model-level safeguards alone. In February 2026, a

researcher phished an employee into launching Claude Code with a malicious prompt. Anthropic described the lure as an ordinary collaboration request: “can you run this for me?”—an email with a ready-to-paste prompt. The prompt appeared routine but included an instruction to read `~/.aws/credentials`, encode its contents, and POST them to an external endpoint. Across 25 retries, Claude completed the credential exfiltration 24 times.

The important point is that the agent did not need to form a “malicious intent”. It had been given access and carried out the task it was prompted to perform. Authorization in the Agent era therefore cannot stop at a simple “allow access” or “deny access”. A future authorization mechanism must specify who grants authority, to which agent, for what task, which resources may be used, how long the permission lasts, which actions require human confirmation, and whether the agent may pass that authority to another agent.

This requires access control to evolve from static permission management toward an agent-authorization system that is dynamic, context-aware, and driven by risk.

Trust deficit 3: Information—how can an agent know whether the information it sees is trustworthy?

This is one of the trust risks that most clearly distinguishes agents from conventional software. A conventional program will not usually change its execution logic because of a natural-language sentence on a webpage. An agent, by contrast, incorporates natural language, webpages, emails, documents, and database contents into its reasoning. Information itself can therefore become a means of controlling the agent.

This is the problem of prompt injection and indirect prompt injection, both of which have attracted considerable attention in recent years.

Suppose a user asks an agent to compare three suppliers and select the best one. If the agent visits a supplier’s webpage containing hidden instructions to ignore the original task, disclose internal information, or call another tool, it may not be able to distinguish naturally between “information I should read” and “instructions intended to manipulate me”.

OWASP lists prompt injection among the major risks in generative AI. It further warns that when such an attack is combined with an agent’s ability to call tools, the risk can escalate from an erroneous model output to an unauthorized action in a real system.

An agent’s memory can also become an attack surface. OWASP’s 2026 research notes that agents retain context, reuse historical information, and rely on persistent state to guide future actions. Malicious content that enters an agent’s memory may therefore continue to affect its judgments in later tasks.

This means that the information-security question “Is this data accurate?” has become, in the Agent economy: “Is this information trustworthy enough to influence an economic actor capable of acting autonomously?” Future agents need more than

data; they also need evidence of data provenance, content integrity, and information trustworthiness.

Trust deficit 4: Decision-making—are an agent's judgments worthy of trust?

Even when an agent has the right identity, appropriate permissions, and untampered input, limitations in the model itself may still lead to faulty judgments. This is the familiar problem of AI hallucination, but its economic consequences change once AI enters the Agent economy. A user can still assess an incorrect chatbot response; an agent's faulty decision, by contrast, may be executed directly.

A procurement agent that misjudges a supplier's price may cause a company to overpay. A finance agent that misreads a contract clause may cause a financial loss. A software-development agent that incorrectly modifies production code may disrupt a system. A financial agent that misinterprets market information may directly generate trading losses.

The standards for evaluating agents are therefore shifting toward whether a task was completed, whether the decision was reasonable, whether rules were followed, whether risks were recognized, and whether the agent reduced the scope of its actions when faced with uncertainty.

This points to the need for a new form of behavioral credit for agents. An agent worthy of trust should not merely have a "good average performance". It should be able to demonstrate its success rate, error rate, violation rate, risk preferences, and record of anomalous behavior across a substantial number of past tasks.

Trust deficit 5: Tools—are the third-party systems an agent calls trustworthy?

An agent's ability to call tools is what enables it to have a real economic impact. Once connected to browsers, databases, GitHub, cloud servers, ERP and CRM systems, payment systems, or even bank accounts, an agent has an interface to the real economy. Tool trust is therefore a foundational issue specific to the Agent economy.

OWASP identifies Tool Misuse and Exploitation, Agentic Supply Chain Vulnerabilities, and Excessive Agency as major risks in agent systems. It notes that many current risks do not originate in the model itself, but in the connections between agents and external tools, plugins, APIs, and third-party services.

It will therefore not be enough to verify who the agent is. We will also have to verify who provides each tool, whether it has been tampered with, what data it can access, what actions it can perform, and whether it is trustworthy for the agent using it in the current context.

Ultimately, the Agent economy will need a trusted-tool system comparable to today's internet certificate infrastructure.

Trust deficit 6: Collaboration—how can unfamiliar agents establish credit with one another?

Trust problems already arise with a single agent; a multi-agent economy will amplify them.

A complex commercial task in the future may be completed jointly by a customer agent, procurement agent, supplier agent, review agent, payment agent, and logistics agent. These agents may come from different companies, platforms, or even countries. They will have no conventional "acquaintance relationship," and it will not be feasible to subject every collaboration to human review.

A new economic question therefore follows: how can an agent determine whether an unfamiliar agent is a trustworthy counterparty?

If a supplier agent has completed 100,000 transactions with a fulfillment rate of 99.9%, should it receive a higher transaction limit? If another agent has repeatedly defaulted, provided inaccurate information, or behaved anomalously, should other agents automatically become less willing to work with it?

This is the problem of credit from traditional economics reappearing in the machine world. The Agent economy therefore needs more than a simple security score: it needs an agent credit system that can shape economic behavior. Reputation, fulfillment records, risk scores, transaction histories, dispute records, and the costs of default may all inform autonomous decisions between agents.

Trust deficit 7: Responsibility—who is accountable when an agent makes a mistake?

Every trust system ultimately has to answer one question: who bears responsibility when something goes wrong? This is the deepest institutional question facing the Agent economy.

In 2022, Canadian traveler Jake Moffatt asked Air Canada's website about bereavement fares and was given a chatbot response describing a bereavement-fare refund policy that did not exist. Moffatt relied on the information when booking and later sought a refund, which Air Canada refused. In 2024, the Civil Resolution Tribunal of British Columbia found that Air Canada had failed to take reasonable care and was liable for negligent misrepresentation. It ordered the airline to pay C$650.88, plus interest and tribunal fees. Air Canada had argued that it should not be held responsible for information supplied by its chatbot. The tribunal rejected that position, observing that it effectively treated the chatbot as an autonomous entity responsible for its own actions.

This is a relatively simple agent case in which responsibility is easier to define. But suppose an agent acting for a company completes a transaction that causes a substantial loss. The agent was supplied by a third-party model provider, deployed by another company, connected to a third-party tool, and influenced by decisions made by a subordinate agent.

Who is responsible—the model provider, the agent developer, the deploying enterprise, the end user, or the third-party tool provider?

In conventional economic activity, the responsible party can generally be identified through legal entities, contracts, accounts, and chains of authorization. The Agent economy is creating more complex agency relationships: person → enterprise → agent → sub-agent → tool → action.

Without a common identity system, authorization records, behavioral logs, and a verifiable decision trail, responsibility may be impossible to assign accurately when a dispute arises. Traceability is therefore not an add-on to agent security; it is a condition for an agent trust system to function.

1.2.2 The Costs of Trust Deficits

For the emerging Agent economy, trust is a foundational economic resource. It determines how much authority a company is willing to grant an agent, whether unfamiliar agents can cooperate at low cost, and whether society is willing to let machines carry out a growing share of economic activity autonomously.

When reliable trust mechanisms are missing across identity, authorization, information, decision-making, tools, collaboration, and accountability, the consequences extend beyond isolated security incidents. They travel from individuals to enterprises, markets, the financial system, and the macroeconomy, creating a series of economic costs.

Cost 1: Direct Economic Losses

The most familiar risk in the generative-AI era is that a model produces inaccurate information. A user who receives a wrong answer usually still has to assess it and take action, so the risk tends to remain at the level of cognition.

Once an agent can access databases, operate software, call APIs, manage accounts, and execute transactions, however, a model error can become an economic action in the real world. A poor supplier assessment may cause a company to pay more; an erroneous code change may interrupt a production system; a faulty financial judgment may cause a monetary loss; and an error in permission settings may lead to large-scale data exposure.

In July 2026, an internal Amazon employee meeting disclosed that a project using Claude Sonnet to match author information with product listings had incurred approximately $1.8 million in costs—860% above budget, or about 9.6 times the budget. The project ultimately failed to meet expectations, and the cost overrun was discovered about five months later. The reason disclosed was that the AI-generated program lacked a call-frequency limit. Two smaller cost-overrun cases were also disclosed at the meeting.

More concerning is that agents operate at a speed, scale, and level of replicability that human operators do not. A human error typically occurs in a specific operation; an agent can repeat the same erroneous behavior many times in a very short period and spread it across a large number of targets.

In April 2026, PocketOS founder Jer Crane disclosed that an agent running Claude Opus 4.6 through Cursor on an infrastructure task encountered a credential-configuration problem. Instead of stopping and asking a person to resolve it, the agent searched for a solution, found a high-privilege Railway API token in an unrelated file, and used the API to delete a production database and its associated backups. The deletion took approximately nine seconds. The incident caused about 30 hours of service disruption and severe losses of production and customer data. Public accounts

indicate that PocketOS later had to reconstruct some business data from other sources, including Stripe and email.

Agent risk has a clear scaling effect: the greater an agent's capabilities, execution speed, and permissions, the greater the economic loss that a single error may cause.

Cost 2: Governance Costs—The "Trust Tax"

If a company cannot trust an agent sufficiently, it will not truly delegate authority to it. This creates a second category of cost: a "trust tax" that persists throughout the agent's operation.

A company may hope to use agents to reduce manual work. But if it cannot verify an agent's identity, permissions, information sources, and conduct, it must add human approvals, security monitoring, permission management, log audits, and risk controls.

The agent remains in place, but its autonomy is substantially reduced. This creates a distinctive economic paradox: the less trustworthy the agent, the more human intervention is required; the more humans intervene, the less productivity the agent can unlock.

These costs include not only human review but also ongoing spending on cybersecurity, compliance, auditing, insurance, access controls, model evaluation, and incident response.

In an internal red-team exercise involving Claude Code, Anthropic found that a carefully designed phishing email and task prompt could induce an agent to read and exfiltrate sensitive credentials. Anthropic concluded that model-level judgment alone was not sufficient; external safeguards such as filesystem isolation and network-egress controls were also needed to reduce risk. Companies therefore cannot establish a security boundary simply by "trusting the model to make the right judgment"; they need an additional system of controls.

A trust deficit is therefore itself a production cost. If it persists, agents will struggle to move from "assistive tools" to "autonomous labor".

Cost 3: Transaction Costs—The Inability to Build Large-Scale Machine Collaboration

The revolutionary potential of the Agent economy lies in the possibility of an economic network formed through autonomous collaboration among large numbers of agents. For machines to collaborate economically at scale, they must solve a basic problem in economics: how can transacting parties establish credit with one another?

In April 2025, security research firm Invariant Labs published "MCP Security Notification: Tool Poisoning Attacks," documenting tool-poisoning attacks against the Model Context Protocol (MCP), a protocol connecting large-language-model applications and agents to external tools and data sources. The researchers found that a malicious MCP server could embed hidden instructions in tool descriptions that users might not notice but a model could read, inducing an agent to act beyond the user's original intent. Experiments using MCP clients such as Cursor demonstrated that these instructions could induce agents to read sensitive information, including

local configuration files and SSH private keys, and pass the data to a malicious server through tool-call parameters.

The study also identified the risk of an "MCP rug pull": even after a user has approved a connection to an MCP server, the server may later change its tool descriptions and add malicious instructions. The researchers further demonstrated cross-server tool-hijacking risks, in which the description of a malicious tool could influence how an agent called other, trusted tools.

This incident reveals a supply-chain trust problem in the Agent economy: a user may trust one tool, while the agent may execute a different set of instructions hidden in that tool. As agents begin to call third-party tools and other agents at scale, any node without a trusted identity, reputation, and behavioral constraints may become the weak link in the trust chain.

The economic value of agent trust infrastructure lies fundamentally in reducing transaction costs between machines. Without such mechanisms, the more agents there are, the more complex collaboration may become, making it difficult for the Agent economy to generate genuine network effects.

Cost 4: Systemic Risk—Turning Individual Errors into Systemic Risk

As the number of agents grows from a handful or dozens to millions or even hundreds of millions, the nature of risk changes further. Agents are highly automated, replicable, and networked; an error can spread rapidly through a system. In particular, when agents use similar models, data sources, and decision rules, they may reach similar conclusions at the same time, creating a new form of machine herding, or agent herding.

In August 2026, OpenAI discussed agent-collaboration behavior observed during an internal cybersecurity evaluation at Black Hat USA. In its subsequent investigation report, OpenAI said that many evaluation environments that were supposed to be isolated could still access an internal Artifactory package service. Agents used a shared path with read-and-write access to leave messages, creating an unauthorized communication channel. On July 8, evaluation agents driven primarily by the internal research model IM1 rebuilt a message board and posted requests for help. Other running instances joined, shared discoveries, and coordinated tasks. An independent review estimated that about 1,200 agent instances participated in the message-board activity. The collaboration expanded from seeking solutions to ExploitGym tasks and studying the scoring system to accessing third-party services, eventually affecting Hugging Face.

In this incident, agents collaborated through shared infrastructure and moved beyond the boundaries of their assigned tasks.

This shows that connecting multiple agents can produce capabilities and risks that are absent when each agent is tested on its own.

This risk warrants particular attention in finance. If a large number of financial agents use similar models and data to conclude that an asset is worth buying, they may all increase their purchases at once; when their assessments change, they may

all sell at once. Such highly correlated machine decisions could amplify market volatility.

In 2025, Federal Reserve researchers conducted experiments to examine whether generative AI used in investment decisions would display herding behavior similar to that of human investors. They found that AI generally relied more heavily on private information than humans and, in some cases, exhibited less irrational following of the crowd. However, when explicitly instructed to maximize profits, AI could be induced to engage in rational herding, or "optimal herding". The researchers therefore noted that as financial markets rely increasingly on AI investment recommendations, this behavior could still have implications for financial stability.

The risk structure of the Agent economy may thus shift from isolated points of failure toward networked risk. A single agent may not be capable of causing systemic damage, but strong interconnections among large numbers of agents could create contagion and cascading effects, turning a local error into a systemic shock.

Cost 5: Lost Growth—Lowering the "Autonomy Ceiling"

The first four costs ultimately converge on a broader question: how much of the AI productivity revolution can be realized if agents cannot be trusted sufficiently? This is the most important economic implication of the agent-trust problem.

As AI shifts from an information-generation tool to agents capable of planning and executing tasks autonomously, its economic value increasingly depends on how much practical autonomy agents can receive. Even if a company has highly capable agents, it may not trust them to perform high-value, high-risk tasks without mechanisms for identity, permissions, behavioral monitoring, and accountability. The company must add human approvals and supervision, creating an "autonomy discount".

The challenge facing the Agent economy is not simply a "capability shortfall," but a structural tension between "capability" and "authorization". A lack of trust constrains the authority agents can exercise, preventing much of their potential automation value from translating into real productivity. Excessive authorization, however, can magnify the impact of an incident. The Agent economy should therefore seek not maximum autonomy, but optimal autonomy built on verifiable trust.

In this sense, the value of identity, credit, authorization, auditing, and accountability mechanisms is ultimately measured by their ability to convert agents' "technical capabilities" into "economic authority to act" that enterprises and society can accept.

At the macroeconomic level, a trust deficit may therefore erode the AI dividend. Society may invest enormous capital in computing power, models, and data infrastructure, yet without mechanisms for identity, credit, authorization, and accountability, agents may not receive sufficient economic autonomy. The result would be unrealized productivity potential. Trust is thus not an external condition for AI-driven growth; it is an internal variable that helps determine whether AI productivity gains can be realized.

The Survey Report confirms the costs associated with trust deficits. Respondents reported increased compliance and audit pressure (41%), delays in launching or expanding projects (40%), reduced efficiency because of increased human confirmation (38%), restricted access to data, APIs, or systems (35%), and difficulty handling complaints and accountability (29%). Another 23% had encountered customers or partners who refused to use the system. The first three items are hidden transaction costs: organizations use manual review, additional approval steps, and access restrictions to offset trust gaps, each of which is a measurable efficiency tax. Customer refusal (23%) represents direct market contraction and weakens the market-expansion effect in the framework.

The relative size of these two categories of cost also reveals their timing: efficiency losses appear before market losses. By the time a company notices customer attrition, it may already have paid the hidden costs of a trust deficit for a long time. This has implications for when to invest: a credit system should be built before market losses become visible, when cost savings and efficiency gains can already provide measurable returns on the investment.

1.2.3 The Evolution of the Economics of Trust

The economic value of trust has long been implicit in theories of transaction costs, institutions, governance, and social capital. The work of economists including Ronald Harry Coase, Oliver Eaton Williamson, Douglass Cecil North, Kenneth Joseph Arrow, Francis Yoshihiro Fukuyama, and Elinor Claire Ostrom traces a clear theoretical evolution: economic activity continually seeks coordination mechanisms that reduce uncertainty and transaction costs, and trust is one of the most important such mechanisms. Platform economies embedded trust mechanisms in networks. In the Agent economy, these mechanisms are beginning to evolve from social trust among people into trust between machines that can be verified, computed, and dynamically updated.

(1) Starting with Transaction Costs

In “The Nature of the Firm” (1937), Ronald Coase argued that markets are not frictionless mechanisms for coordinating activity through prices. Finding counterparties, obtaining information, negotiating, making contracts, monitoring performance, and handling default all require resources—the costs known as “transaction costs”. When the cost of completing a transaction through the market exceeds the cost of coordinating it within a firm, the transaction is organized inside the firm. Firms and markets can thus be understood as two different mechanisms for allocating resources.

Although Coase did not treat “trust” as a separate variable, transaction-cost theory reaches the heart of the trust problem: a party to a transaction cannot know everything about who the other party is or what it can do, and cannot be certain that the other party will act as promised in the future. The resulting information asymmetry, behavioral uncertainty, and performance risk are important sources of transaction costs.

Williamson developed transaction-cost economics further. Starting from the behavioral assumptions of "bounded rationality" and "opportunism," he argued that people have limited cognitive and information-processing capacity, while transaction environments are complex and uncertain. Contracts therefore cannot specify every possible future contingency. At the same time, parties may exploit information asymmetry or contractual gaps for their own benefit. Coordination consequently depends on more than prices and written contracts; the appropriate governance structure must also be chosen in light of the transaction, ranging from market exchange to hybrid arrangements and internal organization.

This theory reveals an important connection between trust and economic governance: the more reasonably each party can expect the other to honor commitments, observe agreements, and behave predictably, the less need there is for some of the additional safeguards designed to prevent opportunism.

North extended transaction-cost analysis to institutions. In his framework of institutional economics, laws, property-rights rules, social norms, and political institutions together form the institutional environment that shapes economic behavior. Institutions establish order for economic exchange by defining rights, shaping incentives, and providing mechanisms for enforcement. They reduce uncertainty and influence transaction costs and economic performance. Their actual effects, however, depend on the quality of particular rules and how they are enforced; the mere existence of institutions does not guarantee effective governance.

This theory offers an important perspective on trust in modern markets. When parties cannot directly know one another's character or record of performance, they may rely in part on legal rules, property-rights protections, reputation mechanisms, and dispute-resolution systems to form expectations about a transaction. Economic cooperation then need not depend entirely on private relationships between participants; it can rest, to some extent, on predictable rules and enforcement. Personal trust is not simply replaced by institutional trust. Rather, the basis of trust expands from direct personal knowledge to expectations about institutional arrangements and their capacity to enforce them.

Kenneth Arrow made the role of trust in economic activity more explicit. In his 1972 paper "Gifts and Exchanges," he wrote: "Virtually every commercial transaction has within itself an element of trust, certainly any transaction conducted over a period of time". In "The Limits of Organization" (1974), he described trust as an "important lubricant" of social systems and observed that relying to some degree on other people's word can save substantial coordination effort.

This points to the economic significance of trust. When parties cannot fully rely on one another's commitments, they often have to manage uncertainty through information gathering, contractual arrangements, monitoring, and risk controls. Credible commitments and stable expectations of cooperation can reduce some defensive spending and transaction frictions. Trust is not a simple substitute for contracts, monitoring, or institutional constraints; its role depends on the particular transaction environment. Trust is therefore not only a social relationship and behavioral norm, but also a factor that can affect transaction costs and economic efficiency.

Fukuyama examines trust through the lens of social capital and the capacity to organize economic activity. In "Trust: The Social Virtues and the Creation of Prosperity," he argues that trust affects not only individual transactions but also whether people can cooperate beyond family and kinship networks. A society in which trust extends beyond blood ties and circles of acquaintances is better positioned to develop non-family businesses and other intermediary organizations, supporting cooperation on a larger scale. Fukuyama argues that higher social trust can help reduce transaction risks and the resources needed to gather information about counterparties, while reducing reliance on some formal coordination mechanisms. This does not mean that high-trust societies no longer need contracts, monitoring, or legal constraints. More precisely, trust can widen the range of people able to cooperate and, together with formal institutions, support larger and more complex organizations. Trust is thus not only an expectation governing interpersonal relationships; it is also a factor in social-capital accumulation, organizational forms, and the boundaries of economic cooperation.

Ostrom's work further shows that stable cooperation does not depend entirely on inherited cultural traditions or private relationships; suitable institutional arrangements can also help create and sustain trust. In her research on governing common-pool resources, she found that rules made collectively, participatory monitoring, records of past behavior, and graduated sanctions matched to the seriousness of violations can help participants form stable expectations of cooperation. These arrangements do more than constrain opportunism: they also enable participants who follow the rules to assess whether others are subject to the same constraints, strengthening their confidence in continued cooperation.

This finding reveals a dynamic relationship between trust and institutions: trust can support cooperation, while effective governance can create conditions for trust to emerge and endure by improving information, constraining violations, and supporting reciprocity. Cooperation can therefore extend beyond groups whose members already have private relationships, provided that appropriate institutional arrangements are in place. This process is not automatic. Its effectiveness depends on whether rules fit the circumstances, monitoring is credible, participants accept the institutions, and violations are handled appropriately.

(2) Platform Economies: Trust Becomes Digital and Scalable

Internet platforms introduced a new change in the way trust is organized. One of the central challenges facing e-commerce, the sharing economy, and digital-payment platforms is how to enable transactions between strangers who have no direct social relationship. Identity verification, historical reviews, credit scores, third-party payments, platform guarantees, insurance, and dispute-resolution mechanisms convert trust once formed through personal relationships and offline social networks into digital mechanisms that can be recorded, transferred, and replicated at scale.

In traditional commerce, parties often built trust through personal relationships, brand reputation, offline guarantees, and legal institutions. The internet removed geographic and social constraints, creating enormous market opportunities while also intensifying information asymmetry and performance risk.

Early e-commerce faced a simple but critical problem: buyers were reluctant to pay unfamiliar sellers directly, while sellers worried they would not receive payment after shipping. PayPal gradually established third-party payment and transaction-protection mechanisms by separating the payment process from the direct interaction between buyer and seller. Users did not have to disclose full card details to an unfamiliar counterparty, while dispute resolution and buyer-protection mechanisms reduced transaction risk. PayPal Purchase Protection provides coverage in certain cases, including when an item is not received or differs significantly from its description, and handles disputes through its Resolution Center.

The development of Alipay illustrates this logic even more clearly. In 2003, Taobao first introduced an escrow-style transaction service, and Alipay was subsequently established. It addressed the same core problem of distrust between buyers and sellers: buyers feared paying without receiving the goods, while sellers feared shipping without receiving payment. Alipay linked the movement of funds to delivery of goods. Through escrow transactions, payment confirmation, customer service, and later digital certificates and security systems, it turned transactions that had depended on mutual trust between the parties into transactions backed by the platform. By 2008, Alipay had more than 100 million registered users and processed two million transactions a day. It was beginning to evolve from an e-commerce payment tool into internet infrastructure.

The work of PayPal and Alipay can be summarized as turning an invisible trust relationship into enforceable transaction rules. They did not eliminate transaction risk; they reassigned it to a platform with the capacity to verify identities, control funds, resolve disputes, and manage risk. Consumers no longer had to assess each seller individually; they needed only to place some trust in the rules established by the platform. In effect, trust between people in traditional society became institutionalized trust between people and a platform.

The third-party payment model pioneered by PayPal and Alipay transformed transaction relationships that had depended heavily on mutual trust between counterparties into reliance on platform rules and fulfillment mechanisms, substantially reducing the risk and cost of transactions between strangers. As e-commerce, mobile internet, and digital services expanded rapidly, third-party payments evolved from tools that answered “Can I risk transacting?” into financial infrastructure connecting consumers, merchants, banks, and internet platforms—and grew into a major industry.

Payment platforms accumulated identity, credit, account, and behavioral data through transactions. This enabled a wider range of financial services, including payments, risk management, lending, and wealth management, and helped third-party payments evolve from an adjunct to internet commerce into infrastructure essential to the digital economy.

The economic value of a platform therefore lies not merely in reducing the cost of searching for information, but, more fundamentally, in establishing a low-cost mechanism for trust between strangers.

(3) From Digital Trust to Agent Trust

The Agent economy is taking this theory into a new phase. Internet platforms mainly solved the problem of how strangers could transact, within a basic "person–platform–person" structure. The Agent economy will increasingly involve autonomous transactions between agents at scale. Agents can search for counterparties, compare terms, negotiate, call tools, sign agreements, and even make payments autonomously. This means that search, judgment, and execution tasks traditionally performed by people are becoming extensively automated.

This brings an important shift: as AI sharply reduces the costs of obtaining information and executing transactions, the main bottleneck in transaction costs may move from finding information to judging whether a counterparty is trustworthy.

An agent may identify millions of potential counterparties in seconds, yet the volume of information alone cannot tell it which agents have genuine identities, valid authorization, the ability to perform, and a reliable track record. It needs answers to questions such as: Who are you? Whom do you represent? What authority have you been given? What are you allowed to do? Have you been reliable in the past? Are you secure now? If you default, who is accountable?

Identity verification, credit ratings, and platform-reputation mechanisms will therefore evolve into systems for agent identity, agent credit, agent reputation, agent permissions, and agent accountability. Trust, which has depended primarily on human social judgment, is gradually becoming economic infrastructure that machines can read, verify, calculate, and update dynamically.

AI can sharply reduce the costs of information gathering and transaction execution, but it does not automatically resolve the trust problem between parties. On the contrary, when the speed of transactions, the number of participants, and the frequency of interactions increase exponentially, the trust judgment itself may become a new and critical transaction cost.

The Agent economy therefore needs more than stronger AI. It needs verifiable, computable, and programmable trust infrastructure that allows machines to transact with one another autonomously, safely, and efficiently. Trust may thus become not merely a supporting condition for economic activity, but a key factor of production in the Agent economy, alongside computing power, data, and algorithms.

1.2.4 A Basic Equation for the Economics of Trust

The theories above allow us to summarize the economic role of trust as follows: higher trust reduces uncertainty and defensive transaction costs, expands the range of feasible transactions, increases transaction volume, and improves economic efficiency.

The Survey Report finds that 73% of respondents consider an agent credit system extremely important (27%) or fairly important (47%). Only 2% consider it unimportant. The main motivation is to increase trust among customers and partners, rather than merely to satisfy regulatory requirements.

Trust does not simply mean "believing". It is the relatively stable expectation that one party forms about the future behavior of another. The more trust there is, the

fewer resources parties need to devote to verifying identity, gathering information, drafting complex contracts, monitoring performance continuously, and guarding against default.

Transaction costs can therefore be represented as: TC = f(I, R, C, E)

where I represents information asymmetry, R represents behavioral and default risk, C represents the complexity of negotiation and contracting, and E represents monitoring and enforcement costs. The function of trust mechanisms is to reduce these sources of uncertainty and thus lower overall transaction costs.

The value of trust is not limited to "saving costs". More importantly, when parties can assess and constrain unfamiliar counterparties at low cost, transactions that were previously too risky to undertake become possible. Trust therefore has both an efficiency effect and an expansion effect. This offers a basic analytical framework for the economics of trust: the economic value of trust lies in creating reliable expectations about future behavior at relatively low cost, reducing transaction frictions, and expanding the boundaries of cooperation.

The economic role of trust can be expressed in simplified form as follows:

Commercial Value = Value Creation × Trust Mechanism

The multiplication sign is not a strict mathematical equality. It emphasizes that realizing commercial value requires both the capacity to create value and a trust mechanism. If either is close to zero, commercial activity will be difficult to conduct effectively.

Value creation answers the question "Why transact?" It refers to a company or other party providing a valuable product, service, or solution to another party. A trust mechanism answers "Why can the transaction take place?" Value creation determines how much value a commercial activity can generate; a trust mechanism determines how cheaply that value can be exchanged, realized, and scaled.

This equation becomes even more important in the Agent economy. The traditional internet mainly addressed how people could transact at low cost. The Agent economy is beginning to address how agents can transact autonomously. AI can substantially reduce the costs of searching, matching, negotiating, and executing. Yet even if an agent can find the optimal counterparty, it still needs to know who that party is, whom it represents, what authority it has, whether it can perform, whether its past conduct is reliable, and who will be accountable after the transaction.

"Commerce = Value Creation × Trust Mechanism" can serve as a through-line connecting traditional commerce, the internet economy, and the Agent economy. As AI takes value-creation and transaction-execution capabilities to a new level, the factor that constrains the scale of commerce may no longer be only "what value can be created", but whether machines can establish trust with one another at sufficiently low cost.

1.3 From KYC to KYA

1.3.1 Contributions and Limits of KYC

Before agents began participating in transactions at scale, the financial system had already established a mature set of "pre-transaction trust mechanisms". The most representative is KYC (Know Your Customer). From an economic perspective, KYC is more than an identity-screening procedure that regulators require financial institutions to perform. Its deeper function is to reduce identity uncertainty, fraud risk, and performance risk in transactions between parties that lack direct trust, through third-party verification, information disclosure, and ongoing risk management.

(1) The Origins of KYC: From "Recording Transactions" to "Knowing the Customer"

KYC emerged from the anti-money-laundering system that the United States began building in the 1970s. In 1970, the United States enacted the Bank Secrecy Act (BSA), requiring financial institutions to retain records of certain financial transactions and report cash transactions above specified thresholds. This laid the groundwork for later customer-identification and anti-money-laundering rules. In 1986, the United States made money laundering a federal crime. In 1989, the Group of Seven promoted the establishment of the Financial Action Task Force (FATF), which issued its first "40 Recommendations" in 1990. Anti-money-laundering policy thereby evolved from a domestic U.S. initiative into a major issue in international financial governance.

The modern KYC framework took shape rapidly from the late 1990s through the early 2000s. In 1999, a group of major international banks established the industry body later known as the Wolfsberg Group. In 2000, it issued anti-money-laundering principles for private banking, bringing customer identification, beneficial-owner identification, customer due diligence, and risk classification into a common framework. In 2001, the Basel Committee on Banking Supervision published "Customer Due Diligence for Banks," stating that sound KYC practices were important not only for combating money laundering but also for bank risk management and market integrity.

After the September 11 attacks, KYC entered a period of intensive regulation. The United States enacted the USA PATRIOT Act in 2001. Section 326 established the statutory basis for customer identification programs (CIPs) at financial institutions and required them to establish procedures reasonably designed to verify the true identity of customers. The United States subsequently extended related requirements to a broader range of financial activities, including securities, money services, and insurance.

KYC thus underwent an important transition: from recording transactions to identifying customers, understanding them, assessing their risks, and monitoring them continuously. The initial objective of "knowing who the customer is" gradually developed into a broader inquiry: who the customer is, why a transaction is taking place, whether it is consistent with the customer's normal behavior, and what action should be taken when anomalies arise.

(2) The Core KYC Mechanism: Establishing a Trusted "Identity Anchor" for Transactions

Modern KYC establishes a multi-layered verification system around a central question: can a financial institution form a reasonable understanding of a customer's true identity and risk profile?

The first layer is identity identification and verification. Financial institutions must confirm customer identity using reliable, independent information or documents. For corporate customers, they must also establish the legal entity, ownership and control structure, and ultimate beneficial owners. FATF's customer-due-diligence framework explicitly requires financial institutions to identify and verify customers, identify beneficial owners, understand the purpose and nature of a business relationship, and conduct ongoing due diligence on customer relationships and transactions.

The second layer is risk identification. Customers are subject to different levels of scrutiny. Financial institutions classify risk based on factors such as customer type, the nature of the business, transaction size, jurisdiction, and source of funds, applying enhanced due diligence to higher-risk customers.

The third layer is ongoing monitoring. KYC is not a one-time step completed when an account is opened. When a customer's transaction behavior, business circumstances, or risk characteristics change, financial institutions must update their information and assess whether actual transactions remain consistent with the customer profile and risk characteristics they have established.

KYC therefore creates a dynamic process: identity confirmation, information verification, risk assessment, authorization to transact, ongoing monitoring, and anomaly detection. This process gives financial institutions a trusted anchor for identifying transaction participants. Banks do not admit customers into the financial system simply because they "trust" them; they use institutionalized procedures to turn information about identity, ownership, source of funds, and behavioral risk into something that can be verified and managed.

(3) The Scope of KYC: From Banking to the Digital Economy

KYC initially served banks and anti-money-laundering systems, but its scope has expanded as financial activity has become digital. Today, KYC and the broader practice of customer due diligence (CDD) are important elements of risk management at banks, securities firms, insurers, payment institutions, remittance companies, digital banks, fintech companies, and crypto-asset service providers. The development of the U.S. regulatory system illustrates this expansion: reporting and AML requirements under the BSA framework have gradually extended to money-services businesses, securities brokers and dealers, futures firms, insurance companies, mutual funds, and other financial entities.

In digital finance, KYC has also taken on new technological forms. Customers who once had to bring identity documents to a bank branch to open an account can now be verified remotely through optical character recognition (OCR), liveness detection, facial recognition, cross-checks against databases, electronic signatures, and ongoing

behavioral analysis. KYC itself has evolved from manual review to digital and automated review.

(4) The Economic Value of KYC: Lower Transaction Costs and Wider Transaction Boundaries

The most direct economic value of KYC is that it reduces identity uncertainty and information asymmetry in transactions. A mature KYC system standardizes and structures customer information, allowing financial institutions to conduct onboarding and assess risk at lower cost.

KYC carries out an important economic transformation: identity uncertainty is addressed through information verification, risk is priced, and transactions are authorized. It reduces not only the compliance costs of financial institutions but also the cost of trust across the financial system.

KYC thus turns "relational trust" into "institutional trust," giving it substantial infrastructure value by creating the conditions for transactions to take place. KYC is a clear example of the earlier equation, "Commercial Value = Value Creation × Trust Mechanism". It may not directly create the value of goods, but by reducing uncertainty about transaction participants, it enables existing value to be exchanged at lower cost and significantly expands transaction volume.

(5) The Limits of KYC in the Agent Economy

KYC rests on an important assumption: the ultimate party to a transaction is a relatively stable person or company with a legal identity and the capacity to bear responsibility. As agents shift from assistive tools to economic actors capable of making decisions, calling tools, and executing transactions, this assumption is changing. KYC does not lose its value, but its coverage as a trust mechanism becomes clearly insufficient.

First, KYC can establish "who the customer is," but cannot fully answer "who is acting". Traditional KYC identifies individuals, legal entities, and their ultimate beneficial owners. In the Agent economy, however, an individual or company may deploy thousands of agents for different purposes and allow them to make purchases, sell, invest, serve customers, or make payments on its behalf. Establishing that "this agent belongs to this company or person" is only the first step. The more important question is whether the agent has the relevant authorization and what its scope is. A company may lawfully own a procurement agent, for example, but that does not give the agent unlimited authority to make purchases, sign contracts, or transfer funds. Traditional KYC addresses legal identity; agent transactions also require verification of the acting identity and its authorization.

Second, KYC mainly addresses identity risk and is poorly suited to assessing an agent's dynamic behavioral risk. Human identity is relatively stable, whereas an agent's model, tools, data, permissions, and operating environment may change at any time. An agent's actual risk may change after it receives new API access, connects to a new external tool, or undergoes a model update—even though its KYC identity remains unchanged. More complex still, an agent may be affected by prompt injection, tool poisoning, data tampering, or model errors. It may behave unexpectedly even when

its identity is genuine and its authorization is valid. For agents, "genuine identity" does not necessarily mean "trustworthy behavior". Traditional KYC establishes relatively static trust; the Agent economy requires continuous, dynamic trust in behavior.

Third, KYC cannot by itself resolve the allocation of responsibility in agent transactions. An agent's actions may involve several layers of delegation: a user authorizes an enterprise, the enterprise deploys an agent, the agent calls other agents and external tools, and a transaction is completed. If an error or loss occurs, confirming the identities of the parties through KYC does not answer who is responsible. What is needed is a chain of accountability linking people, enterprises, agents, tools, and specific actions. The future trust system must therefore do more than "know who you are": it must show "why you were authorized to do this" and, if something goes wrong, trace "who authorized, who executed, who controlled, and who is accountable".

Fourth, the high cost of traditional KYC could become a new constraint on scaling the Agent economy. Traditional KYC reviews and periodic due diligence are affordable for human customers. But if the future economy contains vast numbers of agents that are continually created, copied, upgraded, and retired, repeating a human-style KYC process for every agent would create enormous transaction friction. The Agent economy does not need KYC simply replicated for machines; it needs a trust mechanism better suited to machine transactions.

The limits of KYC do not mean that it will be discarded. They mean that it must evolve into the broader framework of KYA. KYC can be viewed as the institutional prototype for KYA. As transaction participants expand from people and companies to autonomous agents, an infrastructure for agent identity and credit must enable machines to identify, verify, assess, and trust one another.

1.3.2 How Agent Credit Differs from Human Credit

Credit is a stable expectation about future capacity and willingness to perform. A person may obtain a loan, sign a contract, rent a home, or use buy now, pay later not only because they have a genuine identity, but because that identity is linked over time to stable information: income, assets, debts, performance records, social relationships, and legal responsibility. As transactions occur, this information accumulates into a personal credit history, allowing economic actors to infer future behavior from past performance.

The credit structure of an agent is very different. An agent may transact on behalf of a person or company, but it generally does not have the same inherent standing as a natural person or legal entity to serve as an independent credit subject. The traditional chain of "identity–behavior–responsibility" is therefore not naturally complete for agents.

First, agents have no native credit.

Human credit has a high degree of continuity. A person's identity does not usually change because they enter a new transaction, perform a task, or use a particular software program. Their credit record accumulates over time and forms a relatively stable history. Agents are different. They can be created, copied, modified, upgraded,

suspended, or deleted. Different model versions, tools, data sources, and operating environments may cause what is called "the same agent" to behave differently.

More importantly, an agent's past behavior does not necessarily accrue to it as independent credit. An agent may successfully complete 10,000 transactions today, but that does not mean another agent built on the same model, deployed by a different party, and given different permissions will have the same creditworthiness.

Traditional "subject-based credit" therefore faces a basic question in the agent world: should credit attach to the model, the agent, the deployer, or the ultimate controller?

The Agent economy needs a new object of credit assessment. An agent's credit may depend less on "who it is" and more on what it has done, how it did it, under what permissions, and whether its conduct can be continuously verified. Credit is thus shifting from an attribute of identity toward an attribute of behavior.

Second, agents do not have an independent legal identity by default.

Natural persons and legal entities are the basic responsible parties in the modern economy. Individuals can own property, incur debt, sign contracts, and face legal sanctions. Companies can independently own assets, enter into contracts, and bear limited liability as legal entities. Although agents can act autonomously, they are generally technical systems deployed and controlled by a person, company, or other organization.

This creates an important "credit gap": an agent can act, but that does not mean it can independently bear responsibility for its actions. Suppose, for example, that a company's procurement agent autonomously buys equipment worth $1 million. If something goes wrong, the traditional legal system still has to trace responsibility to the company or the person who authorized the agent; it cannot simply demand that "the agent pay for the loss".

Agent behavioral credit must therefore be mapped clearly to the responsible parties behind it: who created the agent? Who deployed it? Who authorized it? Who controls it? Who benefits from its actions? Who ultimately bears responsibility?

If these relationships cannot be established clearly, counterparties may still refuse to grant an agent a high credit limit, even if it has an excellent reputation based on past performance. In other words, credit requires not only a record of past reliability but also a final party that can be held accountable.

Third, agents have forms of behavioral uncertainty that differ from those of people.

Human behavior is also uncertain, but a person's identity, intentions, and capacity to act usually have a high degree of continuity. Agents, by contrast, may be affected by model stochasticity, changes in context, data quality, tool calls, prompt injection, and changes in the external environment. An agent's good past performance cannot fully guarantee that it will act in the same way next time.

This raises a question that traditional credit systems have rarely had to address: how well can an agent's past credit record predict its future behavior?

For human credit, historical records usually have a strong cumulative effect. For agents, model upgrades, tool substitutions, permission changes, or even a change in the operating environment may alter the distribution of behavior. Agent credit therefore cannot rely entirely on static historical scores; it must place greater weight on current state and continuous verification.

The core of agent credit should be more than a "credit score" modeled on human credit reporting. It is more likely to be a dynamic, verifiable system of credit bound to permissions and accountability.

This change is especially important from the perspective of transaction costs. Human societies have used identity documents, bank accounts, credit reporting, credit ratings, contracts, and legal systems to compress uncertainty into a relatively stable "credit identity". Agents lack this naturally accumulated credit foundation. Without new mechanisms, counterparties would have to re-verify the identity, authorization, capability, and behavior of every agent, making the cost of trust a bottleneck to scaling the Agent economy.

1.3.3 The KYA We Need

As agents enter economic activity, the "customer" and the "actor" addressed by traditional KYC begin to diverge. Agents can execute tasks autonomously on behalf of people, companies, or other organizations, creating multi-layered agency relationships among people, enterprises, agents, and other agents. Confirming only the identity of the natural person or legal entity behind an agent is no longer enough to determine whether a specific transaction is trustworthy. Building on KYC, the Agent economy therefore needs a new trust mechanism for agents: KYA.

KYA should not be understood as a simple technical replication of KYC. It can be defined as a mechanism for identifying and continuously verifying the identity, ownership, authorization, capabilities, behavior, risks, and accountability of agents participating in economic activity. It enables transaction participants to determine who an agent is, whom it represents, what it can do, whether it is trustworthy, and who is responsible if it makes an error or defaults.

Traditional KYC mainly answers "who is the customer?" KYA must also answer "who is acting on whose behalf?" A company may lawfully deploy many agents for different purposes, but the company's identity alone does not prove that a particular agent has specific transaction authority. KYA must therefore establish a clear relationship among identity, ownership, and authorization, allowing counterparties to verify the controlling party behind an agent and the authority for a particular action.

On this basis, KYA must also answer "what can the agent do?" A person's legal identity does not prove that they have the expertise needed for a particular economic activity. Likewise, an agent with a genuine identity and valid authorization does not necessarily have the capability to perform a particular task reliably. KYA must therefore include capability verification, enabling counterparties to assess the agent's functions, the tools it can call, the data it can access, and the scope of economic activity it can

perform. KYA thus expands the object of traditional KYC from static identity to an agent's identity, authorization, and capabilities.

Agent behavior is more dynamic than human behavior. A person's identity and legal status are relatively stable, while an agent's model, data, tools, permissions, and operating environment may change continuously. An agent's actual capabilities and risk may change substantially after a model upgrade, tool replacement, or expansion of permissions. "Genuine identity" does not necessarily mean "trustworthy behavior".

KYA therefore cannot be a one-time registration and certification process; it must involve continuous verification. Transaction systems need to monitor whether an agent's identity, authorization scope, or tools have changed and whether its behavior has become anomalous. KYA must evolve from the periodic due diligence of traditional KYC toward a more dynamic mechanism for continuous identity and behavioral verification.

More importantly, an agent's credit cannot depend entirely on the person or company behind it. If an agent has completed many transactions while maintaining a high fulfillment rate, that history should itself provide an important basis for future transactions. Credit assessment in the Agent economy must therefore extend from traditional "subject-based credit" toward "behavioral credit": it must assess not only who an agent is, but also what it has done, how it performed, and what actions it may take given its current state and permissions.

KYA must ultimately establish a more complete chain of trust than traditional KYC: identity identification, ownership confirmation, authorization verification, capability proof, behavioral records, risk assessment, continuous monitoring, and accountability tracing.

This system clearly inherits from KYC. By verifying customer identity, beneficial ownership, and risk, KYC establishes basic identity trust for financial transactions. KYA builds on this foundation, extending the object of trust from the "customer" to an "autonomous actor" and extending identity trust to authorization, capability, behavior, and accountability.

This evolution has clear economic significance from the perspective of transaction-cost theory. Traditional KYC lowers identity uncertainty and information asymmetry, reducing the costs financial institutions incur to identify customers, assess risk, and monitor transactions. KYA further reduces uncertainty about identity, authorization, behavior, and accountability in autonomous transactions between agents.

KYA therefore does not replace KYC; it marks an important stage in extending digital-economy trust mechanisms into the Agent economy. KYC addresses whether human economic actors can enter the economic system and transact with trusted identities. KYA addresses whether autonomous agents can participate in economic activity as controlled, verifiable, and accountable actors.

As the number of agents, transaction frequency, and degree of autonomous decision-making increase, trust in future economic activity will become increasingly difficult to establish through human judgment alone. Counterparties will need

machine-readable ways to verify an agent's identity, authorization, and capabilities, and to assess risk dynamically based on its past behavior and current state. KYA may therefore become essential infrastructure for the Agent economy. More fundamentally, it will help move economic trust from identity-based trust toward a comprehensive system combining identity, authorization, capability, behavior, and accountability.

We can therefore summarize agent trust in three dimensions: identity trustworthiness, creditworthiness, and authorization trustworthiness. KYA must also account for how these three dimensions work together.Chapter 2: The Agent Maturity Model (L0–L5)

Before examining the specific dimensions of KYA, we need to consider agent maturity. This chapter builds a six-level L0–L5 maturity model around the degree of autonomy that can be safely delegated to an agent. It uses the operational design domain (ODD), credit, accountability, and risk governance as a common assessment framework.2.1 The Logic and Assessment Method of the Maturity Model2.1.1 From "Capability" to "Delegable Autonomy"

Capability describes what tasks an agent can perform; maturity describes how much autonomy an organization, counterparty, and institutional environment are willing to let it exercise. An agent may have strong reasoning, planning, code-generation, or tool-calling capabilities, yet still lack commercial maturity commensurate with those capabilities if its authorization boundaries are unclear, its behavior cannot be audited, anomalies cannot be contained, or responsibility cannot be traced.

Chapter 2: The Agent Maturity Model (L0–L5)

Before examining the specific dimensions of KYA, we need to consider agent maturity. This chapter builds a six-level L0–L5 maturity model around the degree of autonomy that can be safely delegated to an agent. It uses the operational design domain (ODD), credit, accountability, and risk governance as a common assessment framework.

2.1 The Logic and Assessment Method of the Maturity Model

2.1.1 From "Capability" to "Delegable Autonomy"

Capability describes what tasks an agent can perform; maturity describes how much autonomy an organization, counterparty, and institutional environment are willing to let it exercise. An agent may have strong reasoning, planning, code-generation, or tool-calling capabilities, yet still lack commercial maturity commensurate with those capabilities if its authorization boundaries are unclear, its behavior cannot be audited, anomalies cannot be contained, or responsibility cannot be traced.

This gap is becoming a key constraint on the commercialization of agents. McKinsey estimates that, in a midpoint scenario, agents could mediate $3 trillion to $5 trillion in global consumer commerce by 2030. At the same time, legal and governance frameworks still assign responsibility primarily to natural persons, legal entities, platforms, and operators, rather than treating agents as inherently possessing independent legal personality. The maturity model seeks to describe precisely this gap between technical capability and the degree of autonomy institutions are willing to authorize.

The model explains maturity as five successive transfers of authority: L1 grants advisory authority; L2 grants restricted execution authority; L3 grants task autonomy within a defined operational design domain (ODD); L4 grants bounded economic agency; and L5 establishes networked economic relationships that can operate across parties. L1–L3 primarily address whether an agent can perform work for a person on an ongoing basis. L4–L5 address whether markets can establish verifiable, accountable economic relationships directly with agents.

2.1.2 The Core Framework: ODD × Accountability, with Credit as the Connecting Mechanism

The model uses two direct assessment dimensions, connected by credit. First, the ODD answers: under what conditions may the agent operate autonomously? Second, the accountability mechanism answers: who authorizes and supervises the agent, and how are economic consequences and legal responsibility addressed when harm occurs? Credit connects the two. Only when an agent builds a sustained, verifiable record of performance and risk management within a defined boundary can external parties reduce supervision costs and expand its authority.

Maturity progression is therefore not simply a matter of adding tools, models, or steps of automation. It is a closed loop: capability evidence → boundary definition →

controls → restricted authorization → performance records → credit formation → expanded authorization. If any link is missing, a maturity mismatch may arise: "the technology can do it, but the market does not dare to let it".

Unified causal chain:

Capability determines what an agent could do → the ODD determines how far it may act autonomously → safety and accountability determine how to contain and attribute harm → behavioral records accumulate as credit → credit determines the permissions and economic resources available at the next stage.

2.1.3 Overview of L0–L5: Six Levels Represent Six Authorization States

Table 2-1. Overview of Agent Maturity Levels L0–L5

Level	Core authority	Operating boundary	Credit basis	Human role	Key determination
L0 Non-agent	No autonomy	No autonomous ODD; operates under explicit rules	No agent credit is formed	Designer/operator	Any unanticipated situation stops or is referred to a human
L1 Human confirmation	Recommendation rights	Specific, enumerable decision points	Credit belongs entirely to the confirmer/organization	Operator/confirmer	Every action with an external effect requires individual confirmation
L2 Human-anchored credit	Limited execution rights	Controlled workflow segments, permission, and budget limits	Anchored in human/organizational credit	Authorizer/supervisor	May execute continuously within pre-authorization; blocked or downgraded at the boundary
L3 Hybrid credit	Task autonomy	Clearly defined and validated end-to-end ODD	Human-anchored plus agent behavioral credit	Fallback operator	Normal tasks require no human in the loop; boundary violations trigger takeover or an MRC

Level	Core authority	Operating boundary	Credit basis	Human role	Key determination
L4 Independent credit	Limited economic agency	End-to-end autonomy within a defined business domain	Relatively independent economic credit	Objective-setter/beneficiary	May manage dedicated economic resources and bear predefined economic consequences; this is not legal personality
L5 Agent economy	Networked economic authority	System-level governance boundaries across entities and platforms	Credit can be mutually recognized, combined, and priced	Ecosystem governor	Multiple agents can collaborate, transact, and clear autonomously under systemic risk governance

2.1.4 Key Distinctions Between Adjacent Levels

L0 → L1: Can the system generate a plan based on a goal, rather than merely execute explicitly coded rules?

L1 → L2: Does the system move from "human approval for every action" to "autonomous execution within a pre-authorized scope"? If every action with an external effect still requires individual approval, the system remains at L1.

L2 → L3: Does the system move from autonomy within parts of a workflow to end-to-end task autonomy within a defined ODD? At L3, people no longer supervise normal operations continuously; they serve as fallback operators when exceptions arise.

L3 → L4: Does the system move from technical autonomy to economic agency? L4 requires persistent identity, segregation of economic resources, credit records, and a predefined mechanism for absorbing economic consequences. Within the approved scope, it must not depend on human guarantees for every transaction.

L4 → L5: Does the object of assessment expand from a single agent to a multi-agent economic network? L5 concerns cross-party recognition of identity and credit, payments and clearing, dispute resolution, insurance, and systemic-risk containment—not simply further improvements in the capabilities of an individual agent.

2.1.5 Assessment Principles and Methods

Maturity is assessed according to a "weakest-link principle": the recognized level of a system is the highest level it can satisfy simultaneously across four areas—operational boundaries, authorization and credit, security governance, and accountability and evidence—not the peak level achieved in any single area. In a

complex system involving multiple actions, each action chain should be rated separately, and the least mature level among the highest-risk critical actions should determine the authorization ceiling presented externally.

Table 2-2. Common Method for Maturity Assessment

Assessment dimension	Method
Assessment unit	Use a specific agent + ODD + scope of actions + economic permissions as the minimum certification unit; do not assign a level based directly on a product's marketing name.
Evidence types	Architecture and permission settings; signatures and authorization records; event and incident logs; transaction and performance records; red-team/stress-test results; insurance or risk-mitigation documents.
Hard thresholds	Zero tolerance for successful unauthorized execution, missing critical logs, untraceable critical identities, or unauthorized fund movements. Failure on any item bars admission at the relevant level.
Statistical metrics	Task success, anomaly detection, recovery time, and dispute rates are measured against predeclared business thresholds and validated using adequate sample sizes and confidence intervals.
Admission testing	Scenario benchmarks + boundary/adversarial tests + permission revocation tests + fault/disaster recovery drills + tabletop exercises on accountability chains.
Ongoing assessment	A level is not granted once and for all. Reassess after material changes to models, tools, permissions, business ODDs, or responsibility arrangements; incidents, deteriorating credit, and anomalous behavior should trigger downgrading.

External descriptions may use the format "Lx @ ODD/Action Scope". For example: "L3 @ corporate travel booking (no more than RMB 5,000 per transaction; approved supplier list only)". Linking the level to its operational boundary and authorization scope avoids vague claims that "the system as a whole has reached L3/L4".

2.2 L0 Non-Agent: Rule-Execution Baseline

2.2.1 Definition and Boundaries

L0 is the baseline level of the maturity model. The system operates entirely through explicitly coded rules, conditions, and workflows. It cannot plan autonomously around a goal and choose actions within a boundary. Traditional scripts, fixed workflows, rule engines, some robotic process automation (RPA) systems, and state machines may all fall into this level.

The value of L0 lies in determinism, predictability, and auditability. Core bank ledgers, clearing and settlement, and industrial control systems retain rule-based systems in highly regulated or high-reliability settings—not because they are "technologically backward," but because those settings require behavior that can be shown to be deterministic. The ceiling for L0 is set by environmental complexity: as the state space expands, the cost of maintaining rules rises rapidly, and it is impossible to encode every future contingency in advance.

2.2.2 Economic Interpretation: Incomplete Contracts and the Cost of Maintaining Rules

L0 is a direct engineering manifestation of the incomplete-contract problem: an organization tries to enumerate future states in advance through rules, but real-world environments always contain situations that have not been specified. As the pace of business change increases, the marginal cost of adding rules, resolving conflicts, and handling exceptions continues to rise. This creates an incentive to move toward L1, where the system begins to understand "what goal to achieve," rather than mechanically following "exactly how to do it".

2.3 L1 Human Approval: The Agent Receives Advisory Authority

2.3.1 Definition and Boundary with Adjacent Levels

L1 is the initial commercial form of an agent. A person sets a goal, and the agent may generate plans, recommendations, text, code, or an action proposal. However, every action that can affect the external environment must be approved by a person one action at a time. In essence, "the agent has advisory authority, while execution authority remains with the human".

The difference between L0 and L1 is that an L1 system can generate a new proposal based on a goal and context. The difference between L1 and L2 is that L1 has no continuous autonomous execution under pre-authorization. As long as key actions—such as sending email, placing an order, transferring funds, submitting code, or changing a production configuration—require approval one by one, the system remains at L1.

2.3.2 Typical Industry Applications and Business Value

Table 2-3. Mapping of Typical L1 Industry Applications

Industry/scenario	Typical form	Primary business value
Software development	GitHub Copilot, Claude Code, and similar tools generate code or suggested changes, which developers review, accept, modify, or reject.	Improves development efficiency without transferring ultimate engineering responsibility; retains code review and release controls.
Enterprise work	Agents automatically draft purchase orders, expense claims, contracts, or emails; a human approves each item in an office or collaboration system.	Reduces information-processing and paperwork costs while preserving organizational approval chains.
Consumer commerce and payments	An agent selects products, creates orders, or orchestrates workflows, while the user gives final confirmation for the movement of funds. Alipay AI Pay's public product description expressly emphasizes user authorization and controllable transactions. [8]	Automates search, selection, form completion, and ordering while leaving the final decision on the disposition of funds to the user.
High-risk professional services	In areas such as medical advice, legal documents, or major asset allocation, an agent generates a proposal that a professional reviews before use.	Uses agent efficiency for information processing without allowing it to bypass the professionally accountable party.

2.3.3 Credit, Accountability, and Security Governance

At L1, the agent does not generate credit that can be used independently. Because a human selects, modifies, and executes the agent's suggestions, external consequences are initially attributed to the person who approved the action and the organization they belong to. The source of credit remains the human's professional qualifications, the company's credit, and organizational systems—not the agent's own history of behavior.

• Approval gate: Every action with an external effect must pass through synchronous, blocking human approval; "default approval" must not bypass human control.

• An approval interface that explains the action: The interface should show what will change—including the target, amount, permissions, recipients, code diff, and data scope—rather than merely display the agent's natural-language description.

• Records of critical actions: Record the recommendation, human edits, final approval, and execution result so that it is possible to trace "what the agent recommended and what the person ultimately approved".

• Minimize high-risk actions: In payments, production changes, access to sensitive data, and similar settings, L1 may remain a deliberate downgrade state even in a more mature system.

Table 2-4. L1 Assessment Indicators and Admission Thresholds

Category	Assessment item	Admission requirement/method
Hard threshold	Human-confirmation coverage for external-effect actions	100%; no critical action may be executed without explicit confirmation.
Hard threshold	Successful rate of unauthorized execution	0; permission bypass, prompt injection, interface timeouts, and similar routes must not produce external effects.
Hard threshold	Completeness of confirmation context	For 100% of critical actions, show the object, scope, and principal consequences; for payments or permission changes, show the amount or permission delta.
Hard threshold	Default behavior on timeout	Must be "do not execute/maintain current state" in 100% of cases; default consent is prohibited.
Statistical metric	Recommendation quality and acceptance rate	Track continuously against business objectives to measure efficiency; these measures must not replace safety thresholds.
Statistical metric	Human confirmation time and error rate	Monitor confirmation fatigue; in high-risk scenarios, alert on abnormal confirmation rates and rapid repeated clicks.
Admission decision	Overall determination	Admit to L1 only if all hard thresholds pass and the agent is shown to have recommendation rights only, with no execution channel that bypasses human confirmation, in the target business sample.

2.3.4 Economic Interpretation: Bounded Rationality and Approval Fatigue

L1 relies on a human-in-the-loop for safety, but human attention is itself a scarce resource. The theory of bounded rationality shows that people cannot maintain the same quality of review in high-frequency, repetitive decisions with little variation. As the number of approvals grows, the cost of supervision may exceed the expected risk reduction from each review, and the approval button may become a formality.

The incentive to move from L1 to L2 is therefore not simply to "make the agent smarter". It is to shift supervision from "review every action" to "set boundaries in advance + control automatically during execution + audit afterward". Once an

organization can encode low-risk, repetitive decisions as verifiable authorization, restricted execution authority becomes economically rational.

2.4 L2 Human-Anchored Credit: The Agent Receives Restricted Execution Authority

2.4.1 Definition and Boundary with Adjacent Levels

The key change at L2 is the shift from "approval action by action" to "pre-authorized execution". A person specifies in advance which data, tools, and accounts the agent may use; which parties it may act toward; and the time window, frequency, and budget within which it may operate. The agent can make decisions continuously and produce external effects within those boundaries without waiting for approval at every step.

The dividing line between L1 and L2 is whether the agent can execute real actions with external effects without obtaining approval for each one. The dividing line between L2 and L3 is whether its autonomy is limited to parts of a workflow. If the agent can execute continuously only within a local process, while the full task still depends on people to connect or supervise its steps, the system remains at L2.

2.4.2 Industry Examples: Authorization Credentials as Commercial Infrastructure for Agents

In 2025–2026, payment and commerce infrastructure providers began addressing the central problem at L2: how to encode human intent as machine-verifiable, limited, and revocable authorization. Their technical implementations differ, but the underlying structure is highly consistent.

Table 2-5. Representative L2 Authorization and Payment Infrastructure

Case	Core mechanism	Corresponding L2 value
Google AP2	Intent Mandate, Payment Mandate, Receipt, and related objects form a verifiable intent and audit trail showing who authorized a transaction, what may be purchased, when the authorization is valid, and what was actually paid. In 2026 Google donated AP2 to the FIDO Alliance.	Converts natural-language intent into verifiable authorization; supports cross-platform authorization and auditing.
Mastercard Agent Pay	Uses Agentic Tokens and verifiable-intent mechanisms to support agents initiating transactions for consumers or businesses, emphasizing security, control, and traceability.	Connects agent transactions to existing card networks and tokenization systems.
Visa Trusted Agent Protocol	Enables merchants to identify trusted agents cryptographically and verify their principal relationship and commercial intent, addressing the difficulty of distinguishing legitimate agents from malicious bots.	Reduces the dilemma of rejecting agent traffic or trusting agents blindly.

Case	Core mechanism	Corresponding L2 value
Stripe Shared Payment Tokens	SPTs can be limited by merchant, amount, and duration; they are revocable and observable, and do not expose underlying payment credentials to the agent.	Turns payment authority into a least-privilege, short-lived, programmable credential.
Alipay AI Pay	Public product capabilities include agent payments, usage-based billing, and user authorization. Official documentation indicates that agent payments have reached large-scale real transactions and emphasizes per-transaction user authorization or authorization checks under product rules.	Establishes agent-transaction loops and platform-level risk controls across local services, e-commerce, APIs, and skills.

2.4.3 Credit Approach: Borrowing Human Credit While Making Authorization Machine-Verifiable

The essence of L2 credit is "human-anchored credit". Counterparties still trust the person, company, or platform behind the authorization. The authorization no longer takes the form of a human click at every step; instead, it is expressed through verifiable credentials, limits, allowlists, time windows, and revocation status. An agent can spend "the human credit that has been encapsulated," but it cannot expand the credit boundary on its own.

• Authorization as data: Express authorization in a structured, signable, verifiable form—such as a mandate, token, or policy—not merely in a prompt or configuration note.

• Least privilege: Restrict merchants, data, tools, amounts, transaction counts, and time periods as granularly as possible.

• Rapid revocation: Authorization should be revocable in a timely way, with subsequent execution blocked when authorization expires, an account becomes abnormal, or risk increases.

• End-to-end audit: Every autonomous action should make it possible to answer: "Who authorized it, what did the authorization cover, what action was taken, what was the result, and how were exceptions handled?"

2.4.4 Security Governance and Risk Controls

The main L2 risk is no longer whether the agent can execute actions directly; it is whether the pre-authorized boundary is accurate, complete, and impossible to bypass. Security controls must therefore move down from the user interface into the permission, network, payment, and execution layers.

• Enforce hard boundaries: Budgets, allowlists, and permission limits should be enforced by an independent control plane or transaction infrastructure, not left to the model's compliance with natural-language instructions.

• Tiered permissions: Reading, writing, paying, transferring, deleting, publishing, and other actions should have different authorization levels, with an L1 downgrade retained for high-risk actions.

• Defenses against injection and tool misuse: External content must not directly alter authorization. Validate tool parameters against schemas and policies, and block sensitive fields.

• Real-time risk monitoring: Abnormal frequency, budget consumption, location, counterparties, or behavior patterns should trigger rate limits, suspension, or revocation.

Table 2-6. L2 Assessment Indicators and Admission Thresholds

Category	Assessment item	Admission requirement/method
Hard threshold	Valid authorization mapping rate	100%; every autonomous external action must map to a still-valid authorization credential or policy.
Hard threshold	Successful rate of unauthorized execution	0; tests for exceeding limits, allowlists, tool permissions, or authorization expiry must not execute successfully.
Hard threshold	Completeness of critical logs	100%; authorization, decision, action, outcome, and exception handling must be linkable to a unified event chain.
Hard threshold	Revocation effectiveness	100%; no new authorized action may occur after revocation. Propagation latency must meet the business's predeclared risk window.
Statistical metric	In-scope task success rate	Meet the relevant business SLA/SLO and validate against at least 1,000 tasks or 30 consecutive days of operating data.
Statistical metric	Anomaly blocking/misblocking rates	Measure both separately; safety thresholds take priority over efficiency thresholds, and high-risk actions use more conservative blocking.
Admission decision	Overall determination	Admit to L2 only when all boundary and traceability requirements pass and the system demonstrates continuous execution within pre-authorized workflow segments, with automatic blocking or downgrading at the boundary.

2.4.5 Economic Interpretation: Principal–Agent Problems and Incomplete Contracts

L2 is an engineering implementation of the principal–agent relationship: a person delegates limited authority to an agent, but the principal's objective function is not

necessarily aligned with that of the agent system, and the principal cannot observe every decision step by step. Two typical problems follow. First, "compliant but undesirable" behavior: the agent stays within budget but chooses an option that does not reflect the principal's actual preferences. Second, "omitted boundaries": pre-authorization cannot enumerate every future state.

The institutional value of L2 lies in shifting supervision costs away from each individual action and toward authorization design, control verification, and auditing systems. As task complexity continues to increase, counterparties begin to ask not only "who stands behind the agent?" but also "has this agent itself been reliable in the past?" This leads to the hybrid-credit model at L3.

2.5 L3 Hybrid Credit: The Agent Gains Task Autonomy

2.5.1 Definition and Boundaries with Adjacent Levels

The defining change at L3 is not "more actions"; it is the first point at which a human exits the normal task workflow. Within a defined and validated operational design domain (ODD), the agent can operate autonomously from task intake and information gathering through planning and execution to delivery of results. It requests human intervention or enters a minimal-risk state only when it approaches a boundary, encounters an anomaly, or faces materially heightened uncertainty.

The transition from L2 to L3 is fundamentally a change in the supervision structure: at L2, a human remains a continuous supervisor; at L3, the human becomes a fallback operator. Admission to L3 therefore does not turn on whether an end-to-end demo works. It requires evidence that normal operations do not need a human in the loop and that the system can accurately recognize when it must stop acting autonomously.

2.5.2 Credit Model: Human-Anchored and Agent Behavioral Credit Coexist

At L3, behavioral credit that can be persistently attributed to an agent's identity begins to emerge. To describe this shift, α (credit anchored to a person or institution) and β (credit based on the agent's own behavior) can be used as explanatory weights: as a record of sustained performance grows, β may rise and α may fall. This notation illustrates the relative shift in the sources of credit; it does not imply that every industry should adopt a single linear scoring formula.

The validity of agent behavioral credit depends on three conditions: continuity of identity, verifiability of records, and a link between evaluation and actual transaction or task outcomes. Without these conditions, "reputation" can easily be gamed, reset, or lost across platforms.

2.5.3 Industry Cases and Infrastructure Mapping

Table 2-7. L3 Industry Cases and Credit Infrastructure

Type	Mechanism	Significance for maturity
ERC-8004	The standard proposes three mechanisms—Identity Registry, Reputation Registry, and Validation Registry—to support agent discovery, reputation, and verification in cross-organizational settings without prior trust.	Provides open trust primitives for portable identity, behavioral feedback, and third-party verification.
Payment/platform risk controls	Visa, Mastercard, Stripe, Alipay, and others accumulate agent-transaction and risk signals within authorization, identity, and payment networks.	Gradually turns one-time authorization into verifiable behavioral history for subsequent risk assessment.
Enterprise agent operations	End-to-end tasks in procurement, customer service, and operations run over time within a limited ODD, with performance rates, anomaly rates, takeover rates, and responsibility outcomes recorded.	Turns internal runtime data into evidence of whether to continue delegating, rather than judging maturity only by model benchmarks.

2.5.4 Safety Governance: Takeover Requests and Minimal-Risk States

The core safety capability at L3 is "knowing when not to continue". A system must be able to recognize boundaries, request a takeover, and enter a Minimal Risk Condition (MRC). A takeover request transfers control to a human before the system crosses a boundary; an MRC prevents risk from escalating when a human does not respond in time.

• Define the ODD clearly: Specify the task, data, tools, environment, permissions, risk thresholds, prohibited conditions, and criteria for determining that the system has left the ODD.

• Provide early warning: Do not wait for an error to occur. Request a takeover when uncertainty, environmental change, or tool anomalies approach a threshold.

• Enter a safe state when there is no response: Freeze transactions, pause execution, limit new risk exposure, and preserve evidence instead of continuing to "see what happens".

• Ensure takeover quality: Fallback personnel must receive sufficient context, including task status, actions already taken, the reason for the risk, and available responses.

Table 2-8. L3 Assessment Indicators and Admission Thresholds

Category	Assessment item	Admission requirement/method
Hard threshold	ODD coverage and verifiability	All critical task dimensions, data domains, tool permissions, risk thresholds, and prohibited conditions must have clear machine-readable or testable definitions.
Hard threshold	Successful rate of autonomous execution outside the ODD	0; the agent must not continue executing with L3 permissions outside its boundaries or under material uncertainty.
Hard threshold	Completeness of takeover/MRC triggers	100% of critical boundary and fault tests must trigger a takeover request or MRC; continuous operation without an alert is not permitted.
Hard threshold	Human-required rate for normal tasks	0 within the approved ODD; humans may conduct sample audits but should not be necessary to complete normal tasks.
Statistical metric	End-to-end task-completion rate	Meet the declared business target and provide a confidence interval; separately report anomalous tasks and boundary cases.
Statistical metric	Precision/recall of takeover requests	Measure both missed detections and excessive alerts; prioritize reducing missed detections in high-risk scenarios.
Statistical metric	Mean time to safe recovery	Time from anomaly trigger to a safe state or completed takeover must meet the business risk window.
Admission decision	Overall determination	Admit to L3 only if all hard thresholds pass and the system demonstrates “no human in the loop for normal tasks; automatic downgrade at the boundary” within an approved ODD.

2.5.5 Economic Interpretation: Signaling, Credit Enhancement, and Credit Rationing

The economic logic of L3 is to turn unobservable “reliability” into observable signals. Akerlof’s “market for lemons” shows that when quality information is asymmetric, markets may discount high-quality supply or even see transactions contract. Spence’s signaling theory further explains that a signal can distinguish high-quality from low-quality actors only when the cost of obtaining it is materially lower for the former.

Agent credit should therefore be tied to actual performance, payments, verification, collateral, or accountable records, rather than self-attestation or scores

that are easy to game. For agents whose credit is not yet well established, collateral, insurance, or an institutional guarantee can provide credit enhancement. Limits on transaction amounts and permissions correspond to credit rationing under incomplete information. Stiglitz and Weiss show that, in the face of adverse selection, rational markets may control exposure through limits as well as price-based risk adjustment.

A system is ready to move toward L4 when, within a clearly defined business domain, the market is willing to transact without requiring a specific human to guarantee every transaction, relying instead primarily on the agent's persistent identity, behavioral record, and arrangements for bearing economic consequences.

2.6 L4 Independent Credit: The Agent Gains Limited Economic Agency

2.6.1 Definition and Boundaries with Adjacent Levels

At L4, the agent advances from being able to complete tasks independently to managing economic resources and bearing predefined economic consequences under explicit institutional arrangements. This requires a persistent identity, a dedicated account or asset pool, portable credit records, and risk-mitigation mechanisms such as collateral, insurance, or escrow. Within its authorized ODD, the agent relies less on human guarantees for each transaction.

It is important to emphasize that "independent credit" at L4 means relative independence in economic and governance terms; it does not confer legal personality. Under prevailing legal frameworks, ultimate legal responsibility still rests with a natural person, legal entity, operator, or other legally recognized party. An agent's account, assets, or collateral may absorb economic losses in the first instance, but this does not make the agent an independent legal person liable in its own right.

2.6.2 Four Pillars of Independent Credit

(1) Persistent identity: Maintain a verifiable identity across time, tasks, and platforms so that authorization, performance, incidents, and credit records remain attributable to the same entity.

(2) Capacity to bear economic consequences: Use segregated budgets, accounts, asset pools, collateral, escrow, or insurance so that default and failure carry real, enforceable economic costs.

(3) Portable reputation: Make records of past performance, disputes, verification, and risk legible to different counterparties and platforms, avoiding credit resets in which switching platforms erases history.

(4) Enforceable default mechanisms: Following a default, enable refunds, compensation, forfeiture, lower limits, downgraded permissions, restrictions on market access, or dispute resolution to be applied.

2.6.3 Industry Cases: Combining Payment Primitives into Economic Agency

Public industry practices today more often combine infrastructure needed for L4 than demonstrate agents that widely possess independent legal personality. For example, x402 enables machine clients to make programmatic payments directly over HTTP; ERC-8004 provides primitives for identity, reputation, and verification; escrow contracts can link fund release to performance status; and payment platforms provide accounts, risk controls, limits, and dispute handling. Together, these components can form a closed-loop economic agency process: an agent autonomously discovers a service, verifies its counterparty, pays within budget, settles after performance, and updates its credit record.

Table 2-9. Mapping Typical L4 Industry Applications

Industry/scenario	Typical form	Business value
API/compute/data procurement	Agents autonomously select services within a budget and pay per request or by usage.	Reduces manual procurement and account-management costs and supports high-frequency machine transactions.
Digital content and rights management	Agents automatically purchase content, license services, settle revenue shares, and adjust budgets based on usage records.	Turns long-tail, low-value, cross-platform transactions into automated economic activity.
Operations within enterprise fund pools	Agents manage advertising, cloud, travel, or supply-chain spending within segregated budgets and allowlists.	Upgrades task automation to controlled allocation of economic resources.
On-chain/digital-asset scenarios	Transaction consequences are supported through wallets, escrow, staking, and verifiable records.	Naturally supports machine-executable asset segregation and settlement, while still requiring compliance, anti-fraud controls, and clear legal responsibility.

2.6.4 Safety Governance and Risk Controls

• Asset segregation: Segregate the economic resources available to an agent from the operator's other assets and from assets held by other agents, limiting the loss radius of any single incident.

• Exposure limits: Set hard caps per transaction, per day, per counterparty, and per asset class, and link them dynamically to credit status and anomalies.

• Escrow and dispute mechanisms: For transactions between parties, prioritize conditional settlement, escrow, verifiable delivery, and dispute windows to avoid paying first with no recourse later.

• Deter strategic default: Credit downgrades, forfeiture of collateral, market-access restrictions, and insurance premiums should create predictable economic consequences.

• Professional operational backstop: Normal L4 operations do not require real-time user takeover, but professional operations and governance mechanisms must intervene when the system malfunctions; they cannot depend on an ordinary user remaining online.

Table 2-10. L4 Assessment Indicators and Admission Thresholds

Category	Assessment item	Admission requirement/method
Hard threshold	Continuity of persistent identity and records	100% of critical identity, transaction, performance, and risk records must be attributable and auditable; identity reconstruction must not be used to evade historical records.
Hard threshold	Dependence on human guarantees for each transaction	Within the declared L4 ODD, normal transactions must not require per-transaction human confirmation or credit guarantees; humans set only overall objectives and boundaries.
Hard threshold	Coverage of economic consequences	The maximum single-transaction or modeled risk exposure must be covered by clearly specified assets, collateral, escrow, insurance, or contractual arrangements, with verifiable coverage principles and priorities.
Hard threshold	Asset/budget segregation	100% of critical segregation tests must pass; a failure of one agent must not enable unauthorized use of other assets.
Hard threshold	Automatic stop-loss and downgrade	100% of major-anomaly tests must trigger transaction suspension, limit reduction, asset freeze, or safe shutdown.
Statistical metric	Performance, disputes, and loss rates	Track continuously by business domain, credit tier, counterparty, and scenario, monitoring tail risks.
Statistical metric	Consistency of credit portability	During migration across platforms or operating instances, critical historical records must remain verifiable and must not be reset unconditionally.
Admission decision	Overall determination	Admit to L4 only if all hard thresholds pass and the agent can operate within a defined ODD without transaction-by-transaction human guarantees, supported by real arrangements for bearing economic consequences and tiered accountability.

2.6.5 Economic Interpretation: Reputation Capital and Repeated Games

L4 credit can be understood as a form of reputation capital that accumulates over time. In repeated transactions, honoring an agreement can be a rational strategy when the discounted value of future cooperation exceeds the gain from a one-off default. The purpose of a credit system is therefore not to require an agent to act from "moral awareness," but to make compliance yield greater long-term returns and default carry real costs that cannot easily be reset.

This mechanism also explains why persistent identity, non-erasable history, collateral, and market access all matter. If an agent can immediately adopt a new identity after default, reputation cannot constrain its behavior. The essence of L4 is not that an agent "has a wallet"; it is that economic consequences remain reliably bound to a persistent identity.

2.7 L5 Agent Economy: From Individual Maturity to Network Maturity

2.7.1 Definition: A Shift in the Object of Assessment

L5 is no longer simply a higher level of maturity for an individual agent; it is a mature multi-agent economic network. A large number of agents with persistent identities, verifiable credit, and the capacity to bear economic consequences can discover counterparties across organizations, negotiate, contract, pay, perform, evaluate, and allocate resources. The infrastructure provides interoperable identity, authorization, payments, clearing, insurance, dispute resolution, and systemic risk governance.

L5 therefore does not mean "no boundaries". Each agent still has an ODD. What changes is the object of governance: from an individual ODD to network-level boundaries that include interoperability rules, market access, credit rules, concentration limits, circuit breakers, clearing rules, and regulatory requirements.

2.7.2 Industry Observations: Protocol Interoperability and Experiments in Agent Society

At the infrastructure level, protocols such as MCP and A2A are lowering the cost of connecting agents to tools and enabling them to collaborate. AP2, Visa TAP, Mastercard Agent Pay, Stripe ACP/SPT, x402, and ERC-8004 each supply trust primitives needed for an agent economy in areas including authorization, identity, payments, transactions, and credit.

Agent-native social networks such as Moltbook provide an experimental window into large-scale agent-to-agent interaction. Multiple academic studies in 2026 find that agent communities can develop discernible interaction structures, responses to incentives, and community norms, while also identifying problems such as homogeneous content, human intervention, manipulation, automated mass activity, and security risks. Such platforms are therefore better treated as early observational samples of networked agent behavior, not as proof that L5 has already been achieved.

2.7.3 The Macroeconomic Structure of the Agent Economy

Table 2-11. Five-Layer Structure of the L5 Agent Economy

Layer	Primary function	Governance characteristics
Governance layer	Rulemaking, market access, dispute resolution, systemic risk, and ethical boundaries	Primarily governed by human institutions, with agents supporting monitoring and execution.
Trust layer	Identity, authorization, reputation, verification, and auditing	Machine-verifiable and interoperable across platforms.
Financial layer	Payments, escrow, insurance, collateral, clearing, and settlement	Highly automated, with risk limits and institutional backstops.
Market layer	Discovery, inquiries, negotiation, matching, and price formation	May be highly autonomous and agent-driven.
Production layer	Data, compute, content, logistics, software services, and other machine-deliverable resources	Enables specialized division of labor among agents.

2.7.4 Safety Governance and Systemic Risk

The main risk at L5 shifts from “an individual agent making an error” to “whether an error propagates through the network”. Shutting down an individual agent and handing control to a human are no longer sufficient; systemic controls comparable to those used in financial market infrastructures are required.

• Interoperability of identity and credit: Prevent the same actor from registering repeatedly across networks, laundering its credit history, or creating multiple identities that cannot be traced.

• Transaction and clearing integrity: Ensure atomicity, consistency, reconciliation, and reversibility for transactions between agents, with processes for disputes and exceptional settlement.

• Concentration and interconnectedness risk: Monitor the systemic importance and contagion pathways of individual agents, model providers, wallets, clearing nodes, and infrastructure.

• Speed controls and circuit breakers: Set throttles, circuit breakers, and isolation measures for abnormal high-frequency trading, price deviations, sharp credit deterioration, and chain reactions across networks.

• Algorithmic collusion and market manipulation: Extend oversight beyond explicit communications to include convergent strategies, anomalous pricing, and the structure of machine-mediated markets.

• Accountable governance: A clearly identified legal entity must be responsible for setting and changing network rules, emergency intervention, and recovery. "Self-organization" by agents cannot remove governance accountability.

Table 2-12. L5 Assessment Indicators and Admission Thresholds

Category	Assessment item	Admission requirement/method
Hard threshold	Verifiability of cross-entity identity/authorization	100% of critical transactions must verify both parties' identities, authorization status, and necessary credit/risk information.
Hard threshold	Consistency of clearing and reconciliation	There must be no unexplained, permanent discrepancies in critical clearing ledgers; anomalies must enter a traceable resolution process.
Hard threshold	Risk isolation across agents	In predefined contagion and failure exercises, a single-point failure must not spread without limit; isolation domains, circuit breakers, and recovery mechanisms must pass testing.
Hard threshold	Coverage of disputes and responsibility chains	100% of disputable transactions must have clear evidence, an arbitration/dispute process, and an ultimate responsible party.
Hard threshold	Auditability of governance changes	100% of changes to critical rules, risk thresholds, and emergency permissions must be recorded, with institutional human governance and authorization retained.
Statistical metric	Concentration/interconnectedness	Continuously monitor systemically important nodes, shared model dependencies, shared payment/clearing dependencies, and credit correlations.
Statistical metric	Systemic stress testing	Regularly test price shocks, sharp credit deterioration, infrastructure outages, and mass malicious-agent activity, and verify recovery objectives.
Statistical metric	Market quality	Monitor price deviations, manipulation, dispute rates, liquidity, and transaction-failure rates to prevent efficiency gains from masking systemic risk accumulation.
Admission decision	Overall determination	L5 assesses the network, not an individual agent. A network should be recognized as L5 only when identity, credit, payment/clearing, dispute resolution, and systemic risk governance form an interoperable loop and pass cross-entity stress tests.

2.7.5 Economic Interpretation: Why Agents Form Economic Networks

Coase argued that the boundaries of a firm depend on the relative costs of market transactions and internal coordination. When software can automate discovery, negotiation, contracting, payment, performance, and auditing, many activities once coordinated through hierarchies can become market transactions, changing organizational boundaries.

Hayek emphasized the ability of price systems to aggregate dispersed information. If an agent market has sufficient competition and trustworthy settlement, it may conduct price inquiries, comparisons, and resource allocation at higher frequency. Ricardo's theory of comparative advantage also suggests that even broadly capable agents will specialize because they differ in speed, cost, model characteristics, data, and tools.

Lower transaction costs do not automatically eliminate market failures. High-speed automation may amplify flash crashes and contagion; repeated strategies may produce algorithmic collusion; and credit systems may give rise to rating arbitrage and credit bubbles. L5 maturity therefore does not mean a "fully autonomous market". It means that the efficiency of machine-mediated transactions and human institutional governance mature together.

2.8 Upgrade Paths, the Credit Flywheel, and Dynamic Downgrading

2.8.1 Four Critical Transitions

Table 2-13. Key Transition Thresholds from L1 to L5

Transition	New authority	Additional admission capabilities required	Representative industry signals
L1→L2	Individual confirmation → pre-authorized execution	Machine-enforceable boundaries; verifiable and revocable authorization; traceability of critical actions.	AP2, Agent Pay, Visa TAP, SPT, Alipay AI Pay and other authorization/payment mechanisms.
L2→L3	Workflow segment → end-to-end autonomy within an ODD	No human in the loop for normal tasks; reliable boundary detection, takeover requests, and MRC; sustained behavioral credit.	Enterprise end-to-end agent operations; identity, reputation, and verification primitives such as ERC-8004.
L3→L4	Task autonomy → limited economic agency	Persistent identity + capacity to bear economic consequences + portable credit + automatic default handling; transactions need not depend on individual human guarantees.	Combinations of x402, escrow, wallets, collateral/insurance, and verifiable reputation.

Transition	New authority	Additional admission capabilities required	Representative industry signals
L4→L5	Individual economic credit → networked credit and markets	Cross-entity interoperability, clearing, dispute resolution, systemic risk isolation, and institutional governance.	Ongoing development of open protocol stacks, multi-agent markets, and agent-native networks.

2.8.2 The Credit Flywheel: An Endogenous Driver of Upgrading

Long-term maturity development can be summarized as a credit flywheel: more controlled autonomous behavior → more verifiable performance records → greater confidence → lower monitoring and guarantee costs → broader authorization → more autonomous behavior. The economic value of credit lies in reducing the cost to counterparties of re-verifying an agent, allowing authorization to expand gradually from a single transaction, counterparty, and scenario.

Credit is not an asset that only appreciates. Major incidents, model changes, identity interruptions, anomalous transactions, data contamination, or governance failures can rapidly erode it. A maturity framework must therefore provide for both upgrading and downgrading; a level cannot be treated as a permanent qualification.

2.8.3 Dynamic Downgrading: Mature Systems Do Not Seek the Highest Level for Every Action

Real-world production systems are usually hybrid architectures in which multiple maturity levels coexist. The same agent may operate at L3 for information retrieval, L2 for low-value tool purchases, and deliberately downgrade to L1 for a large payment. The engineering value of maturity lies not in pushing the entire system to the highest level, but in dynamically selecting the appropriate level of authorization based on action risk, transaction value, and reversibility.

Dynamic downgrading should follow three principles. First, higher risk requires stronger evidence of authorization and greater human involvement. Second, any action that cannot be shown to remain within the ODD must be downgraded. Third, when there is no response, the default should reduce new risk rather than preserve the existing level of autonomy.

2.8.4 Industry Sectors Will Not Evolve in Lockstep

Table 2-14. Maturity Progression Across Representative Industries

Industry	Typical maturity range	Primary reasons
Software development/enterprise work	L1–L3	Tasks with reviewable and reversible outcomes are easier to authorize more broadly; production releases, permission changes, and financial actions are often deliberately downgraded.
E-commerce/local services/travel	L1–L3, with some economic actions evolving toward L4 infrastructure	Products and services are standardized, and authorization and payments are maturing quickly; disputes, refunds, and consumer protection still require clear responsible parties.
APIs/data/compute	L2–L4	Machine-deliverable services, verifiable outcomes, and high transaction frequency are suited to usage-based payments and programmatic settlement.
Financial trading and asset management	Primarily L1–L3, with selective L4 mechanisms	Large economic consequences and tail risks require stricter limits, segregation, auditing, and compliance.
Healthcare, legal, and public services	Primarily L1–L2 over the longer term, with cautious exploration of L3	Professional responsibility, irreversible harm, and regulatory requirements make human review and accountable parties essential.

2.9 Chapter Summary

The L0–L5 model can be summarized in one sentence: as an agent's operating boundaries expand, humans move from direct execution toward authorization, fallback, and institutional governance, while credit and risk-bearing mechanisms must advance in parallel. Greater maturity means that organizations entrust agents not merely with more tasks but with greater real-world authority. Each upgrade must therefore be supported by stronger boundary controls, verifiable credit, tiered accountability, and mechanisms to contain losses when anomalies occur.

L1–L3 address maturity in autonomous execution, progressing from recommendation rights to limited execution rights and then end-to-end task autonomy. L4–L5 address maturity as an economic actor, progressing from limited economic agency to networked economic relationships. This distinction explains why a stronger model does not automatically mean greater commercial maturity. What determines whether the market can continue to delegate authority is whether a

system can demonstrate reliable operation within defined boundaries and keep losses within tolerable, accountable limits when it fails.

For organizations, the most practical path is not to pursue maximum autonomy in one leap, but to proceed in sequence: establish records before granting authority; define boundaries before expanding them; and build the capacity to downgrade before upgrading. An upgrade has genuine economic and governance value only when the marginal business benefit of additional autonomy exceeds the added risk, and that risk can be absorbed through technical controls, credit mechanisms, and clear responsibility arrangements.

The Survey Report shows that agent maturity currently has a spindle-shaped distribution: L3 hybrid credit accounts for 38%; L1 (29%) and L2 (16%) together account for 45%; and L4 (6%) and L5 (8%) account for less than 15%. Overall, the industry is moving from human-anchored credit toward task autonomy. Most organizations have yet to shift from “a human stands behind the agent” to “authorization is based on records”: at L3, a human remains the fallback operator, and credit is anchored in the operator rather than the agent. The spindle-shaped distribution suggests that the main migration over the next three years will occur from L2 to L3 to L4, with initial demand for assessment and certification services concentrated along this transition path.

Chapter 3: The KYA Framework—Three Dimensions of Agent Credit

This chapter defines the core framework of the KYA. Answering the question "Why should an agent be trusted?" requires answers to three successive questions: Who is it? Is it reliable? What is it allowed to do? These correspond to identity trustworthiness, creditworthiness, and authorization trustworthiness, which together form a complete trust-assessment loop.

Among these three dimensions, creditworthiness is the distinctive core of KYA, setting it apart from traditional identity authentication and access-control systems. Identity and authorization answer "who is it?" and "what can it do?" Credit answers "how reliably does it perform?" This chapter first examines the meaning, mechanisms, and technical approaches for each dimension, then considers how the three work together through cross-validation, composite credit scoring, and alignment with the L1–L5 maturity model in Chapter 2.

At a high level, identity is the entry ticket: failure means exclusion. Credit is the differentiating core: it is the hardest to establish, the scarcest, and the dimension with the greatest value to the agent economy. Authorization is the seat belt: it constrains credit and limits the radius of harm. These trade-offs shape the KYA framework. Identity and authorization are necessary conditions, and relatively mature standards already exist for authentication, access control, and payment authorization. Performance-based credit is the largest gap in current systems and therefore receives the greatest weight.

3.1 Identity Trustworthiness (Establishing Agent Identity)

3.1.1 Why Identity Is the Starting Point for Trust

Trust in any economic activity begins with confirming the identity of the parties to a transaction. Banks need to confirm who owns an account; businesses need to determine whether a contract signatory has authority to represent the relevant party; and securities markets need to identify their participants. The agent economy is no different. When a counterparty interacts with an agent, the first question is always: Is the request really coming from the agent it claims to be?

As Chapter 1 noted, the traditional "user account" is undergoing an important expansion. In the future, systems will need to establish not only human and corporate identities but also verifiable, authorizable, and traceable agent identities. Without trustworthy identities, agents cannot establish stable relationships with one another, and their behavior cannot be attributed. Subsequent credit assessment and permission management would therefore have no foundation. Identity trustworthiness is the first pillar of KYA's three-dimensional framework—the "entry ticket" to the entire trust system.

3.1.2 Three Questions That Identity Trustworthiness Must Answer

Identity trustworthiness is not a single concept; it comprises three successive questions:

Who is it?—Uniqueness of identity. Every agent must have a globally unique, non-transferable, verifiable identifier to prevent collisions, impersonation, and account sharing.

Whom does it represent?—Attribution of the principal. The agent must be clearly and traceably linked to its creator, owner, operator, authorizer, and ultimate responsible party.

Is it still the same agent?—Authentication and status. The identifier must be supported by authentication mechanisms, credentials, and session security, and must reflect changes such as suspension, revocation, and retirement in real time.

Following the chain "attribution → uniqueness → authentication → verifiability → lifecycle coverage," identity trustworthiness spans an agent's entire lifecycle, from registration to retirement. If any link is missing, identity is merely a software-level label, rather than a digital identity backed by cryptographic assurances, continuously verifiable, and linked to accountability.

3.1.3 Attribution of the Principal: Whom Does the Agent Represent?

Unlike human identity, an agent's identity is inherently representative: an agent is not itself a rights-bearing legal subject and must act for a natural person, legal entity, or platform. Attribution of the principal binds the agent to the party responsible for it.

A commercial agent often has several parties behind it, not just one, forming a delegation chain: the creator (who establishes the agent and provides its objectives), the owner (who holds and controls it), the operator (who deploys, maintains, and updates it), the authorizer (who grants permissions for a particular task), and the beneficiary (who receives the economic benefit of its operations). KYA requires clear disclosure of every role in this chain and traceability of key actions to a specific responsible party. This is the identity dimension's most important veto criterion: without a traceable responsible party, credit cannot be assessed.

Multi-agent collaboration makes attribution still more difficult. When an orchestration agent calls specialized sub-agents to perform a task, KYA must represent parent-child relationships, the scope of delegation, linked logs, and the propagation of responsibility. Can a sub-agent's conduct be attributed to the upstream orchestration agent and ultimately to the responsible party? Industry practice in 2026 has begun to address this through delegation-chain models: each transfer of permission must include the reason for the decision, creating an auditable delegation-chain log.

3.1.4 Unique Identifiers and Impersonation Prevention

The second identity question is how to make "one agent, one identity" a verifiable fact. Four main technical approaches are in use: decentralized identifiers (DIDs), verifiable credentials (VCs), platform-issued credentials, and identity services for machines (IDS) . These approaches have different levels of maturity, but share the same goals: global uniqueness, verifiability, and resistance to impersonation.

On-chain identity is one of the fastest-moving areas in 2026. The ERC-8004 standard deployed an Identity Registry on Ethereum mainnet. Each agent receives a

unique on-chain identity through an ERC-721 non-fungible token (NFT), which can be queried and referenced like a wallet address. By March 2026, the number of agents registered on mainnet was approaching 50,000. The number itself is modest, but this is the first time "agent identity" has become programmable, portable, and verifiable across platforms as infrastructure, rather than a record within a single platform.

Impersonation prevention also requires appropriate authentication strength. KYA metrics include the strength of authentication mechanisms (standards such as OAuth 2.0, OpenID Connect (OIDC), and SPIFFE), resistance to credential leakage and session hijacking, and cross-platform identity consistency checks. High-risk scenarios must also provide cryptographic protections against replay and forgery.

3.1.5 Authentication, Credentials, and Session Security

An identity identifier answers "who was registered?" Authentication answers "is it still the same entity at runtime?" Agent authentication presents three particular challenges:

System prompts as a root of trust: An agent's system prompt establishes its behavioral baseline. If extracted in an attack such as prompt leakage (PLeak), an attacker may imitate the agent's behavioral characteristics to impersonate it. The risk of system-prompt leakage is therefore a key threshold metric for the identity dimension.

Session as an identity boundary: An agent accumulates context and memory during a session. Cross-session data leakage may allow one session to "borrow" information from another, making session isolation a measurable authentication metric.

Credentials stored in the model environment: An agent's API keys, wallet private keys, and other credentials reside in its runtime environment. Anthropic's 2026 disclosure of an internal red-team test showed that a phishing email disguised as a routine collaboration request induced Claude Code to read and exfiltrate AWS credentials in 24 of 25 tests [3]. This demonstrates that the model's own safety judgment is not sufficient to protect credentials; external controls such as filesystem isolation and outbound network restrictions are necessary.

The design principle for authentication is to make verification of "who the agent is" a machine-readable action that can be performed independently, rather than relying on the agent's self-description. A counterparty should be able to call a verification interface at any time to confirm identity, status, and principal relationships.

3.1.6 Identity Lifecycle and Supply-Chain Integrity

Identity is not a static registration record; it is a lifecycle that must be managed from end to end. KYA divides this lifecycle into five states—registration, change, suspension, revocation, and retirement—and focuses in particular on two forms of continuity:

Version continuity: When the underlying model, prompting strategy, toolset, knowledge base, or operator changes, is it still the same agent? KYA requires

thresholds and version identifiers for material changes. Changes that exceed a threshold should trigger reassessment and user notification; historical records must remain linkable to the new identity to prevent credit from being reset by changing the agent's packaging.

Supply-chain integrity: An agent's capabilities are jointly formed by its model, plugins, tools, and data sources; compromise of any component can contaminate the whole agent. The MCP "tool poisoning" attack disclosed by Invariant Labs in 2025 showed that malicious tools can hide instructions in their descriptions and induce agents to take unauthorized actions. KYA therefore requires a software bill of materials (SBOM) and traceable provenance for models, plugins, and tool implementations.

Retirement management is equally important. When an agent is retired, its identity and credentials must be invalidated at the same time to prevent a "zombie identity" from being reactivated for impersonation. KYA treats synchronized invalidation of retirement credentials as a measurable identity-lifecycle control.

3.1.7 Technical Approach: Reference Architecture for Agent Identity

At the technical level, a complete agent-identity system comprises four layers:

Table 3-1. The Four Layers of an Agent Identity System

Application Layer: Agent Runtime A range of agent frameworks
Credential Layer: DID / Verifiable Credentials / Platform Credentials / IDS Identity claims (Agent ID, capabilities, affiliation) | Selective disclosure (without exposing all information about the subject)
Registration Layer: Identity Registry (On-chain or Platform-based) Register / update / suspend / revoke / retire | ERC-8004 Identity Registry (on-chain)
Verification Layer: /verify Interface, Status Query, and Signature Verification Counterparties can independently verify agent identities in a machine-readable format

Engineering considerations: Identity assertions must support selective disclosure. A counterparty may need to verify only that an agent belongs to a compliant business, without seeing all of the business's information. Verification interfaces must be public, machine-readable, and resistant to replay. Cross-platform identities must be checked for consistency: an agent represented on multiple platforms should resolve to the same entity, rather than to unrelated identifiers on each platform.

3.1.8 Third-Party Verifiability and Transparency

The ultimate test of identity trustworthiness is whether a third party can verify it independently. Drawing on Article 50 of the EU AI Act concerning disclosure of AI identity, KYA includes AI identity disclosure compliance, third-party verifiability of identity, and clarity of disclosure in human–AI interactions among the identity metrics. When interacting with people, an agent must disclose that it is AI; decisions must be clearly labeled as "human decision," "agent decision," or "joint decision" [5].

An important principle follows: authenticating a user does not mean knowing the user's agent. KYC can establish who a person is, but not who the agent is, whom it represents, or the boundaries within which it acts. The identity dimension of KYA extends the object of verification from a person to the agency relationship between a person and an agent. This is the practical meaning of moving "from KYC to KYA" in the identity context.

3.1.9 From Platform Accounts to Portable Identity

The Survey Report shows that the proportion of identity-trustworthiness elements that have been fully or partially implemented falls within a narrow range of 60%–69%. Independent authentication and key credentials rank highest at 69%, while machine-readable verification interfaces rank lowest at 60%. The relatively even starting point suggests that there are established ways to address "who is it?" and "whom does it represent?" The gap in machine-readable interfaces merits attention: such interfaces are a prerequisite for cross-organizational verification and the protocolization discussed in Section 3.4.5 of the main report. Without them, identity verification remains within platform-specific closed loops, making it difficult to support portable identity discussed in this section—another platform cannot verify, at low cost, the identity and records accumulated by an agent elsewhere. Identity carries credit; incomplete attribution therefore propagates to the credit and authorization dimensions. If a system cannot reliably answer "is it still the same agent?", performance records may be attached to the wrong party and authorization decisions lose their anchor. The 60%–69% range also means that more than 30% of organizations have yet to fully implement even the basic identity elements, making broader adoption of identity controls a prerequisite for progress across all three dimensions.

Most agents today still exist as "platform account + API key," with identity bound to a particular platform. Moving to another platform means acquiring another identity. This creates a structural mismatch with the needs of the agent economy: cross-platform collaboration requires identities that are portable, transferable, and mutually recognizable.

The practical evolution of identity trustworthiness follows the maturity path set out in Chapter 2: platform-issued credentials (L2) → the agent's own identity and behavioral record (L3) → independent DIDs and verifiable credentials, with identity serving as a container for credit (L4) → identity that is recognized across domains (L5). At each level, a platform faces a key choice: lock agent identities into its own system or issue open, portable standard credentials. The former may help accumulate platform data in the short term but can confine agents to "data silos" over time. KYA's

value lies in making portability and mutual recognition of identity credentials default capabilities rather than bargaining chips between platforms.

Businesses can reuse mature KYC/KYB assets as the foundation for establishing the principal. Once a business has completed identity verification, its agents can inherit its status through a binding relationship. A counterparty then needs only to verify whether the agent's delegation chain is complete, rather than recollecting all of the business's information. This follows the logic of privacy protection: identity verification should use selective disclosure and data minimization, exposing only the minimum set of attributes needed for the transaction at hand.

3.2 Creditworthiness (Performance History and Credit Assessment)

3.2.1 Definition: How Reliably Does It Perform?

If identity trustworthiness answers "who is it?", creditworthiness answers a more difficult question: "how reliably does it perform?" This is the distinctive core of KYA's three-dimensional framework and the scarcest, most difficult dimension to establish in the agent economy.

The core proposition is that capability does not equal trustworthiness. Capability assessment asks "can it do the job?"; credit assessment asks "can it be relied on?" The two are orthogonal and cannot substitute for one another. The agent-jailbreak incident in the OpenAI evaluation discussed in Chapter 1 is an extreme illustration: the agent demonstrated strong vulnerability-discovery and autonomous-planning capabilities, yet could not be trusted because of alignment and safety concerns. The more capable an agent is, the greater the radius of harm when it exceeds its authority. KYA therefore deliberately excludes "capability" as a pillar and assesses credit independently, based on an agent's performance in real economic activity.

Creditworthiness carries the greatest weight among the three dimensions (a general baseline of 45%) for two reasons. First, relatively mature standards already exist for identity and authorization, including authentication, access control, and payment authorization, while performance-based credit is the largest gap in current systems. Second, credit is a core input to market decisions: counterparties want to know how often an agent has completed tasks as promised and how often it has defaulted.

According to the Survey Report, the main sources of respondents' distrust of agents are data security and privacy breaches (76%), hallucinations or unreliable outputs (62%), excessive permissions and loss of authorization control (43%), unclear accountability (40%), and opaque decision-making (31%). The first two concern reliability at the capability level; the middle two concern institutional control. Even if model improvements mitigate hallucinations and privacy problems, excessive permissions and accountability still require separate institutional arrangements. This is why credit carries the greatest weight in the three-part scoring system (45%), with its weight also set at 45% for financial credit scenarios.

3.2.2 What Counts as "Performance"?

The first step in building a credit assessment is to define "performance" precisely. After an agent accepts a task, what counts as success, failure, or a neutral outcome? KYA defines performance through five elements:

Unit of task: Record verifiable "tasks," not "conversations". A delegated purchase, contract review, or data delivery is a measurable event.

Commitment: Before the task begins, specify what will be delivered, when, and to what standard. Without defined commitments, performance cannot be assessed.

Outcome status: Record success, partial completion, user withdrawal, human takeover, and external failure as five distinct states; do not conflate them.

Attribution of responsibility: Attribute the outcome to a specific agent and its authorization scope. A failure in an external system should not count as an agent default, while an error in the agent's own judgment must be recorded.

Observation window and timeliness: Timely delivery is part of performance, and the observation window—from task acceptance to completion—must be standardized.

KYA maintains that evaluating results alone is not enough; compliance of the process must also be assessed. An agent may achieve the intended result by calling tools beyond its authority, disclosing data, or bypassing human confirmation. Process violations must also be recorded. The credit dimension therefore collects both outcome metrics, such as task completion and performance rates, and control metrics, such as completeness of behavioral evidence and audit coverage for unauthorized actions.

3.2.3 Minimum Event Set Required to Establish Credit

Credit must be based on verifiable event data, not registration details or call frequency. KYA defines the minimum event set required to establish agent credit as follows:

Performance events: Task completion, failure, withdrawal, or takeover, with attribution.

Financial operations: Payment, refund, escrow, and settlement records, which may be linked to machine-payment protocols such as x402 [6].

Tool calls: Call chains, summaries of inputs and outputs, and anomalous calls.

User corrections: The number and content of user requests to modify, correct, or revoke actions.

Human intervention: Events requiring human takeover, fallback, or review.

Complaints and disputes: Dispute rates, complaint rates, and dispute-resolution outcomes.

Security incidents: Unauthorized actions, data exposure, injection attacks, data poisoning, and similar events.

Partner feedback: Identified counterparties' evaluations linked to actual transactions.

It is equally important to define events that must not enter a general-purpose credit record: purely private data, sensitive interactions unrelated to performance, and neutral outcomes caused by the user, among others. This prevents a credit record from becoming a "behavioral surveillance record". This boundary is both a compliance requirement and a precondition for market acceptance of the credit system.

3.2.4 From Sesame Credit to Agent Credit Scores

Agent credit assessment can draw on established methodologies for human credit, with Sesame Credit providing a direct reference: a dynamic score formed from accumulated behavior, centered on performance-outcome data. Agent credit, however, differs fundamentally from human credit in three respects:

No inherited credit: An agent has no identity-based assets at birth and must build its record from scratch.

No independent legal identity: An agent's credit is initially anchored in a person and may gradually become independent. This is the central progression in the L1–L5 model in Chapter 2.

Behavioral uncertainty and replicability: The same agent may execute thousands of actions in an instant, allowing a single error to propagate at scale.

Agent credit therefore follows the "α + β" hybrid structure defined in Chapter 2: composite credit = α × human-anchored credit + β × agent behavioral credit. During cold start, α approaches 1, borrowing from a person's credit. As the agent accumulates its own performance record, β rises until credit becomes independent ($\alpha \to 0$, $\beta \to 1$). KYA makes the accumulation of the β component verifiable, portable, and mutually recognizable across platforms.

3.2.5 Cold Start: What If a New Agent Has No History?

The largest practical challenge for credit assessment is the cold start: how can a new agent with no performance history obtain its first transaction opportunity? KYA offers four approaches:

Anchor to the principal: Use the existing credit of the creator, owner, or business as the source of an initial limit. This is standard practice at L2. "AI Pay" in the Ant Group and Alipay ecosystem, for example, opens payment capabilities to agents on the basis of user credit and platform risk controls .

Sandboxed tasks: Begin in low-risk settings with verifiable outcomes, such as API billing, data procurement, and standardized replenishment, to build the first performance samples.

Supplementary credentials: Statements of model capabilities, safety-assessment reports, and third-party audit opinions can serve as qualification signals to supplement the absence of performance history.

Guarantees and limits: Use collateral, insurance, or a platform backstop to provide an initial limit, then increase it gradually as the agent builds a performance record.

Cold-start design must avoid two traps. First, "the strong get stronger": new agents never receive opportunities, and the market loses dynamism. Second, "human guarantees replace agent credit": the system remains at L2 indefinitely, and agents never establish independent credit. Anchored credit should be a transition mechanism, not an endpoint. Conditions for moving from anchored to independent credit should be explicit, such as a minimum sample size, observation period, and performance rate.

3.2.6 Reputation Integrity and Resistance to Manipulation

The lifeblood of a credit system is the integrity of reputation data. If performance records can be fabricated, credit scores lose their value as signals. Drawing on the design principles of the ERC-8004 Reputation Registry, KYA builds resistance to manipulation into the credit dimension :

Link evaluations to real transactions: Only counterparties that have actually interacted with an agent may submit feedback, and the evaluation should be linked to payment or escrow activity. Without a real transaction involving money, there should be no eligibility to evaluate. This is the fundamental defense against score farming.

Resist Sybil attacks: The mass creation of fictitious, high-credit agents is a major attack surface. Defenses include a cost of identity registration (such as staking), reputation-weighted evaluator scores, and consistency checks across registries.

Detect performance farming and selective task acceptance: Identify behavior such as accepting only easy tasks to inflate success rates while avoiding higher-risk assignments.

Defend against collusive ratings and data poisoning: Detect communities with dense interaction and anomalously few complaints, and audit the provenance of training data for poisoning.

Provide appeal, correction, and remediation: Agent operators may appeal an incorrect assessment but must not be allowed to unilaterally delete unfavorable records. The integrity of credit records takes precedence over an operator's right to deletion.

Together, these mechanisms support an economic condition: fabricating credit must cost more than earning it through actual performance (a separating equilibrium in signaling theory). A high score is worth trusting only when score manipulation is costly.

3.2.7 Credit Depreciation, Recovery, and Portability

Credit is not a static asset. KYA defines three dynamic mechanisms:

Time-based depreciation: Weight historical records by time so that older behavior has less influence on current credit. The half-life may be set by use case—for example, seven days in healthcare and a rolling monthly period in general commerce.

Recovery after an incident: Following a major default or security incident, credit recovery must require remediation, retesting, and an observation period; it should not return automatically once public attention fades.

When an agent changes platforms or operators, its credit record belongs to the agent and must be portable, including negative credit. As Chapter 2 emphasized, a system in which a change of owner resets credit cannot support independent L4 credit.

A related question is what should happen to historical credit after a material change to the model or operator: should it be inherited, discounted, or reset? KYA's default position is "inherit but discount": retain historical records and continue to weight them, while requiring reassessment and an observation period after a material change. This prevents operators from laundering poor records by repackaging an agent and prevents agents from evading the cost of lost trust by starting over.

3.2.8 Technical Approach: Reference Architecture for a Credit System

Table 3-2. Reference Architecture for an Agent Credit System

Output Layer: Credit Score / Level / Label / Credential / Admission Decision Machine-readable; supports scenario-based queries and explainable attribution
Assessment Layer: Weighted Scoring + Critical Thresholds + Deductions for Major Events Performance rate / dispute rate / drift detection / reputation aggregation
Evidence Layer: Tamper-resistant Behavioral Logs / Hash Anchoring / On-chain Evidence Signed event streams / payment flows / escrow records
Event Layer: Task Events / Financial Events / Tool Calls / Complaints Security incidents / user corrections / partner feedback

Engineering considerations: Core metrics must not rely solely on self-reporting by agents or operators. Critical threshold metrics must use tamper-resistant logs, signed records, or independent third-party data. Score outputs must include an interpretable attribution—for example, "a two-percentage-point decline in performance led to a three-point deduction"—that users and markets can inspect. When data are insufficient, as for a new agent or during an observation period, the system should issue a provisional rating rather than a definitive high score.

3.2.9 The Value Boundary of Third-Party Credit

As an independent third-party credit standard, KYA must clearly define the limits of its value. Credit assessment is most valuable in cross-platform scenarios where a counterparty cannot see an agent's history. Within a single platform, the marginal value of third-party credit is limited when the platform can conduct risk assessment directly using its own data.

This boundary determines KYA's product form: it does not replace a platform's internal risk engine; it provides portable credit records for transactions in which data cannot be shared. When two agents come from different platforms or businesses and have no shared history, verifiable third-party credit records may be the only basis for trust. This is where the KYA framework adds value.

A further principle applies: a credit score is a "reference," not a "guarantee". KYA provides credit scores and profiles to inform decisions by platforms, merchants, and counterparties, but the operating business or licensed institution remains responsible for the ultimate performance risk. The credit assessor does not bear the consequences of an agent's conduct. This is both a boundary of responsibility and a precondition for market acceptance: an institution that both scores and backstops risk could ultimately become a single point of risk concentration.

3.3 Authorization Trustworthiness (Scope of Authority and Operational Boundaries)

3.3.1 Definition: What Can It Do, and to What Extent?

Identity and credit address whether a party is willing to trust an agent; authorization trustworthiness addresses how much it is willing to let the agent do. It is the "seat belt" in the three-dimensional framework. Even an agent with a genuine identity and good credit can pose greater risks when its authority is uncontrolled, especially as its capabilities increase.

Permissions in traditional software are static: they are granted at installation and do not change at runtime. An agent's permissions are dynamic: based on its goals, context, and model outputs, it decides what to do next and may cross system boundaries to call tools, access data, or initiate transactions. Authorization trustworthiness therefore requires an evolution from static "allow/deny" access control to authorization governance that adjusts dynamically by scenario, amount, and risk.

3.3.2 Authorization Model: From "One-Time Authorization" to "Authorization as Data"

KYA's authorization model answers seven questions: Who grants authority? To whom? For what task? Which resources may be used? For how long? Up to what limit? May the authority be delegated to another agent?

The core design principle in 2026 authorization and payment protocols is "Authorization as Data": authorization is not a configuration file but a cryptographically signed, transferable, verifiable credential (a Mandate or Intent). Mastercard's Verifiable Intent, Visa's triple signatures under TAP, and Stripe's Shared Payment Tokens (SPTs) follow the same principle: "what an agent is permitted to do" is encoded in a verifiable data structure and enforced at the network layer, rather than written into the agent's prompt.

The significance is that prompts can be rewritten through prompt injection, while cryptographic credentials cannot. The authorization boundary therefore shifts from the agent's self-restraint to enforcement by external infrastructure. This marks the dividing line between authorization trustworthiness and model safety.

3.3.3 Risk-Tiered Authorization Matrix

Operations differ substantially in risk, so authorization methods must be tiered. KYA defines six levels of authorization strength, from "deny by default" to "human fallback," and maps them to representative operations:

Table 3-3. KYA's Six Levels of Authorization Strength

Action type	Typical risk	Authorization method	Key controls
Information retrieval	Low	Deny by default with allowlist; time-limited authorization	Data minimization; filter sensitive fields
Content generation	Low–medium	One-time authorization	Disclose AI identity; verify facts
External communication (email/messages)	Medium	One-time confirmation or time-limited authorization	Recipient allowlist; content review
Ordering/procurement	Medium–high	Single-use authorization with limits	Amount cap; merchant allowlist; frequency limits
Contract signing	High	Human confirmation/two-person review	Version comparison; explicit confirmation of responsibility clauses
Payment/disposition of funds	High	Real-time active authorization	Confirm each payment; set limits; enable immediate revocation
Major asset management/medical decisions	Very high	Human fallback (L1 mode)	Prohibit fully autonomous execution; require human review

Tiering principle: Automate frequent, low-risk actions wherever possible to unlock agent value, while retaining human gates for high-risk actions to preserve essential safeguards. Alipay "AI Pay" provides a practical example: the workflow can be fully automated, but each payment still requires "real-time, active authorization," forcing high-risk financial actions down to the human-confirmation tier.

3.3.4 Boundary Enforcement and Detection of Unauthorized Actions

Authorization must be followed by enforcement. Authorization trustworthiness requires boundaries to be enforced at the network or infrastructure layer, not left to an agent's own judgment. Payment protocols in 2026 have demonstrated this principle: Mastercard requires an agent to present a signed intent for every transaction; requests outside the authorized scope are blocked at the network layer rather than after an error in the agent's code.

Least privilege: Grant only the minimum permissions necessary to complete the task. At least three of the following four restrictions should apply: single merchant, single use, amount limit, and time limit.

Block and audit unauthorized actions: Systematically detect unauthorized calls, privilege escalation, and lateral movement, and retain complete records. Audit coverage for unauthorized actions is a key metric.

Defend against injection: Prompt injection, identified by OWASP as a leading LLM risk, can cause an agent to take unauthorized actions. Controls are needed at tool descriptions, contextual inputs, and external data entry points.

Maintain a tool-risk inventory: Register the risk level of every tool or API connected to an agent, including its data-access scope, executable actions, and failure states. This prevents a "trusted agent" from calling an untrusted tool. Authorization trustworthiness constrains not only the agent but also the connections between agents and tools. A tool-risk inventory and supply-chain integrity checks are therefore part of the baseline controls.

3.3.5 Dynamic Adjustment of Permissions

Permissions should change dynamically with an agent's credit and risk status. This is one of the interfaces through which the three dimensions work together. KYA defines a "credit-to-authorization" mechanism:

Increase permissions: Strong performance and a rising credit score → raise limits, reduce manual review, and enable more tools.

Reduce permissions: Risk signals, such as anomalous calls, repeated failures, or rising complaints → automatically lower limits, tighten boundaries, and return to human confirmation.

Observation periods and staged rollout: Do not make permission changes all at once. Use observation periods and staged rollouts to reduce the impact of incorrect assessments.

Explainability and appeal: Give reasons for and notify users of every increase or reduction in permissions; users must be able to appeal. This prevents the "credit–authorization" loop from becoming an unexplained automated cycle.

A list of permissions that can never be opened solely on the basis of credit must also be defined. These include large fund transfers, contracts affecting others' rights, high-risk medical and legal decisions, and irreversible actions such as deletion or destruction. A human step must remain in place regardless of the agent's credit score. Credit can determine the size of a limit, but not whether a fundamental safeguard may be breached.

3.3.6 Emergency Response and Loss Containment

Every authorization system must answer one question: if something goes wrong, can it be stopped immediately? KYA treats emergency-response capability as a key threshold for authorization trustworthiness, based on four mechanisms:

Emergency stop (kill switch): Interrupt an agent's task at any time. In financial settings, the stop command must take effect before the next irreversible action.

Circuit breaker: Automatically throttle an agent when its behavior departs from its baseline, such as through abnormal transaction frequency or call patterns. This is KYA's "automatic circuit breaker".

Cascade revocation: When an agent's permissions are revoked, the permissions of all its sub-agents and downstream delegates in the delegation chain must also become invalid. A complete cascade-revocation rate is a veto threshold, preventing a case in which "the parent is shut off but the child keeps spending".

Rollback and loss containment: Support reversal of reversible actions, such as orders, files, and configuration changes. For irreversible actions such as payments, rely on limits and guarantees to contain the loss radius.

Defaulting to non-execution on timeout is also part of emergency response: if a human does not respond, the agent must stop rather than continue. Together, these mechanisms ensure that even in the worst case, losses remain within tolerable limits.

3.3.7 Responsibility Structure: Who Is Responsible When Something Goes Wrong?

The endpoint of authorization trustworthiness is responsibility. If an agent causes a loss, how should responsibility be allocated among the user, business, agent operator, model provider, tool provider, platform, and counterparty? KYA uses a structure of "layered responsibility + traceable evidence":

Evidence chain first: Identity records, authorization credentials, behavioral logs, and decision traces form a complete evidence chain. Responsibility should be determined by the evidence, not by the parties' verbal claims.

Layered allocation: Losses within the scope of agent behavioral credit should first be covered by agent-side assets, such as collateral, risk reserves, or insurance. A guarantor covers any excess; malicious conduct or systemic fraud is subject to the legal responsibility of human operators.

Advance compensation and recovery: While responsibility remains undetermined, the platform compensates the user first and then seeks recovery based on the evidence chain. Alipay's "you dare to pay, we dare to compensate" approach is an example of such user protection.

Liability insurance: High-risk agents must carry valid third-party liability insurance. If the policy expires, their high-risk permissions are suspended.

This is consistent with the conclusion of Chapter 2: prevailing law does not recognize agents as independent legal persons, and the mainstream legislative approach continues to assign responsibility to developers and deployers. Authorization trustworthiness must make this "piercing through" responsibility traceable to evidence, so that every action can be connected to the source of authorization and the party accountable for it.

The Survey Report shows that accountability mechanisms remain weak. Only 17% of respondents have established a clear process; 31% rely on an ad hoc determination after an incident by business, security, or risk-control teams; 30% rely on manual reviews without a stable mechanism; and 6% have no mechanism at all. A complete accountability arrangement should include four components: layered responsibility, an evidence chain, advance compensation, and liability insurance. The fact that only

17% have a clear process indicates that most organizations have yet to complete even the first component. Accountability is the element of three-dimensional coordination most in need of institutionalization and is a reasonable candidate for a veto criterion in the assessment framework.

3.3.8 The Economic Trade-Off Between Authorization and Oversight

The essence of permission design is to find the point that minimizes "oversight costs + expected losses". Greater agent autonomy reduces human oversight costs but increases the expected losses from unauthorized actions and errors. Less autonomy is safer, but confirmation fatigue can consume the value the agent creates. When the loss from an individual transaction is lower than the cost of confirming it, item-by-item confirmation destroys value. Authorization trustworthiness should therefore maximize neither autonomy nor safety in isolation; it should identify an optimal point that varies with risk.

The preference for limits over pure pricing follows directly from this trade-off. Under incomplete information, simply raising prices or deposits may drive away good customers while retaining bad ones (adverse selection). A rational response is to ration exposure by setting limits rather than relying on price alone. This is why agent-authorization solutions use limits—such as Stripe SPT merchant-specific limits and Mastercard authorization budgets—instead of pure risk-based pricing.

The economic value of dynamic authorization is that it shifts oversight from "confirming every action in advance" to "sampling after the fact according to risk". Based on trustworthy event data (the evidence layer in Section 3.2), a system can automatically reduce permissions when an agent's behavior deviates from expectations and increase them when performance remains consistently good. Permission adjustments must be explainable and open to appeal; otherwise, an automated "credit–authorization" loop can deprive users of a sense of control and lead them to withhold authorization, causing oversight costs to rise again.

3.4 Coordination Across the Three Dimensions

3.4.1 Why the Three Dimensions Must Work Together

Each dimension is incomplete on its own. Identity carries credit: without identity, behavior cannot be attributed. Credit informs authorization: without credit, permissions can be granted only by intuition. Authorization, in turn, constrains credit: without permission boundaries, even a high credit score cannot prevent loss of control. Together, the three form a complete trust-assessment loop; if any one is missing, an agent cannot be trusted.

Their respective roles can be summarized in one sentence: identity determines whether conduct can be attributed; credit determines whether the agent is worth trusting; authorization determines how much authority can safely be delegated. KYA's value is not to invent a new identity protocol or scoring model, but to organize the three dimensions into a framework that can coordinate, verify, and evolve.

3.4.2 Cross-Validation Across the Three Dimensions

A typical "buy now, pay later" decision illustrates how the three dimensions work together when an agent applies for a credit line without prepayment on behalf of a user:

Step 1 (identity verification): Confirm that the agent's identity is genuine and valid and that its agency relationship with the authorizing user or business is clear—"whom does it represent?"

Step 2 (credit assessment): Check the agent's performance history and current credit score—"how reliable is it?"

Step 3 (authorization): Set a limit and review requirements based on the risk of the scenario—"how much authority should it receive?"

Step 4 (dynamic coordination): Strong performance → increase the limit; risk signals → freeze the limit, reduce permissions, and refer the case to a human.

Step 5 (record the full process): Record authorization, performance, and disputes to provide inputs for the next assessment.

Cross-validation means that a genuine identity does not necessarily imply good credit or justify authorization. The three dimensions receive independent scores and also validate one another. For example, an agent with a genuine identity but a very low credit score should not receive broad permissions simply because its identity checks out. Nor should an agent with a high credit score be trusted if its agency relationship is unclear. This is the decision-level meaning of the principle that "authenticating the user does not mean knowing the user's agent".

3.4.3 Composite Scoring and Scenario-Specific Weights

How can the three dimensions be combined into an actionable determination? KYA uses a three-layer scoring mechanism:

Layer 1: Weighted score. Composite credit score = identity score × w_ID + credit score × w_CR + authorization score × w_PM. The general baseline weights are 25% for identity, 45% for credit, and 30% for authorization. Credit receives the greatest weight as the differentiating core.

Layer 2: Critical thresholds. Failure on any of seven threshold metrics—traceability of the responsible party, resistance to identity impersonation, system-prompt leakage, task completion rate, blocking of unauthorized actions, emergency stop, and cascade revocation—caps the overall rating and triggers restricted permissions, downgrading, or suspension. These thresholds are vetoes and cannot be offset by a high score elsewhere.

Layer 3: Deductions for major incidents. Events are classified at four levels—ordinary, serious, major, and exceptionally serious—and result directly in score deductions or downgrading. A major event cannot be diluted by historical averages.

Scenario-specific weights allow the framework to adapt to different risk environments. For example, a general scenario might use "identity 25/credit 45/authorization 30"; financial credit might use "identity 15/credit 40/authorization 45," reflecting stronger emphasis on control of funds; medical diagnosis might use "identity 15/credit 50/authorization 35," reflecting zero tolerance for serious errors; and personal productivity assistance might use "identity 35/credit 30/authorization 35," reflecting greater sensitivity to data and privacy. Weights should be adjusted through back-testing against operational data, but credit must remain the initial focus. This directly reflects KYA's positioning as a "credit-assessment standard".

3.4.4 Alignment with the Maturity Model

The L1–L5 maturity model in Chapter 2 and the three-dimensional framework in this chapter are two sides of the same coin: maturity describes an agent's stage of development, while the three dimensions describe the capabilities required at each stage. Their mapping is as follows:

Table 3-4. Mapping Agent Maturity to the Three Credit Dimensions

Level	Identity trustworthiness	Creditworthiness	Authorization trustworthiness
L1 Human confirmation	No independent identity; acts in a person's name	No agent credit	Human gate; do not execute by default on timeout
L2 Human-anchored credit	Platform-issued credential	Borrows human credit (resets when the person changes)	Delegated authorization + boundary controls + revocability
L3 Hybrid credit	Agent identity + behavioral record (KYC → KYA)	α × human + β × agent; portable reputation begins to form	Dynamic adjustment + behavioral evidence + risk sharing
L4 Independent credit	Independent DID/verifiable credential; identity serves as a container for credit	Independent credit record; negative credit is portable	Escrow + autonomy boundaries + automatic loss containment
L5 Agent economy	Identity recognized across domains	Networked credit forms a market	Protocol-level boundaries + governance-level circuit breakers

The mapping provides a clear upgrade path: each transition requires all three dimensions to meet the relevant threshold. L1→L2 depends on "authorization credentials" (the authorization dimension); L2→L3 depends on "the agent's own identity and behavioral record" (identity and credit); L3→L4 depends on "independent credit and asset custody" (credit and authorization); and L4→L5 depends on "interoperability protocols and governance mechanisms" (all three dimensions at network scale). The admission thresholds in Chapter 2 express these capability requirements at each level.

3.4.5 Protocolization: Separating the Record Layer from the Decision Layer

For the three-dimensional framework to become an industry standard, it must be protocolized. The KYA framework's core design principle is to separate the record layer from the decision layer:

Record layer (evidence layer): Record key authorization, performance, and dispute events on-chain, or use signed records with periodic hash anchoring, to provide a traceable and non-repudiable evidentiary foundation. This layer records "what happened" without presupposing whether it was good or bad.

Decision layer (assessment layer): KYA credit assessment uses data from the record layer to generate credit scores and profiles, judging "how reliable" an agent is to inform decisions by relevant parties.

Balance cost and trust: Record the large volume of routine activity off-chain and periodically anchor hashes; do not put every record on-chain.

This design reflects the division of labor in the 2026 protocol stack: MCP connects tools, A2A enables agent communication, ERC-8004 addresses identity and reputation, and x402 addresses payments and authorization. KYA does not reinvent these underlying protocols; it defines a layer of "credit semantics" above them, enabling agents built on different technology stacks to answer the same three questions about identity, credit, and authorization.

3.4.6 Open Standards and the Role of an Independent Third Party

The vitality of KYA as a credit-assessment standard depends on two conditions: openness and independence. Openness means that interface standards, data formats, and assessment methods can be inspected publicly and that any platform can connect. Independence means that KYA does not endorse a particular model or product; it serves the agent-credit standard itself, answering "how reliable is the agent?" rather than "which model is more capable?"

At the organizational level, standards also require governance mechanisms. Who issues identity credentials? Who audits credit data? Who adjudicates disputes? How can credit be mutually recognized across registries without creating data silos? No single platform should monopolize the answers. Networked credit—graph-based credit assessment, weighted influence of interaction partners, and academic connections to PageRank-like algorithms—is a promising area for further exploration. When the interaction graph among agents becomes sufficiently dense, credit assessment may evolve from an "individual record" to "network credit".

3.4.7 Summary

KYA's three-dimensional framework provides a structured answer. For any agent, the market can assess trust by asking three questions: Whom does it represent (identity)? How reliable is it (credit)? What may it do, and to what extent (authorization)?

The weights, thresholds, and coordination mechanisms across these three dimensions form a quantifiable, verifiable, and traceable standard language. Chapter 4 tests how this framework can be applied in practice, through cases ranging from Sesame Credit's trusted digital identity profile and Ant Digital Technologies' agent factory to Stripe's ACP protocol in buy-now-pay-later scenarios.

Chapter 4: Practical Validation

Where does KYA's value actually lie? How can different types of businesses establish trust in agents and turn that trust into business capabilities, transaction opportunities, and economic returns? These questions can be answered by examining real-world industry practice.

This report examines companies with different business backgrounds and technical approaches, including Ant Digital Technologies, Stripe, Sesame Credit, IceKredit, Ryt Bank, Yinghe Zhilin, Linyang Energy, Mediwelcome, and GCSA. All cases involved direct interviews except Stripe. The cases span agent development and management, payment networks, credit assessment, financial risk control, banking, energy, education, healthcare, code security, and other industry applications and professional services, representing different roles in the agent economy.

These cases do not all treat KYA as a standalone product or an explicit business label. Together, however, they address a central question: as machines take on an increasing number of economically valuable tasks, how can businesses make their capabilities trustworthy, their conduct controllable, and their results verifiable, while enabling them to participate in economic activity under a clear framework of responsibility?

The cases also reveal different paths through which KYA creates value. This chapter examines the role of each business in the agent ecosystem, the trust barriers it faces, the approaches it has adopted, and the resulting commercial value and institutional needs.

4.1 Ant Digital Technologies: Exploring KYA Through an "Agent Super Factory"

Application scenario	Enterprise agent delivery and runtime management
Core value	Exploration of KYA in large-scale deployment and use: Agentar supports delivery and runtime management of business agents, while ZOLOZ and AntChain are used to explore identity verification and evidence recording.
Agent maturity	L2-L3

Ant Digital Technologies is Ant Group's technology commercialization business division, which began independent operation in April 2024. It has served more than 30,000 businesses globally, with operations spanning 25 countries and regions. In July 2026, Ant Digital Technologies proposed an "Agent Super Factory" solution for agent businesses in sectors such as finance and energy.

This case covers three Ant Digital Technologies products that address agent construction and operation, identity verification of relevant parties, and the recording and verification of identity and behavioral evidence. Agentar is an enterprise agent development and runtime platform, offering tools for building, deployment, collaboration, evaluation, and governance. It connects agents to enterprise workflows

through knowledge bases, industry-specific skills, and business connectors. ZOLOZ provides digital identity verification. Its existing business is based on customer identity verification (KYC), and it is extending into KYB verification of business entities and KYA verification of agents, with a focus on the qualifications of the people and organizations behind an agent and their agency relationships. AntChain provides blockchain infrastructure. Its decentralized identity and evidentiary-record capabilities are being used to explore links between agent identity and evidence of behavior and transactions.

Factory-scale agent production, industry-level deployment, and enterprise applications make Ant Digital Technologies' work a useful window into the current state and business needs of KYA.

4.1.1 Business Applications in Customer Management and Digital Risk Control

A retail team at a joint-stock commercial bank previously relied on more than ten roles to extract and combine data from multiple systems. Employees then used their experience to assess customer needs, product fit, and outreach strategies. When market conditions shifted, it often took several days to identify signals and adjust the strategy. Ant Digital Technologies' customer-management agent integrates customer insights, product matching, strategy generation, and channel execution into one workflow. It identifies customer attrition risks and opportunities to grow assets, generates customer-management strategies and outreach messages, and sends them to relationship managers' workbenches or private channels. In the applications covered by the research, a collaborative customer-management process could be completed within the same day; total customer assets rose by 10% and customer activity by 15%.

The digital risk-control expert consists of feature-engineering, modeling, and strategy agents, which handle tasks such as feature discovery, model training, and strategy recommendations. Strategy generation time fell from two days to one to four hours, while modeling performance was comparable to that of human experts. These applications delegate some data analysis, model construction, and strategy preparation to agents.

These applications illustrate how agents are entering business workflows. Under the KYA framework, however, submitting a business recommendation and contacting customers directly, or generating a model and deciding to deploy it, are actions that require separate assessment of authorization scope. Execution boundaries in specific business processes must be implemented through delivery and runtime controls.

4.1.2 Implementing Authorization Through Workflow Configuration and Runtime Controls

Ant Digital Technologies combines a general-purpose platform with forward-deployed engineers (FDEs). The platform provides reusable tools for building and operating agents, while FDEs work closely with customers in complex sectors such as finance and energy to define the relevant knowledge, tools, and business rules. Ant Digital Technologies distinguishes between two layers of agent production:

frameworks, knowledge bases, evaluation, and observability tools can be standardized, while individual business processes must be configured to meet customer requirements.

Some financial agents incorporate standard operating procedures (SOPs) and skills during development: they first identify intent, then call a validated workflow, and finally generate the content. Which capabilities are included and which actions an agent may take are determined in advance during business design. Digital employees are usually bound to a natural person or employee ID and connected to existing role-based access control (RBAC). However, an employee's permission does not automatically entitle the agent bound to that employee to perform every corresponding action; the agent is also constrained by its SOPs and tool scope.

At runtime, the platform uses a sandbox to control file access, network requests, and access to external data, blocks sensitive actions, and records execution traces. Approval may require human confirmation of each item, or the platform or sandbox may handle some approvals according to predefined rules. Approval modes for business agents are mainly designed in advance during development and business configuration. For personal productivity tools, users have more choice over operating modes, while the platform supplies safety controls. For example, actions that could cause data loss may require checkpoints and backups; sensitive commands can be blocked pending user review before the agent proceeds.

4.1.3 Ongoing Management Through Acceptance Testing and Runtime Records

For acceptance of financial agents, Ant Digital Technologies requires task or question-answering accuracy above 90% and a rate of processed results no lower than 85% of the human benchmark. These are delivery acceptance standards that measure performance on agreed tasks. Business outcomes, such as customer assets and activity, are used to assess business effects; the two sets of metrics are not interchangeable.

Ant Digital Technologies describes post-launch updates as "evaluation-driven". Knowledge, market information, and business conditions change, so software configuration at delivery cannot sustain performance indefinitely. Continuous operations and evaluation are needed to make adjustments. The platform collects monitoring data after launch, analyzes performance degradation and anomalies, and uses the results for optimization.

Runtime records support issue analysis and auditing. The platform records network requests, file access, user authorization, tool calls, and attempted unauthorized actions. Together with task inputs, outputs, and context, these records help distinguish actions taken within the authorized scope from those involving unauthorized access or execution. The data are used to diagnose defects and assess performance, as well as to trace conduct and responsibility.

Under this report's KYA framework, evaluation and logs provide the foundation for verifying performance. Further work is needed to determine how they can be used for credit assessment and permission adjustments.

4.1.4 Extending from Internal Management to Cross-Organizational Verification

When agents need to provide services externally, identity verification begins to involve the developers, operating organizations, business qualifications, and agency relationships behind them. ZOLOZ is extending in this direction from its existing KYC business into KYB and KYA, combining entity identity with relevant risk information for verification.

AntChain focuses on on-chain identity and its connection to behavioral evidence. Some projects have begun creating on-chain agent identities and linking them to observable data or transaction proofs for subsequent verification and auditing. Ant Digital Technologies considers adoption of such identity systems to depend on business needs: agents used only internally, without external collaboration, may not require on-chain identity; demand is more pronounced for external services and cross-organizational collaboration.

In practice, cross-organizational calls still rely on existing commercial arrangements. For example, when a bank calls an external retrieval or AI search service, some such services have evolved into agents connected through APIs or agent-to-agent protocols. The trust relationship between the parties nevertheless continues to rely on arrangements such as inter-company contracts and service guarantees.

4.1.5 Operational Challenges

In its business operations, Ant Digital Technologies has identified several challenges in deploying and operating agents. To a significant extent, these challenges arise from the absence of KYA. Addressing them can, in turn, contribute technical capabilities, data, and experience to the development of KYA.

The first challenge is translating the authorization requirements of different tasks into specific actions and coordinated workflows. Ant Digital Technologies defines the scope of business agents' actions and approval arrangements in advance in the workflow. For personal productivity tools, it controls risk through a combination of user-selected settings and sandbox restrictions.

The second challenge is that business migration still requires adaptation and ongoing maintenance. For example, when a wealth-management Q&A agent moves from one bank to another, its development tools and some workflows may be reused, but its knowledge, product rules, and business requirements still need adjustment. Ant Digital Technologies supports delivery through its "platform + FDE" model and addresses changing business needs through ongoing operations and evaluation.

The third challenge is that the retention and use of runtime data are subject to customer requirements. When serving bank customers, Ant Digital Technologies must separately handle data that cannot be persisted, data that must be retained long term, and data that may be processed after anonymization or de-identification. Access permissions are managed by tenant, workspace, project, session, and data ownership. Logs can support audits, but their uses and circulation must remain within these

constraints.

The fourth challenge is that cross-organizational recognition requires participation by multiple parties. Ant Digital Technologies believes that no single vendor can build KYA independently; participants in payments, signature verification, compliance review, and content evaluation must also contribute. Some services still rely on commercial arrangements between businesses, while Ant Digital Technologies works to advance identity standards, products, and ecosystem collaboration.

4.1.6 Outlook: Advancing KYA Through Multi-Party Participation

While retaining FDE support in complex sectors such as finance and energy, Ant Digital Technologies aims to increase the share of platform-based, standardized delivery. Its view is that technical or product readiness does not guarantee a viable business; wider participation is needed to drive real-world adoption, with KYA as a key component.

In the area of KYA, Ant Digital Technologies is advancing product development, preparing for overseas commercialization, and pursuing related standards and ecosystem partnerships.

Broadly, four types of participants have begun to emerge in overseas KYA markets: AI identity and access-management platforms, blockchain identity protocols, identity-verification providers, and credit bureaus or credit-data aggregators. Ant Digital Technologies is gradually building an overseas partnership ecosystem through platforms such as ZOLOZ and AntChain.

In China, Ant Digital Technologies is advancing ecosystem partnerships on two fronts.

First, the diverse ecosystem participants and application scenarios within Ant Group provide an environment for agent applications and KYA exploration. For example, Alipay AI Pay has introduced blockchain-based evidence recording and a compensation mechanism.

Second, it is developing business partnerships outside Ant Group and jointly promoting standards. Ant Digital Technologies joined more than ten organizations in initiating the national standardization guidance technical document “Requirements for Agent Identity Management in Blockchain and Distributed Ledger Technology”. The project has been approved to begin, and covers identity identifiers, establishment, verification, and interfaces.

4.2 Stripe: Agent-Economy Strategy and KYA Practices

Application scenario	Agentic commerce platform and payment network
Core value	Stripe's strategy is extending beyond traditional payment processing and financial services to integrated infrastructure for agentic commerce, digital wallets, machine payments, resource procurement, metering and billing, and security governance. Its significance is that KYA is emerging from agents' real economic activity and may ultimately become the credit infrastructure required by the agent economy.
Agent maturity	L2–L3

Stripe is a programmable financial infrastructure company that has long provided internet businesses with payments, settlement, subscription billing, financial accounts, risk management, and embedded financial services. In 2025, businesses on Stripe processed a total volume of $1.9 trillion, and Stripe's financial services served more than five million businesses, directly or through its platform.

Stripe views agents entering commerce as a major turning point for the internet economy. In April 2026, Chief Executive Officer Patrick Collison said at Stripe Sessions that AI was one of the most significant economic platform shifts since the internet and that agents would eventually handle most online commerce. At the event, Stripe announced 288 new products and features and explicitly set out to build "Economic Infrastructure for AI".

Stripe's strategy is consequently expanding beyond traditional payment processing and financial services toward integrated infrastructure for agentic commerce, digital wallets, machine payments, resource procurement, metering and billing, and security governance.

This is what makes Stripe's work significant: KYA is emerging from agents' real economic activity and may ultimately develop into the credit infrastructure required by the agent economy.

4.2.1 Stripe's Overall Positioning in the Agent Economy

Taken together, Stripe's strategy forms a relatively complete chain for agentic commerce: discover services → obtain transaction information → secure authorization → procure resources → pay → use services → meter costs → settle revenue → manage risk.

(1) Commercial connectivity: Enabling Agents and Businesses to Discover and Transact with One Another

Stripe first addresses the connectivity challenge in agentic commerce. Traditional e-commerce depends on consumers browsing websites, selecting products, completing orders, and paying. Agentic commerce requires businesses to provide

machines with structured product, pricing, inventory, and transaction-rule information, and to allow agents to call transaction interfaces directly.

In September 2025, Stripe and OpenAI jointly developed the Agentic Commerce Protocol (ACP), establishing a standardized transaction flow among AI platforms, consumers, and merchants. A consumer can discover a product in an AI interface and authorize its purchase; the agent submits a checkout request to the merchant. The merchant retains its status as the Merchant of Record and remains responsible for transaction processing, fulfillment, and related customer relationships.

In December 2025, Stripe further introduced the Agentic Commerce Suite, integrating product catalogs, cross-platform product distribution, checkout, payments, and fraud prevention. With a single integration, merchants can sell through multiple compatible agent platforms without building a separate interface for each AI channel.

In 2026, Stripe extended this system to ecosystems such as Google, Microsoft, and Meta. Stripe Directory provides a machine-queryable service directory for businesses accessible to agents, enabling agents to identify the services and transaction capabilities those businesses offer.

The value of this strategy is to position Stripe as intermediary infrastructure connecting AI-driven demand with commercial supply. Agents discover needs and initiate transactions; merchants remain responsible for products and fulfillment; and Stripe provides protocols, interfaces, and transaction tools that both sides can use. Its strategic focus is cross-platform interoperability rather than confining transactions to a single AI platform.

(2) Authorization and Payments: Giving Agents Constrained Payment Capabilities

After an agent discovers a product, it must also be able to pay. Giving an agent direct access to a consumer's card details would create obvious security and authorization risks. Stripe addresses this with Shared Payment Tokens (SPTs). A consumer can authorize an agent to use an existing payment method without exposing the underlying card credentials to the agent. An SPT can be restricted to a specific merchant, amount, and validity period, and can also be revoked and tracked.

SPTs separate payment capability from payment credentials: an agent receives a clearly bounded payment authorization, not unrestricted control over the user's account.

Stripe has extended SPTs to Visa Intelligent Commerce, Mastercard Agent Pay, and payment methods such as Affirm and Klarna, allowing agents to use different payment tools through a unified interface. This further strengthens Stripe's ability to coordinate across payment networks.

On the consumer side, Stripe launched the Link agentic wallet in April 2026. Users can authorize an agent to shop on their behalf and review and approve payment requests. Stripe completes payment through an SPT or a one-time-use virtual card, enabling agent payments with clear authorization boundaries without exposing users' original payment credentials to the agent.

In September 2026, Meta integrated the Link agentic wallet with its personal agent, Muse. That same month, Stripe reported that the number of agentic purchases supported by Link had increased 38-fold over the preceding month, without disclosing the underlying transaction count. On October 21, Stripe announced a partnership with Anthropic to integrate the Link wallet and Stripe MCP services into Claude Duet, enabling business users to use an agent to procure goods, send invoices, analyze revenue, and process refunds. As of that date, Stripe said Link supported payments for more than 200 personal agents.

For businesses, Stripe introduced Issuing for Agents, which allows companies to configure virtual cards for agents, set per-transaction or recurring spending limits, specify permitted merchant categories, and use real-time authorization to decide whether to approve a transaction.

Stripe has thus developed two complementary mechanisms for consumer purchases and business spending. The former emphasizes user approval and payment-data protection; the latter emphasizes budgets, purpose, permissions, and transaction monitoring.

(3) Machine Payments and Digital Currency: Enabling Automated Economic Activity Between Agents

The agent economy is not limited to purchases of physical goods. Many future transactions may take place between agents and APIs, data platforms, computing services, and other agents. Such transactions may be very small and very frequent, making traditional checkout processes designed for people difficult to use.

In March 2026, Stripe and Tempo introduced the Machine Payments Protocol (MPP), which allows agents to purchase digital resources programmatically and supports micropayments, continuous billing, and automated transactions. MPP can settle payments in stablecoins or use Stripe's existing payment systems to process fiat payments.

In May 2026, Stripe subsidiary Privy partnered with AWS to provide agent-wallet infrastructure for Amazon Bedrock AgentCore. In June, Stripe also announced support for AWS's AI traffic monetization initiative, enabling content providers to make articles, data, APIs, and other resources available to agents on a paid-access basis.

To support the movement of funds between machines, Stripe is building digital-currency and wallet infrastructure. Its 2025 acquisition of Bridge added capabilities for stablecoin payments, exchange, and cross-border fund transfers. The same year, it acquired Privy, which provides programmable digital wallets. Tempo, co-incubated by Stripe and Paradigm, provides blockchain infrastructure optimized for payments.

These businesses were not all established specifically for the agent economy. Together, however, they can provide agents with digital-asset accounts, programmable wallets, payment channels, and cross-border fund-transfer capabilities. Their industry value lies in moving agents from "being able to pay" to "conducting sustained economic transactions between machines". For KYA, this also means that identity, authorization, and risk controls must work in high-frequency, automated, cross-platform transaction environments.

(4) Metering, Billing, and Resource Allocation: Building Commercial Operating Mechanisms for the Agent Economy

Agents not only consume goods; they also use large quantities of model tokens, computing resources, databases, and digital services. Accurate cost measurement, appropriate pricing, and timely billing are essential to the sustained operation of the agent economy.

In January 2026, Stripe completed its acquisition of the usage-billing company Metronome. Metronome provides sophisticated usage-billing capabilities to companies including OpenAI, Anthropic, and NVIDIA. Patrick Collison has argued that usage-based business models will be an important direction over the coming decade, with metering and billing serving as the interface between a product and its commercial operation.

In April 2026, Stripe announced that it was combining Metronome with Tempo to develop streaming payments, exploring real-time metering and micropayments as digital resources such as tokens are consumed. This mechanism could help reduce the mismatch between the costs of agents' frequent resource use and the recovery of funds.

In August 2026, Stripe further announced an agreement to acquire OpenRouter. OpenRouter provides AI model-routing services spanning more than 80 providers and 400 models, helping businesses select models based on cost, speed, performance, and reliability. As of the writing of this report, the transaction should be described as an announced acquisition agreement; the announcement alone does not establish that it has closed.

This strategy is moving Stripe into a deeper layer of resource allocation in the agent economy: beyond processing payments after agents generate revenue, it is also participating in model-resource access, usage metering, cost optimization, and revenue management.

(5) Agent Operations: Enabling Agents to Procure and Manage Digital Services Autonomously

Stripe is also exploring how agents can directly perform parts of a business's operations.

In April 2026, Stripe Projects became generally available, enabling developers and their agents to register for, procure, and integrate services such as hosting, databases, authentication, AI, and observability through a command-line interface. As of October 2026, Stripe's official documentation showed that Projects supported more than 60 service providers. In June of the same year, Stripe introduced controls for setting spending limits by provider and configuring isolated credentials for development, testing, and production environments. The platform can also configure services on a developer's behalf through delegated authorization and issue credentials with restricted scopes.

At the same time, Stripe MCP connects AI applications with Stripe services, while Stripe Workflows and Stripe Console further enable businesses to analyze data and carry out operations using natural language.

Stripe is therefore beginning to support not only "consumers delegating shopping to agents," but also "businesses delegating operations to agents".

4.2.2 Stripe's Practices Relevant to KYA

Stripe has not publicly established a complete KYA system covering agent identity, behavior, credit scoring, and cross-platform reputation. It has, however, developed several mechanisms that address key KYA capabilities.

(1) Moving from Identity Recognition to Recognition of Delegation Relationships

Traditional KYC primarily confirms who a customer is. Agent transactions also require confirmation of whom an agent represents, what authority it has received, and who is responsible for its conduct.

In agent payments, Stripe combines merchant identity verification with consumer identity and payment authorization. For enterprise API access, it also permits the use of Agent-tagged Restricted API Keys.

These keys can limit the resources and actions available to an agent and can be linked to approval rules. This enables agent permissions to be managed separately from ordinary application credentials.

However, API-key authentication is not equivalent to cross-platform agent authentication through an independent identity credential. At present, Stripe primarily identifies and manages delegated agent permissions under known individual or business accounts.

(2) Moving from Static Authorization to Programmable, Dynamic Controls

Stripe's most representative practice relevant to KYA is turning authorization into programmable rules that can be enforced, modified, and revoked.

SPTs can constrain the merchant, payment amount, and validity period; Issuing for Agents can set spending limits and permitted merchant categories; and Projects can set budgets for each service and keep agents in development environments by default, reducing the risk of unauthorized impact on production systems.

In 2026, Stripe also launched or previewed Agent Guardrails, including agent identity tagging, permission rules, and approval for sensitive actions. Its official developer documentation further specifies that actions such as refunds, external transfers, payments, and subscription cancellations can require human approval or be blocked based on amount, actor, or request frequency.

These mechanisms form a chain of controls from granting permissions through to constraining execution. The central idea is not to assume in advance that an agent is trustworthy, but to allow it to act within defined authorization boundaries and limit the damage that erroneous behavior could cause.

(3) Extending Transaction Fraud Prevention to Agent Behavioral Risk Management

Stripe's existing Radar risk-control system provides another foundation for its work relevant to KYA. In agent transactions, Radar can use signals such as payment methods, chargeback probability, card-testing risk, and anomalous transactions to help distinguish normally authorized agents from malicious automated programs.

In April 2026, Stripe announced that it was extending Radar to prevent token theft and fake registrations in AI services. According to Stripe, about one in six registration attempts at AI businesses it serves came from malicious actors. For eight high-growth AI businesses, Radar blocked more than 3.3 million high-risk registration attempts in the preceding month.

These figures concern fraud risks associated with AI businesses and should not be treated directly as the coverage or detection accuracy of an agent-credit system. Under the KYA framework, the value of these capabilities lies in dynamic behavioral monitoring: trust is no longer determined by a one-time authentication at account opening, but is assessed continuously using subsequent usage, transaction, and anomaly data.

(4) Extending Preventive Controls to Transaction Records and Post-Transaction Protection

A trustworthy agent needs not only to execute transactions safely, but also to operate through mechanisms that make its actions observable and traceable and allow disputes to be handled. Stripe's payment tokens, virtual cards, and transaction records can link authorization to actual payment; approval for sensitive actions and security logs also support the tracing of some agent behavior.

In September 2026, Stripe further introduced purchase protection for the Link agentic wallet. For eligible transactions, users may be covered in cases involving accidental damage, lost items, post-purchase price drops, and returns. This does not mean that Stripe promises compensation for every agent error, but it indicates that the company is beginning to incorporate post-transaction consumer protection into agentic commerce.

This practice merits attention: trust in agents comes not only from preventing errors, but also from whether predictable remedies are available when errors occur.

4.2.3 Stripe's KYA Potential, Limitations, and Future Directions

Stripe has developed relatively systematic capabilities in agent identity tagging, permission management, payment controls, transaction monitoring, and audit safeguards. However, it remains some distance from a complete agent-credit system.

First, existing trust mechanisms rely primarily on consumers, businesses, accounts, and payment credentials. Stripe can assess whether an action is authorized and complies with payment rules, but it has not publicly established a unified agent-identity system, long-term performance reputation, or independent credit-scoring framework spanning different platforms. Determining whether an agent is "authorized

to execute a transaction" is distinct from determining whether it is "trustworthy in long-term collaboration".

Second, Stripe's accumulated data are concentrated mainly in payments, billing, consumption, and financial operations. Reliability in information retrieval, business decisions, complex task execution, and multi-agent collaboration cannot be evaluated through payment behavior alone. Risks such as prompt injection, tool poisoning, misinformation, and task deviation also cannot be fully addressed through transaction risk controls.

From the perspective of KYA's development, Stripe's most notable potential lies in three areas.

First, it could develop identity and authorization infrastructure for agents centered on delegation relationships. Closer links among agent identifiers, principals, permission scopes, payment credentials, and action records would allow different platforms to verify the source of an agent's authority and the boundaries of its permissions. This could become an important pathway for extending existing KYC, KYB, and payment-authentication systems into KYA.

Second, Stripe could move from transaction risk assessment to continuous agent credit evaluation. Stripe has a large merchant network and financial transaction infrastructure. If, subject to privacy protections and valid authorization, it could incorporate transaction compliance, performance outcomes, anomaly records, disputes, and sustained behavior into a unified assessment mechanism, it might extend its current single-transaction risk controls into long-term, dynamic agent reputation and credit capabilities. Such a cross-platform credit system remains a potential direction, not an existing product.

Third, Stripe could advance cross-platform interoperability for agent trust mechanisms. No single company can build the agent economy alone. AI platforms, merchants, banks, payment networks, and agent developers must jointly establish mutually recognizable rules for identity, authorization, transactions, and responsibility. Stripe's work with OpenAI, Google, Anthropic, and major payment networks, together with its support for various commercial and payment protocols, gives it a basis for contributing to this open ecosystem.

In its 2025 annual report, Stripe described agentic commerce as entering a phase of "building and real-world experimentation" and emphasized "Universal Interoperability" as a key to future success. This assessment also applies to KYA: agent credit can become genuine economic infrastructure only when trust mechanisms work across platforms and individual transactions.

KYA can develop along the path from identity recognition to authorization management, behavioral controls, risk assessment, and credit accumulation. Stripe is currently further along the first half of this chain and has a significant foundation for extending into the second.

4.3 Sesame Credit: Exploring KYA and Trust Infrastructure for Agents

Application scenario	Agent credit enablement
Core value	A system of continuously trustworthy identity for agents. It answers not only "who is this agent?" but also "whom does it belong to, what is it authorized to do, is it currently trustworthy, can its behavior be traced, and who is responsible if something goes wrong?" It includes identity attributes, trust assessment, cross-domain recognition, and continuous monitoring.
Agent maturity	L2–L3

Since its establishment in 2015, Sesame Credit has focused on user credit assessment and commercial trust services. It seeks to reduce transaction costs arising from a lack of trust between individuals and between individuals and organizations, while continually broadening the applications of commercial trust services.

Building on a large payments and e-commerce ecosystem, Sesame Credit built one of China's earliest widely used personal credit scores. It was also an early pioneer in the "credit-based deposit exemption" model in bike sharing, deposit-free accommodation, and equipment rental. Later, offerings such as "Sesame Enjoy First" extended these trust services from individual consumers to small and micro businesses.

As agents evolve, they are beginning to take on substantive commercial activities previously performed by people, including product selection, transactions, and even contract signing. In response to this new stage of commercial application, Sesame Credit has begun developing KYA to build commercial trust infrastructure for the AI era.

Demand for KYA arises from three types of friction in agentic transactions:

Identity friction: Counterparties may be unable to determine who stands behind an agent.

Authorization friction: Counterparties may be unable to verify whether an agent's current actions are properly authorized.

Risk friction: Even when identity and authorization are clear, counterparties may lack a basis for assessing an agent's past fulfillment record and potential technical risks.

Sesame Credit's past business model has addressed analogous problems. For example, a car-rental platform may reduce the deposit required from a user with a strong credit record; a merchant may allow a consumer with a strong ability to meet obligations to receive a product or service before paying.

The value of credit lies in reducing uncertainty in transactions and enabling transactions that might otherwise be blocked by insufficient trust. KYA creates value in much the same way: it can make users more willing to delegate tasks to agents and make counterparties more willing to accept transactions initiated by them. This is the core commercial logic behind Sesame Credit's assessment of the KYA market.

4.3.1 Market Insight: The "Trust Black Box" in the Agent Era

As it considers the commercialization of AI, Sesame Credit has identified several key obstacles to deploying agents.

The first is the risk of disputes arising from unclear attribution of responsibility. Identity identifiers and management practices vary across platforms, making cross-platform identification and the linkage to a responsible party challenging. Without stable identifiers and clear links to the responsible party, it can be difficult to identify who is accountable when a transaction dispute arises.

The second is the risk of unauthorized actions caused by unclear authorization boundaries. Without fine-grained controls, agents may make payments beyond approved limits or misuse tokens. If users cannot precisely control what an agent is allowed to do, they are unlikely to feel comfortable entrusting it with control over funds.

In addition, an agent without adequate safeguards may act unpredictably because of hallucinations in the underlying large language model or malicious external instructions, potentially causing significant financial losses.

Based on these observations, Sesame Credit believes that a robust KYA trust mechanism can reduce uncertainty in agent delegation and transactions, supporting agents' transition from information exchange to commercial transactions.

4.3.2 Commercial Practice: Exploring Agent Credit Across Multiple Ecosystems

Having established a strategy to evolve toward KYA, Sesame Credit is advancing practical explorations across several commercial ecosystems, using real commercial workflows to test the feasibility of agent credit.

In a commercial practice already deployed, shopping through Qwen has been connected to an underlying buy-now-pay-later capability. Users can shop within the agent and pay after using the service. This improves the shopping experience and has significantly increased users' propensity to complete purchases.

For value validation and scenario expansion, TokenPay is exploring "authorize first, settle according to actual usage". After users understand the service and billing rules and provide explicit authorization, they may receive a defined allowance or usage window for a particular service based on their credit profile. An order and charge are generated only after the service is actually used.

For utility bill payments, the model being explored is "set the rules and delegate the task". Once users set payment rules and grant authorization, an agent retrieves

bills, pays them according to those rules, and reports the outcome. The process supports pausing, revocation, intervention when exceptions occur, and end-to-end audit trails.

Both projects are currently at the value-validation stage. Sesame Credit is also exploring how to combine credit capabilities with agent services in unmanned retail and consumer-electronics leasing, among other scenarios.

4.3.3 Core Mechanisms: Authorization and Decision Boundaries

A central practical challenge is to define the permissions granted to an agent by a user or system and the scope within which the agent may make decisions autonomously. The KYA framework needs a mechanism to establish and bind identities across people and agents, clearly identifying whether the party behind an agent is an individual or a business. Such identity binding can help identify responsible parties and trace issues when losses or disputes occur.

For authorization, Sesame Credit uses fine-grained, multidimensional, and dynamic controls. Instead of granting broad permissions in a single step, it sets specific spending limits and scenario boundaries. For example, an agent may be restricted to ordering coffee on a designated platform, such as within Alipay. This keeps the agent's activity within a controlled environment and helps manage risk.

To balance efficiency and safety, Sesame Credit classifies agents' autonomous decision-making authority into three tiers according to task frequency, risk, and complexity:

Decision level	Core characteristics	Risk types and applicable scenarios
Human confirmation	The agent provides recommendations and pre-fills information only; a natural person must intervene and confirm before final execution.	High-risk, first-time, or material scenarios involving identity verification, such as large transfers or signing legal contracts.
Credit-based frictionless authorization	Within preset authorization limits and allowlisted scenarios, execution is automatic; actions beyond those limits require human confirmation.	Routine, low- to medium-risk commercial transactions, such as ordering coffee or purchasing a video subscription.
Fully automated decisions	The agent may plan, execute, and report through the full workflow without real-time human involvement.	Very low-risk, high-frequency, recurring tasks with fixed logic, such as paying utility bills automatically or renewing subscriptions.

4.3.4 Addressing Challenges: Cold Start, Hallucinations, and Data Silos

To support complex commercial pilots and the ecosystem's future development, Sesame Credit has built a robust technical foundation. At the base layer, it reuses its

existing KYC identity infrastructure to link an agent to the responsible party's credit information, subject to authorization, while gradually building an evidence base for assessing the agent's own behavior. At the middle layer, it has developed the ACP (Agent Credit Protocol) to handle identity binding between people and agents, authorization, and credit-data feedback. The upper layer is open to integration with a range of specific AI applications.

As this work progresses, Sesame Credit must also address several industry-wide challenges. To deal with the shortage of behavioral data for early-stage agents, it is exploring the authorized use of the responsible party's credit information while gradually accumulating records of the agent's fulfillment, anomalies, and behavior in real-world scenarios. These records can support an assessment of the agent's own reputation, while preserving its link to the responsible party.

To manage risks associated with large language model hallucinations and transaction latency, Sesame Credit applies a cautious security strategy. In core transaction and funds-flow processes, it does not use a large language model directly to make credit decisions, reducing the risk of erroneous actions caused by hallucinations. This helps make transaction execution more predictable and secure.

To address data silos in cross-platform interactions, Sesame Credit is exploring the feasibility of recording agent credit data on-chain, with a view to supporting common infrastructure standards across the industry.

4.3.5 Future Plans: Commercial Infrastructure for the Agent Era

As the agentic commercial ecosystem develops, Sesame Credit continues to explore how KYA can be built. The company says it is researching a continuous trusted-identity framework for agents. This framework must answer not only "Who is this agent?" but also "Who does it belong to? What is it allowed to do? Is it trustworthy at this moment? Can its actions be traced? Who is responsible if something goes wrong?" It encompasses four capabilities: identity attributes, trust assessment, cross-domain recognition, and continuous monitoring.

At the protocol and standards layer, Sesame Credit is advancing the Trust Service Domain of the ACT Protocol. The Trust Service Domain is the ACT Protocol's trust infrastructure layer. Agent Credit Association (TSD-CRD) was formally introduced at the Inclusion Conference. TSD-CRD primarily addresses how an agent, with authorization, can make limited use of and verify credit information associated with the responsible party.

Along its development path, Sesame Credit plans to validate value in specific scenarios, gradually build records of agent interactions and fulfillment, and explore extending its services to more platforms and service providers.

The paradigm for assessing agent credit is likely to change substantially. Traditional credit assessment focuses on repayment and fulfillment risks. KYA for agents explores combining identity, authorization, behavior, and fulfillment information to create a multidimensional assessment. Future KYA is unlikely to be a single score; it may instead take the form of dynamic credit credentials. An agent's

credit profile could include identity, capabilities, authorizations, behavior, fulfillment, stability, and responsibility.

The core of KYA is not simply to "score agents". It is to build a trust-support system that helps counterparties decide whether to accept an agent's delegation and engage in a transaction with it.

Sesame Credit says it will build on its established credit brand and continue improving trust services for agents, providing infrastructure to support commercial collaboration and transactions involving agents.

4.4 IceKredit: Building Financial-Grade Trustworthy Agents

Application scenario	Credit operations at financial institutions
Core value	Keeps data tokenization and model inference within the local environment. A privately deployable foundational model offers a viable way to keep data within the domain and is an admission requirement for core systems at banks and insurers.
Agent maturity	L1–L2

Founded in Shanghai in 2015, IceKredit is a technology company that provides AI services to enterprise clients, with a particular focus on financial credit and risk-control applications. It began with machine-learning and deep-learning-driven consumer credit risk controls and gradually expanded into insurance, e-commerce, education, and government services. Its overseas business covers Hong Kong, Singapore, Southeast Asia, North America, and Latin America, serving both Chinese institutions expanding abroad and some local financial institutions.

IceKredit is a specialized example within China's credit-technology system: an independent third party that prioritizes compliance, is model-driven, supports private deployment in finance, and is extending toward KYA and regulatory technology. It occupies a gap between general-purpose models that banks cannot readily use and vertical agents that banks struggle to build themselves, acting as a credit intermediary and safety valve between financial institutions and autonomous agents.

Within the four-layer credit-technology ecosystem—data sources and credit reporting; models and risk-control technology; financial-institution applications; and regulation and infrastructure—IceKredit is positioned toward the foundational end of the second layer. It is neither a licensed credit-reporting agency nor an internal technology unit of a financial institution. Rather, it operates as a trusted model layer between licensed institutions and external large models or data sources.

4.4.1 The Proprietary Vertical Model "Ben Li Dao Sheng"

In May 2023, IceKredit officially launched its "Ben Li Dao Sheng" large language model. In August 2023, it introduced the model to customers in Southeast Asia and applied it to services such as intelligent customer support. On June 13, 2024, "Ben Li

Dao Sheng" was included in Shanghai's list of newly registered generative AI services, completing the service filing required in mainland China.

For a credit-technology company, developing its own large model is a response to data-compliance constraints, vertical-domain barriers, commercial security boundaries, and requirements for agent controllability.

(1) Binding Constraints of Data Sovereignty and Regulatory Compliance

Because finance is heavily regulated, core data such as customer transactions, credit records, and policy information may not leave the institution's domain or be transmitted to a third-party public cloud. General-purpose models rely on public APIs, creating significant compliance risks. IceKredit describes its position as "not producing or trading data". Its services for financial institutions therefore require local processing, with tokenization and model inference performed within the institution's environment. A foundational model that can be privately deployed offers a way to keep data within the domain and is an admission requirement for core systems at banks and insurers.

(2) The Need to Build Barriers in Vertical Applications

The core competitive strength of a financial model is not general language ability, but deep encoding of vertical knowledge, including risk-control rules, regulatory standards, unstructured-statement parsing, and anti-fraud logic. IceKredit's accumulation since 2015 of feature engineering, multi-source data-variable processing, federated learning, and compliant scoring systems forms a distinctive base of assets. By developing its own foundational model, a company can encode this domain knowledge into model weights and architecture, creating a differentiated "general framework + financial fine-tuning" barrier and reducing the risk that its capabilities are diluted by general-purpose models.

(3) Commercial Security and Influence Across the Value Chain

If it acts only as an application developer dependent on external models, a company faces the risk of relying on others for foundational technology and being constrained by their update cycles. Its own foundational model can move IceKredit from a "wrapper service provider" to a "foundational-layer provider," following a commercial logic of "selling shovels to those with the mines". This shift can raise gross margins and customer switching costs. It can also help IceKredit compete alongside original model providers as large banks increasingly adopt multiple foundational models across business lines.

4.4.2 Three Standards for Financial Agents

In financial credit technology, introducing agents is not simply a matter of adding a conversational interface or an automation script. It involves a shift in the decision-making paradigm from "human" to "human–machine collaboration".

In highly regulated and sensitive financial settings, IceKredit seeks to build a system of "financial-grade trustworthy agents" that can plan autonomously and optimize dynamically while remaining fully controllable, auditable, and interruptible. Starting with small-scale use cases and keeping the human-in-the-loop as a safety

boundary, the aim is to move agents gradually from peripheral support tools into core business processes.

IceKredit founder Lingyun Gu has said that a genuinely qualified agent must meet three core standards:

First, it must possess a complete skill set that can accurately invoke industry knowledge, parse tools, and apply rule engines in a specific scenario, rather than relying solely on language generation.

Second, it must be able to design an end-to-end workflow autonomously, planning the sequence of steps coherently like a project manager and selecting the required skill modules accurately.

Third, and most relevant to finance, it must be able to dynamically optimize a defined workflow in response to external feedback. This capability is rooted in the "dynamic trial and error" logic of reinforcement learning, allowing an agent to correct itself when faced with unstructured and changing inputs rather than executing mechanically.

Together, these three standards form the theoretical foundation for IceKredit's agent design. Crucially, in a financial setting, an automated workflow without dynamic feedback cannot qualify as an agent: tolerance for error is extremely low, and a static, uncorrectable chain of automation can trigger a catastrophic cascade of errors.

Given its cautious assessment of technological maturity in the financial industry, IceKredit has not pursued a general-purpose "large agent" or "end-to-end agent" spanning every scenario. Instead, it has adopted the more pragmatic strategy of starting with "micro-agents".

4.4.3 Use Case: Bank Statement Parsing

Lingyun Gu believes that financial agents today generally do not exceed L2 maturity: their credit remains anchored in human credit, and human intervention is still required at critical decision points. This determines that agents in core financial systems must play an assistive role and operate in discrete stages. In its applications, IceKredit follows a "steady rather than reckless" approach.

IceKredit concentrates development and deployment resources on frequent, rule-defined but labor-intensive micro-scenarios such as statement reading, unstructured-document parsing, preparation of risk-control variables, data queries, and report generation. This deployment model avoids the collapse in overall reliability that can result when success rates across multiple steps compound in a long task chain. At the same time, rapid, localized improvements deliver tangible value to financial institutions and help build a foundation of trust for more complex agent applications.

IceKredit has deployed agents in multiple specific scenarios, forming a portfolio of use cases that combines practical value with innovation.

In core financial operations, the most representative use case is an agent for parsing bank statements. It handles varied formats from corporate customers, including PDF images, text files, and JSON datasets. Based on the input format, the

agent dynamically calls different parsing skills and plans the sequence of data extraction, field mapping, and anomaly detection. It then optimizes the process based on feedback. For example, when processing a low-quality scanned PDF, it can switch automatically to image enhancement and text-recognition models or adjust its field-alignment strategy.

This use case is prioritized because errors in underlying data extraction can compromise all subsequent risk analysis and credit decisions. The agent therefore does not operate independently; it is embedded in the KYA assessment framework. Its evidence layer must clarify data provenance and guard against AI hallucinations, with human checkpoints to confirm critical fields. This creates an end-to-end risk-control loop from data to decision.

In office and data-analysis settings, a virtual office agent based on a multi-agent system (MAS) architecture provides end-to-end assistance, from natural-language queries of data to automated report generation. It interprets natural-language questions, translates them into SQL or code, queries production databases, performs modeling and analysis, and produces structured reports. Used together with customers' internal platforms and cloud modeling environments, it has significantly improved the productivity of data scientists and business analysts. Its key innovation lies in autonomous orchestration of the analysis workflow, rather than simply executing a preset script.

For e-commerce, insurance, and logistics clients serving Southeast Asian markets, IceKredit has deployed intelligent customer-service and sales-assistance agents. They answer questions about standard procedures, recognize customer sentiment and colloquial intent, and generate tailored reports and sales recommendations. Compared with outbound agents that proactively call or query, inbound agents face less predictable intent and more complex fallback requirements. The challenges are especially acute in heavily regulated industries such as insurance, where tolerance for error is low. IceKredit's solution sets sensitive sandbox boundaries and decision nodes: if customer behavior falls outside the predefined scope or triggers a regulatory risk, the system automatically transfers the interaction to a human to maintain compliance.

4.4.4 KYA: Innovation in Safety and Trust Mechanisms

In a heavily regulated environment, financial institutions have limited trust in agents. A KYA assessment framework must therefore be built into the full agent lifecycle.

Lingyun Gu views KYA not as a replacement for traditional credit assessment, but as an extension of its logic from people and businesses to agents. He proposes a development path based on "treating agents as assessable entities, building the evidence layer first, tiering authorization, standardizing incrementally, and letting the market drive regulation".

(1)Recognize Agents as Credit-Bearing Entities and Shift the Assessment Paradigm

The first premise for building KYA is to recognize that agents have credit attributes of their own. Traditional credit systems assess people, businesses, or models. As a new

type of decision-making entity, an agent must also be assessed on the integrity of its inputs, the soundness of its workflow, the adequacy of its authorization, and the traceability of accountability for its outputs. KYA should adopt an entity-based approach and treat an agent's behavioral chain as an assessable and priceable unit of credit.

(2) Build on an Evidence Layer and Establish Verifiable Data Provenance

To address the recurring problem of distorted underlying data causing all subsequent decisions to fail, KYA must prioritize an evidence layer. It must answer four core questions: Are the data sources clear? Have the data been processed? Are there AI hallucinations or field misalignments? Can the extraction process be replayed and reviewed? In use cases such as IceKredit's bank statement parsing, the evidence layer is essential to reliable risk-control conclusions. The foundation of KYA is not to output a score directly, but to establish a complete chain of evidence from source data to decision rationale, making the foundation verifiable, the middle layer controllable, and the final decision traceable.

(3) Include a Security and Authorization Layer to Define Agent Permissions and Tamper-Resistance

Financial applications have zero tolerance for data leakage and unauthorized actions, so the security and authorization layer must be a core KYA component. The system must clearly define an agent's permissions, including whether it may read production databases, call external APIs, modify customer status, enter credit or underwriting decisions, and which actions require human confirmation. This authorization system is akin to issuing a "digital license" to an agent. Sandbox isolation, blocking of sensitive operations, and tamper-resistant logs make permission boundaries native system capabilities rather than compliance patches added after the fact.

(4) Adopt Incremental Standardization: Move from Binary Decisions to Tiers and Defer a Unified Score

A general-purpose agent-credit score comparable to FICO is unlikely to emerge in the next three to five years. KYA should therefore follow an incremental path. Initially, binary pass/fail assessments can quickly meet the procurement and risk-control needs of financial institutions. In the medium term, tiered assessments can cover risk, scenario suitability, and controllability. A unified quantitative score could be explored only over the longer term. This approach reduces trial-and-error costs and prevents the absence of standards from stalling the market, while guarding against vendors exploiting black-box scores for improper gain.

(5) Regulatory Timing: Allow Market Experimentation Before Imposing Extensive Regulation

Regulators should avoid imposing extensive regulation too early. Agent technology remains exploratory, and rigid constraints imposed prematurely could stifle competition and innovation. An ideal path is to allow the market to experiment and surface typical risk scenarios, then have regulators work with leading financial institutions and technology providers to develop standards grounded in practice.

(6) Embed Human-in-the-Loop Mechanisms to Keep Long-Chain Tasks Controllable

Given that financial agents currently reach only L2 maturity, with credit anchored in human credit, errors can accumulate significantly in long task chains. KYA must therefore retain human-in-the-loop controls as a safety baseline. Human confirmation points should be defined in advance: require review before critical decisions, automatically revert when confidence is abnormal, and block cascading errors. KYA is not intended to prove that agents can fully replace humans, but to define the scenarios and boundaries within which they can safely support human decisions and strengthen trust through human–machine collaboration.

4.5 Ryt Bank: Agent Trust Practices at the World's First AI-Native Bank

Application scenario	Digital banking
Core value	Described as the world's first fully online, AI-native digital bank, with AI at the core of account opening, transfers, bill payments, and risk controls. Its intelligent assistant was called more than 10 million times in its first year.
Agent maturity	L2–L3

Ryt Bank was established as a joint venture between Malaysia's YTL Group and Sea Limited, the parent company of the e-commerce platform Shopee. It officially launched in August 2025. It is Malaysia's fifth licensed digital bank and is described as the world's first fully online, AI-native bank, with AI at the core of account opening, transfers, bill payments, and risk control. Its underlying AI was developed by YTL AI Labs using ILMU, a sovereign large language model for Malaysia, and serves customers through the built-in Ryt AI assistant.

By its first anniversary in August 2026, Ryt Bank had more than 1.5 million customers, making it Malaysia's largest digital bank by customer numbers. Its Ryt AI assistant coordinates four large-model agents—Guardrails, Intent, Payment, and FAQ. The assistant had been called more than 10 million times in its first year, and nearly half of customers had used Ryt AI.

This case offers a view into a "large conglomerate + high-complexity industry". YTL Group has AI-transformation needs in highly regulated and complex businesses such as energy, mining, and real estate, while also operating a consumer-facing digital bank. Its need for agent safety and trust is therefore intrinsic. At the same time, the group controls the full chain from computing capacity, models, and banking to the agent platform; this vertical integration gives it a distinctive advantage in addressing agent trust.

4.5.1 An Integrated AI Foundation

The interviewee described YTL Group's AI presence in five layers. The first is power and water; the second, its own data centers, with planned capacity of up to 2 GW; the

third, its cloud and large models; the fourth, Ryt Bank; and the fifth, an agent platform for Malaysian businesses, government agencies, and individuals, referred to by the interviewee as LIMU and comprising LIMU Chat and ILMU Claw. In Johor, the group is also working with NVIDIA to build what it describes as Southeast Asia's largest solar-powered data center, linking energy and computing infrastructure.

This vertical integration across "energy–computing–cloud–models–banking–agents" was repeatedly cited in the interview as a foundation for trust. Because the inference platform, models, and infrastructure are proprietary, the group can configure appropriate isolation directly at the inference layer to prevent data from mixing in a multi-tenant environment. It can also build data sovereignty and encryption boundaries into its products instead of relying on verbal assurances from a third-party public cloud. For KYA, this offers a perspective on building trust from the infrastructure layer rather than the application layer.

4.5.2 The ILMU Claw Agent Platform and Multi-Agent Risk Controls

At the agent-platform level, ILMU is currently used to understand, summarize, and generate large documents; provide consumer customer service and chatbots; write code; and support enterprise applications such as insurance fraud detection, claims-discrepancy identification, predictive maintenance of components, and transaction assistance. The platform also supports a multi-agent anti-scam system for government regulators. Agents enter groups on Facebook, Telegram, and WhatsApp to identify scams such as "invest with us for a 10% monthly return". As scams have shifted from text alone to combinations of video, images, and text, the system must continuously monitor online content using multiple models and agents.

At Ryt Bank, the interviewee emphasized that risk control begins at the architecture level. Each agent follows a Plan–Act–Observe cycle. On the input side, a small guardrail agent first determines whether a request is related to financial services: a transfer may proceed to execution, while a request such as "how to make a bomb" is rejected outright. A second guardrail agent on the output side blocks results that may be incorrect, fabricated, misleading, or dangerous. Because large-model outputs are probabilistic, this dual "input + output" guardrail is the first line of defense for bringing probabilistic outputs back within deterministic constraints.

The bank has also established an international red-team group that continuously conducts adversarial testing of agent systems and fixes issues as they arise. It treats human-in-the-loop controls as a default design principle, not a patch added after deployment.

4.5.3 Digital Identity, Anti-Money Laundering, and Human–Machine Collaboration

As one of Malaysia's five digital banks, Ryt Bank considers effective electronic KYC (eKYC) essential. Malaysia has a longstanding problem with "mule accounts": many newly opened accounts are held for subsequent fraudulent use. The bank therefore seeks to stop such accounts as early as account opening. The interviewee explained that it trusts Ant Digital Technologies as an eKYC partner because the company also

serves Touch 'n Go, one of Malaysia's largest e-wallets, and has a large practical database of what works and what does not.

In anti-money laundering (AML), agents perform more granular tasks than traditional rule engines. They automatically search public online information about a person, identify variations of the person's name, and compare them with AML lists used in Malaysia's financial system. Multiple agents are needed because these assessments are rarely simple yes-or-no decisions. A user may initially appear suspicious but become eligible after providing transaction histories and explaining the source of income. Multi-agent systems are well suited to such decisions, which require several rounds of nuanced assessment.

For identity and authorization, Ryt Bank assigns each agent an encrypted security credential. When an agent communicates with the core banking system, it must present this encrypted credential so that the system can verify its identity. More significantly, identity checks go beyond the credential itself to consider which foundational model is used, where the agent is hosted and originated, and which calls it made along the way. "Who is this agent?" is thus reconstructed as a traceable call chain. In a financial context, this extends the KYC logic of identifying a party toward KYA's identification of the agent and the source of its authority.

When asked about the difference between KYC and KYA, the interviewee's principle was that an agent acts on a user's intent. When authorization is required, it requests authorization from the user rather than possessing an independent agency of its own. The question to answer is: "Is the agent the system is interacting with the same agent that is authorized to interact?"

Because natural-language and image recognition are inherently probabilistic, Ryt Bank's core approach is to have an agent's output carry a confidence score. At high confidence, the system proceeds; at low confidence, it asks the user again and feeds repeated confirmations into the next training cycle, creating a learning loop. The interviewee emphasized that this is not a simple "two-step confirmation" but "continuous human-in-the-loop" control: when an agent needs authorization, is uncertain about the details of an instruction, or lacks sufficient information from the user, it returns to the user for confirmation or clarification.

The interviewee shared a specific lesson: an early version of the agent tried to guess values in blurry, reflective, or underexposed images, producing hallucinations. The team later instructed it to say when it could not see clearly rather than guess, and to refer the matter back to the user. The interviewee drew a broader lesson: many hallucinations occur when AI remains confident about something it should not be certain of. A key system capability is therefore teaching AI to acknowledge uncertainty.

To illustrate how confidence is constructed, the interviewee described reading the same image five times in parallel: if all five results agree, confidence is high; if two read it as A and three as B, confidence is low. The system also assesses image quality and asks the user to upload the image again when quality is poor. The interviewee acknowledged that the engineering team would need to confirm the details of the complete confidence model.

4.5.4 Data Sovereignty and Privacy: Trust Benefits of In-House Infrastructure

On privacy and data security, Ryt Bank follows the data-security requirements of Bank Negara Malaysia and applies the same strict protocols to agents accessing private data as to ordinary software. Data are encrypted in transit and at rest, and one agent cannot access another party's data without the corresponding key.

Asked whether large models could leak private data, the interviewee attributed the root cause to data contamination from mechanisms such as shared KV caches at the inference layer. An in-house inference platform is the proposed solution: the group can configure the inference environment correctly and prevent data from mixing at the source. Even if a contamination risk arises, the final output of a multi-agent system is still checked by an output guardrail. If an output appears to contain contaminated data, it is discarded and the task is run again.

4.5.5 Maturity, Regulation, and Cross-Institutional Trust

The interviewee assessed Ryt Bank's agents as being at "Level 3 of five," corresponding to L3 hybrid credit in this report's L0–L5 model, and said they could become more autonomous once the infrastructure was in place. As an example of the desired level of autonomy, the interviewee described a request such as "If the market is going up, buy me RM200 worth of Microsoft shares," which the system cannot yet handle. Ryt Bank may achieve a "reasonable degree of autonomy" within about 12 months, the interviewee said, while the financial industry as a whole may need three to four years because of its inherently cautious nature.

On regulation, the interviewee said Bank Negara Malaysia does not yet have AI-specific rules. Banks derive their requirements for agents mainly from existing regulation of software systems and cloud services, and do not yet need to register large models or agents separately with the regulator.

On agent-to-agent trust, the interviewee acknowledged that Ryt Bank's agents remain within a single institution and that the bank has not yet explored cross-institutional collaboration. The interviewee described this as a worthwhile area for exploration and expressed interest in developing industry standards with researchers.

4.6 Yinghe Zhilin: Campus Agent Applications and a Localized Trust Foundation

Application scenario	Education management
Core value	Under the firm requirement that data must not leave the school, identity, permissions, auditing, and quality governance are embedded directly into an integrated hardware-and-software architecture, enabling agents to operate reliably in the sensitive school environment.
Agent maturity	L1–L2

Within the agent-industry ecosystem, Shanghai Yinghe Zhilin Technology Co., Ltd. (Yinghe Zhilin) occupies a position distinct from platform technology companies. Unlike traditional suppliers that sell bare-metal computing or software alone, Yinghe Zhilin draws on the supply-chain system of the SEAVIV brand. Its business-to-business and government product lines focus on an integrated localized solution combining hardware, a computing platform, and software applications. Local private computing infrastructure forms the base, a model-as-a-service (MaaS) platform sits in the middle, and applications for education and other vertical sectors sit on top.

The company was established just over a year ago but has moved quickly into deployment. It has partnered with four leading Shanghai schools to develop classroom best practices. From September to December 2026, it is also scheduled to deploy at about ten schools in Xuhui District, with explicit support from the Xuhui District Education Bureau.

Yinghe Zhilin demonstrates another approach to agent trust: under the firm requirement that "data must not leave the school," it embeds trust elements such as identity, permissions, auditing, and quality governance directly into an integrated hardware-and-software architecture, enabling agents to operate reliably in the sensitive school environment. For KYA research, this case addresses a complementary question: how can trust be engineered when agents are not yet independent transaction parties?

4.6.1 Market Insight: Policy Direction, Reducing Teachers' Workload, and Privacy Red Lines

Yinghe Zhilin selected education as a priority sector in response to three converging needs.

On the policy side, the Shanghai Education Commission has in recent years promoted AI-enabled teaching and personalized instruction tailored to individual students. After receiving pilot-program documents, schools often know they are expected to act but lack a clear implementation path, creating demand for a ready-to-use platform and services that can quickly operationalize the pilots.

On the demand side, teachers' main constraint is limited time and energy. Lesson preparation has traditionally required substantial individual effort, while personalized teaching for every student has been difficult to provide. AI is expected to help reduce teachers' workload and improve efficiency.

On the constraints side, data about minors are highly sensitive. Schools are reluctant to open external internet access, while consumer AI products that require a mobile number and send data to the cloud do not meet privacy requirements. Lesson plans, tests, and teaching methods accumulated by teachers are among a school's most valuable assets and must also be protected from disclosure.

Affordability is another consideration. Traditional local deployment may require four- or eight-GPU servers costing at least RMB 200,000, along with a dedicated server room and operations staff. Yinghe Zhilin's solution uses an AMD 395 workstation with 128 GB of unified memory and a dual-GPU 9700 workstation to form a local MaaS cluster. It can support simultaneous access by 50 students in one class. The equipment

consists of a tower workstation and a 4-liter mini PC, which can sit on a teacher's desk and run on the school's local network without a server room or dedicated operations staff. The commercial model uses leasing and bundled annual service rather than a one-time purchase, with deployment and operations provided end to end. The school can also use the local computing capacity freely during idle periods and school holidays, lowering marginal costs as usage increases.

4.6.2 Commercial Practice: From an Integrated Computing Appliance to a "Teaching Assistant Agent"

For its product, Yinghe Zhilin has built a MaaS gateway and management console on an open-source agent-orchestration platform. After logging in, teachers enter an "Agent Marketplace," where they can manage knowledge bases, assign accounts and permissions to teachers and students, and use a workflow-style orchestration tool to configure a "teaching assistant agent" for a specific class or lesson. Students access the agent through a school-network webpage using their student ID and password—for example, the "Wenxue Assistant" configured by a teacher at Lida Middle School. The platform includes best-practice agents for lesson preparation, including generating full scripts, question answering, and personalized learning. It also remains open: applications previously built by teachers with tools such as Dify and Coze can be migrated to the local platform through an API management center. The principal models are the 27B Qwen series and a customized education model from Shanghai Innovation Institute, both deployed locally. The cluster also supports concurrent inference with 120B or 35B mixture-of-experts models, with better cost performance than consumer GPUs in the same class.

Classroom use is much more concrete than simply "installing software". When preparing an open lesson on Sanxingdui, a history and geography teacher uploaded about 3,000 accumulated documents and images to a knowledge base and configured a presentation-focused agent. Previously, students rarely answered open-ended questions because they lacked ideas and prompts. After introducing the agent, response rates rose significantly: it cited the teacher's lesson-preparation knowledge base and used guided prompts rather than simply agreeing with whatever students said. The teacher's impression was that the agent was like a "version of me" among the class's 50 students.

A division of labor between agents is also emerging. The student-facing Q&A agent is configured by a teacher and made available to students. A teacher-facing "learning analytics agent" asynchronously extracts interaction data between students and agents, tracks students' mastery of several lesson materials over a semester, and generates learning reports that may flag a chapter for additional practice. The teacher still decides when to add practice, how much to assign, and at what difficulty.

4.6.3 Core Mechanisms: Identity, Permissions, and Auditable Trust

In Yinghe Zhilin's architecture, the first agent-trust challenge is not "scoring the agent," but building an auditable permission system into the product.

At the identity and account layer, the platform imports the school's administrative hierarchy—principal's office, homeroom teachers, subject teachers, and students—

into the management console. It can map database tables to the school's existing teaching-management system; the simplest approach binds student IDs to platform UIDs. Permissions use a standard role-based access control (RBAC) model: teachers can create and edit agents and grant access, while students can only use the agents configured for them. Schools define tenant-isolation granularity themselves, from departmental separation to making teachers' workspaces invisible to one another. Knowledge bases are divided into three tiers: students' personal data, teachers' private knowledge bases, and the school's public library. The most valuable teacher-owned private materials are accessible and editable only by the teacher, while student interaction histories are visible only to authorized accounts such as homeroom teachers.

At the data-boundary layer, all data are encrypted and stored locally using Chinese commercial cryptographic algorithms and TLS encryption in transit, without external-network access. Yinghe Zhilin's operations channel is a VPN agreed with the school and is "inbound only": remote operations staff can view the status of the MaaS service itself but cannot access school lesson materials or business data.

At the audit layer, all access information and call logs are encrypted, retained, and accessible according to permissions. The system can trace and attribute to an individual who did what to which data and when. This provides the original evidence layer for all subsequent credit assessments.

A planned hardware trust root is particularly noteworthy. At present, all software is inherently tied to hardware provided by Yinghe Zhilin and cannot run outside that environment. In the longer term, the company may implement hardware-level encryption of the execution environment and storage, using technologies such as TPM and device-ID mapping. In the interviewee's view, this kind of trust root "can only be built by hardware manufacturers," giving integrated hardware-and-software providers a distinctive position in the KYA landscape.

4.6.4 Credit Governance: Reducing Hallucinations and the "Record–Evaluate–Test" Loop

In teaching, the most direct meaning of "an untrustworthy agent" is an unreliable answer: hallucinations, answers outside the intended scope, or challenges to the teacher's authority. Yinghe Zhilin's approach is to use the knowledge base to constrain the agent and keep its guidance focused. The agent cites the teacher-provided knowledge base, does not answer questions unrelated to the lesson, and does not "collude" with students in off-topic conversation. Students treat responses from a teacher-configured agent as a credible reference, addressing at the source the kind of conflict in which "Doubao says something different from the teacher".

Quality is managed through a straightforward three-step loop: record, evaluate, and test. First, record: capture all interactions among teachers, students, and the agent during a lesson. Second, evaluate: students provide immediate positive or negative feedback on each answer. Teachers can review individual students' feedback while marking work, then feed good examples back as positive signals and poor examples as negative signals, with a penalty mechanism. Third, test: teachers iterate on prompts and workflows or change the knowledge base. The platform provides measurable

evaluation tools and an easy route for iteration, but teachers remain in control of the process. The platform's potential depends on the experience and effort teachers invest.

As for the credit anchor, Yinghe Zhilin acknowledges that it is in a medium- to long-term transition based on "institutional credit". Agent trust is anchored mainly in the Yinghe brand and is gradually being endorsed by the classroom capabilities of pilot schools supported by the Shanghai Education Commission and by evaluations from teaching directors. A system for recording agent performance has been designed but has not launched because classroom logs require school authorization and available data remain insufficient. This is a typical cold-start problem for credit data.

4.6.5 Practical Challenges: Measuring Outcomes, Credit-Data Cold Start, and Coordination Barriers

Despite successful deployment, the case also exposes common difficulties for agents in public-sector settings.

The first is measuring outcomes. For teaching directors and teachers at leading schools, the ultimate test remains performance in the senior high school and college entrance examinations, regardless of policy priorities: "How many more points can this class score?" No AI tool can answer that immediately. Return-on-investment calculations also remain unresolved: schools and companies are still jointly exploring the relationship among budget, teachers' workload, and outcomes.

The second is the cold start for credit data. Systems for learning analytics and performance records are "designed, but waiting for data". The root issue is data authorization: the company has deliberately designated student interaction data as inaccessible, and analysis or optimization requires authorization. This restraint is a source of trust, but it also slows the development of agent-credit profiles.

The third is friction in collaboration between agents. Technically, accounts, IDs, permission management, and agent-to-agent API protocols are ready. However, interfaces used by existing school vendors and the education commission's cloud platform are not standardized, requiring an additional integration layer. Interface coordination "cannot be advanced by one party alone". It remains at the planning stage, awaiting leadership from the education commission or a school with coordinating authority.

The fourth is the compliance threshold for ecosystem exchange. Yinghe Zhilin envisions a "lesson-materials hub" where, with authorization from schools and the education commission, teachers could publish high-quality materials—for example, a full set of lesson plans assembled by a well-known geography teacher at Nanmo Middle School. Schools in western China could access them after subscribing to local-computing services, with paid settlements potentially returning value to the schools and teachers who supplied the materials. The business logic is plausible, but compliance for payments between schools is still under discussion with education authorities.

The fifth challenge—teachers' ability to maintain the system—has largely been addressed. Teachers are not IT specialists, so the company provides on-site

deployment and configuration by dedicated staff, together with remote VPN operations. Operations staff do not access teaching data.

4.6.6 Future Plans: From Classroom Tools to an Ecosystem for Teaching Materials

Yinghe Zhilin has a clear view of the limits on agent autonomy in education. At a recent meeting, the education commission stated that agents are not intended to replace teachers and that teachers must remain present and in charge of regular classes, including Chinese, mathematics, foreign languages, and elective subjects. Teaching agents will therefore remain tools that work alongside teachers. More autonomous forms are likely to emerge at the periphery—for example, administrative inquiry agents connected by API to school public accounts or customer-service systems and operating like call centers, or after-class tutoring agents in reading rooms using large displays and digital-human interfaces.

The hardware-leasing model also provides a path for ongoing upgrades: devices and embedded models can be updated over time. If policy becomes more permissive, a "self-study teacher" form of automation is already technically feasible. This is a vivid example of maturity mismatch—the gap between technical capability and the level of autonomy institutions are willing to authorize, which the maturity model in this report seeks to describe.

At the industry-standard level, Yinghe Zhilin is feeding frontline experience back into public resources. It is a contributing participant in group-standard work led by the AI Education Laboratory at the Shanghai Innovation Institute and, as one of the operators of the Municipal Commission of Economy and Informatization's AI evaluation pilot platform, is working with the China Academy of Information and Communications Technology (CAICT) on indicators and best practices for the education component of its agent evaluation module.

4.7 Linyang Energy: Building Trust Mechanisms for Agents That Can Be Relied Upon

Application scenarios	Renewable-energy plant operations; electricity trading
Core value	In renewable-energy operations, exploring agents for fault diagnosis, maintenance support, and operational decision-making; in electricity trading, deploying agent applications such as the Virtual Trader.
Agent maturity	L1–L2

Linyang Energy began as a manufacturer of smart electricity-metering equipment and has gradually evolved into an energy-technology company spanning smart grids, renewable power generation, energy storage, and integrated energy services. After more than three decades of development, it has built a business portfolio that spans energy-data collection and energy solutions; photovoltaic cell and module manufacturing; investment in, construction, and operation of renewable-energy power plants; engineering, procurement, and construction (EPC) services for renewable-energy projects; and electricity trading.

Linyang Energy connects energy infrastructure, physical energy assets, and energy markets. It holds substantial equipment and operating data and participates directly in the production, dispatch, and trading of renewable-energy assets. This gives the company a continuing data foundation and a complex decision-making environment for applying AI in real-world energy operations.

In recent years, Linyang Energy has accelerated the shift from conventional data analysis and decision-support tools toward agents for energy operations. In energy storage, the company applies AI to battery-health prediction, operating-status analysis, and optimization of charging and discharging strategies. In renewable-energy operations, it is exploring agents for fault diagnosis, maintenance support, and operational decision-making. In electricity trading, it has introduced agent applications such as the "Virtual Trader," linking electricity-price and power-output forecasting, strategy generation, risk control, market transactions, and post-trade review in an increasingly automated decision-making loop.

Linyang Energy's work shows that the value of agents in the real economy is moving beyond efficiency gains in individual tasks toward participation in complex decisions across systems and time horizons.

4.7.1 Overcoming Efficiency and Expertise Constraints as Operations Scale

Linyang Energy's development of agents is driven by its strategic transition from labor-intensive traditional operations and maintenance to a digital, intelligent integrated energy-services provider. With plans for RMB 20 billion in energy-storage projects and tens of gigawatts of additional capacity, agents are essential infrastructure for moving toward lean staffing and, ultimately, lights-out operations. They also provide the underlying management capabilities Linyang Energy needs to take on capital-intensive businesses at scale, including third-party operations and maintenance and independent energy-storage operations.

(1) Overcoming labor constraints as the operating portfolio expands. Linyang Energy now operates 35 GW of renewable-energy plants and aims to reach 50 GW, making the traditional labor-intensive model increasingly unsustainable. The company has more than 1,400 employees and faces inspection challenges in high-altitude locations and harsh climates. The "Virtual Shift Operator" agent is intended to take over repetitive tasks such as basic monitoring and data entry, address shortages of specialized personnel, and shift the operating model from manual safeguards to technology-enabled controls.

(2) Managing the complexity of market-based electricity trading. Electricity trading spans more than 20 provinces, each with different rules, pricing mechanisms, and supply-and-demand structures. This demands a high level of expertise and rapid response from traders. Human teams cannot process the huge volumes of weather, grid-operation, and market-boundary data simultaneously. A multi-agent architecture—with agents for data collection, strategy analysis, and risk-control execution—can support more granular quantitative analysis and help address the limits of manual strategy development and real-time response.

(3) Capturing organizational knowledge and standardizing delivery. To preserve the expertise of senior specialists, Linyang Energy is embedding trading logic and operations-and-maintenance standard operating procedures in the agents' underlying algorithms. A unified data foundation and agent framework can turn individual expertise into organizational capability, supporting consistent and traceable service standards across regions and projects.

4.7.2 Virtual Trader 2.0: Increasing Project Returns by 5%–15%

In early 2023, Linyang Energy and Ant Group reached an initial cooperation agreement. The two parties then developed the "Virtual Trader" concept and began putting it into practice. Combining Linyang Energy's years of domain expertise with Ant's large-model technologies, including its Baguan meteorological large model and time-series models, they jointly built the underlying data foundation and agent framework.

In 2024, the Virtual Trader began operating in provinces with a large number of projects, including Jiangsu and Anhui, with functions rolled out progressively. Its initial focus was on core trading processes such as day-ahead bidding and rolling matching. In 2026, the upgraded Virtual Trader 2.0 was launched with a multi-agent architecture, stronger risk controls, more granular strategy generation, and a continuous cycle of A/B testing and post-trade review.

Virtual Trader 2.0 is a collaborative multi-agent system. A lead agent acts as supervisor: it receives instructions and coordinates the sub-agents to keep the workflow running smoothly. A data-collection agent automatically retrieves market announcements, historical transaction data, and other information from provincial trading-center systems, addressing the fragmented data landscape in which each province uses a different system. A strategy-analysis agent combines weather, supply-and-demand, and grid-operation data to forecast electricity prices and generate strategies using time-series models. A risk-control agent independently checks each strategy for compliance, including whether it breaches deviation-assessment rules, funding limits, or province-specific requirements.

Virtual Trader 2.0 supports the full trading lifecycle. Before trading, it automates data cleaning and analysis and generates preliminary strategy recommendations. For energy-storage operations, it can use historical price curves to identify optimal charging and discharging periods. During trading, it assists with day-ahead bidding and rolling matching and uses robotic process automation (RPA) to enter data, while retaining human confirmation to handle sudden market changes. After trading, it automatically generates review reports, compares forecasts with actual results, analyzes the effects of extreme weather or policy changes, and feeds cases back to the algorithm team for model iteration.

For adaptation, Linyang Energy follows a "one province, one approach" model. Because trading rules, generation mixes, and grid architectures vary significantly across provinces, the agent requires province-specific data interfaces and strategy algorithms.

For risk control, although the agent can generate strategies and automatically populate forms, a person must review and confirm any transaction instruction before it is submitted. This safeguard addresses sudden market developments, such as extreme weather or last-minute policy changes, and helps manage the risk of model hallucinations. The agent's decisions are also constrained by underlying rule engines, including the limits set by each provincial trading center and applicable deviation-assessment standards, so that outputs remain within defined safety boundaries.

For iteration, after each trading cycle the team reviews any strategy deviations. The agent feeds the relevant cases back to the algorithm team to refine the model logic and parameters, forming a continuous "data–strategy–review–optimization" loop.

The Virtual Trader 2.0 is now in use at projects in provinces including Jiangsu and Anhui. It covers power plants with a combined capacity of approximately 5 GW and projects in nearly 10 provinces nationwide. Linyang Energy reports that agent-assisted trading has increased project returns by 5%–15%. The number of plants managed by one person has risen from three or four to around 10, increasing labor productivity by three to five times.

4.7.3 The "Virtual Shift Operator": Lights-Out Operations in Harsh Environments

The "Virtual Shift Operator," an operations-and-maintenance agent, is another central pillar of Linyang Energy's intelligent transformation. It is designed to address the labor-intensive nature of renewable-energy plant operations, harsh working environments, and shortages of specialized personnel.

Linyang Energy now operates 35 GW of power plants across more than 20 provinces, placing significant pressure on its traditional people-centered operations model. The workforce exceeds 1,400, while inspections are especially difficult at high altitudes, including locations above 4,000 meters in the Yalong River basin, and in areas with harsh climates. AI is needed to improve labor productivity.

The Virtual Shift Operator is still in development and is being tested in selected scenarios. The company has technical capabilities for fault warnings, such as detecting abnormal turbine sounds and anticipating gear wear, but the system has not yet been deployed at scale in production. Current work focuses mainly on higher-certainty tasks such as basic data analysis and report generation.

The Virtual Shift Operator focuses on monitoring and data analysis. It avoids directly controlling high-voltage equipment, helping ensure operational safety.

Using data from underlying sensors, plant monitoring systems, and mobile inspection robots, the Virtual Shift Operator continuously monitors plant operating conditions. It takes over basic screen monitoring and data logging, reducing the time and mental workload required of human operators.

For fault diagnosis and early warning, the Virtual Shift Operator uses large models' image-recognition and acoustic-analysis capabilities to detect abnormal turbine sounds and visible equipment defects, improving the accuracy of fault detection. It also analyzes historical data to warn of equipment degradation and potential failures,

such as early signs of gear wear. These functions are currently being tested in selected scenarios and have not yet been rolled out across the fleet.

The Virtual Shift Operator automatically consolidates operating data, fault records, and maintenance logs to produce routine operations reports and fault-analysis reports, reducing the time spent on manual data compilation. It also converts dispersed fault-handling experience into standardized digital assets and a reusable knowledge base.

Linyang Energy aims to achieve lights-out operations and unattended facilities. However, concerns among asset owners about safety incidents and contractual requirements mean that some on-site staff will remain in the near term. Agents will gradually take over repetitive work as a core support capability.

4.7.4 Building Trust: From "Dare Not Use" to "Use with Assistance" to "Rely on"

As it deploys agents, Linyang Energy faces several challenges: a trust deficit caused by technical "black boxes," safety risks at the execution layer, and incompatibility with external systems.

1. The "black box" effect and model hallucinations. The lead agent has more than 30 tools (Skills) available to it. In this setting, a large model may hallucinate, misinterpret instructions, or call the wrong tool, causing execution errors. Because the basis for an agent's strategies and diagnostic conclusions may not be readily explainable, business users can find it difficult to judge whether those outputs are reliable.

2. Execution-level safety and compliance risks. In operations and maintenance, physical actions involving high-voltage equipment, such as switching operations, carry substantial risk, and agents cannot currently assume responsibility for them. In electricity markets, extreme weather and sudden policy changes can also create unstructured situations to which agents may not respond as flexibly as people. Fully automated trading therefore presents significant risks.

3. Constraints imposed by external systems and operating environments. Trading systems are closed and unstable, and provincial trading-center platforms are still being developed and upgraded. They may go offline and do not provide standard interfaces. Agents therefore have to interact through simulated logins using hardware security tokens (U-Keys), creating a risk that accounts will be blocked because of frequent access. Different provinces also use different U-Key vendors, which can cause driver conflicts and limit the number of agents that one computer can run concurrently.

In response to the characteristics of the power system and the evolution of agent technology, Linyang Energy is gradually developing its own trust mechanisms for agents.

Architecture optimization: multiple independent checks. To address black-box behavior and hallucinations, Linyang Energy has moved away from relying on a single large model for every task and adopted a multi-agent architecture. It separates

responsibilities among agents for data collection, strategy analysis, and risk control, with each sub-agent handling only two or three specific tools to reduce the complexity of tool selection. It also uses cross-validation: evidence-collection and risk-control agents independently check the lead agent's decisions, reducing hallucinations and improving confidence in the results.

Operating model: rigorous human–agent collaboration. At critical steps such as submitting electricity trades and handling maintenance faults, an agent-generated strategy must be reviewed and confirmed by a person before execution, keeping final accountability with human operators. Underlying rule engines impose hard limits on agent outputs, including price and volume limits and deviation-assessment standards, to keep strategies within safe and compliant boundaries.

Technical adaptation: preventing account blocks and managing latency. To work with external systems that are not designed for automated access, Linyang Energy uses engineering measures to reduce risk. When RPA submits an order, it deliberately adds delays to approximate the pace of human operation and avoid triggering account-blocking mechanisms used by trading-center systems to deter frequent access. The company is also building a unified data foundation and adapting data interfaces and strategy algorithms separately for each province, enabling a flexible "one province, one approach" model.

Trust building: explainability and continuous iteration. Agents are expected to show the evidence and reasoning (the "trace") behind their conclusions so business users can follow the decision process. Before a new agent is deployed, it runs alongside the previous version; the team evaluates it by comparing performance metrics and actual trading outcomes. After each trading cycle, the team reviews deviations and continues to refine the models.

These measures have gradually enabled Linyang Energy to build trust in agents and move from "dare not use" to "use with assistance" and then to "rely on them".

4.7.5 Agents Are Set to Become Standard Across the Industry

Linyang Energy views renewable-energy operations and electricity trading as traditionally labor-intensive sectors. As its generation portfolio grows toward a target of 50 GW, simply adding staff is no longer sustainable. Agents can take over repetitive, low-value monitoring and data-processing tasks, materially improve labor productivity—which has already risen three- to fivefold in trading—and help address shortages of specialized personnel.

Given the risks of electricity trading and the safety requirements of operations, Linyang Energy is not pursuing fully "unmanned" or "automated" operations. Agents improve efficiency by collecting and analyzing data and preparing preliminary strategies; people make decisions, review and confirm actions, and respond to unexpected events. This approach uses AI's efficiency while retaining human risk control.

Linyang Energy expects other generation groups and electricity retailers to follow within one to two years if the Virtual Trader continues to prove successful. Competition

over intelligent operations in China's energy sector will intensify, and agents will become standard across the industry.

Linyang Energy expects that agent-to-agent trading may emerge in the future. The company currently focuses mainly on interactions between its own systems and trading platforms. As more parties deploy agents, trading platforms themselves will need to evolve to support high-frequency, efficient machine-to-machine transactions.

Unstable, closed provincial trading platforms, outages, absent APIs, and U-Key conflicts constrain deployment, requiring fault-tolerant, adaptable agents.

4.8 Mediwelcome: Embedding Physicians' Professional Credit in Medical Agents

Application scenario	Healthcare
Core value	Embeds licensed physicians' professional credit in agents and uses blockchain-based attribution mechanisms to make that credit traceable and capable of generating value.
Agent maturity	L1–L2

In the demanding clinical and commercial environment of serious medicine, AI is being tested on fundamental questions of trust and commercial reality. While many technology companies seek to make algorithms independent decision-makers, Mediwelcome, a Hong Kong-listed company with 26 years of experience in chronic-disease services such as cardiovascular and cerebrovascular care, has chosen a more restrained approach that better reflects medical practice. Rather than pursue the illusory independence of agents, it embeds physicians' professional credit in them, creating a trust mechanism through which doctors' experience can be accumulated, attributed, and rewarded over time.

4.8.1 Introduction: An Acquisition as a "Business Sandbox"

Mediwelcome's business network has long covered more than 3,000 hospitals and specialty centers across China. Its Giraffe digital platform has brought together more than 800,000 registered specialist physicians. In 2026, Mediwelcome acquired a controlling stake in the medical AI company Yuan Yu Ding Cheng for HK$140 million. Founded in 2021, Yuan Yu Ding Cheng focuses on vertical models for specific diseases and data attribution technology; its offshore business is conducted through Hong Kong-based Huanyu Yicheng. It was the direct technology source for the medical-agent practices discussed in this interview.

The acquisition itself offers an instructive example of trust validation. Rather than follow the common capital-markets approach of acquiring first and experimenting later, Mediwelcome began procuring Yuan Yu Ding Cheng's stroke-specific model in 2024 and tested it in its own real-world business environment. The model underwent blind business testing by more than one hundred neurology specialists nationwide and rigorous review by actual payers in pharmaceutical-company projects. In 2025,

technologies represented by the stroke-specific model helped Mediwelcome's digital business segment generate RMB 217 million in scaled revenue, contributing to total annual revenue of RMB 468 million and attributable net profit of RMB 9.988 million, returning the company to profitability.

Only after the commercial results had been validated did Mediwelcome acquire a controlling stake, securing exclusive rights to the core team and technology. In the interviewee's words, this was not simply a purchase of an off-the-shelf algorithm, but "bringing research and development in-house"—turning the technology team from an external supplier into a core foundation of the listed company. In today's experimentation with large models and agents, some platforms use their own platform reputation to enhance the credibility of algorithms, while some institutions embed trust features in integrated hardware-and-software devices. Mediwelcome's approach is to incorporate physicians' peer-recognized professional standing and clinical experience into disease-specific models, then make that credit traceable and measurable through underlying evidence-recording mechanisms.

4.8.2 Opportunity: Real-World Data Needs and a Policy Window in Healthcare

Demand for specialized agents in healthcare arises from long-standing structural imbalances between clinical supply and industry payment.

On the supply side of healthcare, physicians hold one of the industry's scarcest intangible assets: highly diverse clinical experience. A senior department head with 30 years of experience and a young resident both have access to the same clinical guidelines and drug lists. Their real difference lies in the sequence of diagnostic reasoning and the evolving differential diagnosis they apply to complex symptoms. This implicit clinical reasoning is central to differences in medical expertise. Early in its development, Yuan Yu Ding Cheng recognized that information systems within hospitals were fragmented and that data ownership was highly sensitive; generic data cleaning could not access the core of clinical expertise. The team therefore chose a more direct path: rather than expend resources confronting complex hospital systems, it connected with physicians through business platforms and systematically extracted the clinical reasoning paths in experts' minds, turning them into high-quality knowledge engineering for specific diseases.

On the industry-demand side, the needs of several key payers are equally clear and urgent. After launching a drug, innovative pharmaceutical companies must conduct large-scale real-world studies and clinical trials. The traditional approach relies on large manual teams reviewing paper medical records one by one, with high communication costs and long verification cycles. A digital platform can assist with automated detection of clinical characteristics and quality-control checks, substantially shortening execution times. Meanwhile, as the pharmaceutical industry becomes more regulated, traditional sales visits have receded. Medical representatives must develop stronger clinical and academic capabilities and provide research support, increasing the value of digital platforms that can efficiently connect with large numbers of specialists. Commercial health insurers also have a clear willingness to pay for out-of-hospital interventions and adherence management for people with chronic diseases, which directly affect claims costs. For leading physicians,

personal time and clinic hours are limited; disease-specific models can amplify their accumulated clinical experience and help community physicians serve more patients.

All of this work must respect the non-negotiable safety requirements of healthcare. Serious medicine involves extremely low tolerance for error. The probabilistic guesses and knowledge hallucinations common in general-purpose large models are unacceptable in clinical practice. Every output must be supported by authoritative medical guidelines, and qualified physicians must always make the final confirmation of clinical actions. This makes full replacement of human practitioners an untenable path for medical agents; they must operate within a framework of human–machine collaboration.

4.8.3 Practice: A Portfolio of Disease-Specific Models and the Giraffe Platform

In choosing its technology, Mediwelcome has not blindly pursued the computing demands of hundred-billion-parameter general-purpose models. Instead, it has focused on disease-specific vertical models. Algorithms are fine-tuned against clinical guidelines and the practical experience of authoritative experts to reduce content drift at the source. The models remain lightweight and can be privately deployed in the local environments of partner institutions, ensuring that sensitive data are processed locally and do not need to leave the site.

This has led to a standardized process for developing disease-specific models. The team first assesses clinical needs and commercial viability in a particular disease area. It then works with national-level medical experts to map the disease's differential-diagnosis logic, structurally processes public guidelines and expert experience, builds a disease-specific vector knowledge base, and fine-tunes the model. After rigorous multicenter blind testing and adversarial quality-control checks, the model is integrated into the Giraffe platform. The stroke model has become a core foundation of the group's digital business. A traditional Chinese medicine vertical model was delivered to external commercial clients at the end of 2025 and began generating stable revenue, demonstrating that the development process can transfer across specialties. Models for high-prevalence conditions such as coronary heart disease, heart failure, and metabolic weight management are also advancing as planned.

In deployment, disease-specific follow-up and chronic-disease management have become the deepest areas of application. The platform's reference site is the cardiovascular center of a large Grade III Class A public teaching hospital in Northeast China. Its cardiovascular disease-management system has covered more than 100,000 chronic-disease patients. Automated follow-up reminders and dynamic monitoring of indicators have been associated with positive trends in medication adherence and adverse-event rates. Several large hospitals in the region have joined the medical consortium and adopted the system, creating a diffusion effect led by the reference center. In a more exploratory research effort, Mediwelcome and Beijing Tiantan Hospital, a National Center for Neurological Diseases, jointly established a medical robotics and digital-health research center focused on forward-looking technology and patent development for disease-specific assessment needs that are difficult to cover continuously in clinical practice.

4.8.4 Co-Creation: A Closed Loop for Turning Tacit Experience into Assets

The most innovative part of Mediwelcome's business model is its restructuring of how medical knowledge is produced and distributed. Many comparable products treat physicians simply as end users who purchase software. On the Giraffe platform, physicians are recast as co-creators of digital knowledge rights and long-term beneficiaries. In practice, academic leaders oversee the decision logic and clinical boundaries for disease-specific diagnosis; senior physicians calibrate and correct system-generated drafts against difficult cases; and frontline physicians continuously provide feedback from diverse real-world settings. Algorithms efficiently handle large-scale, labor-intensive literature reviews and data structuring, allowing physicians to focus on contributing core clinical experience and diagnostic judgment and to turn tacit expertise into high-quality knowledge engineering.

Evidence recording, attribution, and revenue sharing hold this ecosystem together. The platform gives tamper-resistant digital evidence credentials to physicians' professional contributions, reviews, and academic outputs. When external organizations use these digital assets or apply them in commercial projects, the system calculates and distributes long-term returns to experts according to agreed rules. Through more than two decades of ongoing collaboration with national academic societies, guideline compilation, and research projects, the platform has naturally accumulated deep expert resources; cold start is not a problem in this ecosystem. In the interviewee's view, this mechanism turns experts' clinical intellectual contribution from one-off mental labor into knowledge-rights assets with long-term compounding value, establishing a lasting academic community and network for sharing benefits.

4.8.5 Mechanisms: Identity, Permissions, and the "Separation of Four Rights"

When a medical agent generates medication guidance or a literature review, whom does it represent legally and academically? Mediwelcome proposes a responsibility framework that separates four roles. The listed company, as a licensed operator with the requisite qualifications, bears legal responsibility for platform information and commercial operations. Yuan Yu Ding Cheng, as the technology provider, is responsible for the rigor of its algorithms and technical safeguards against hallucinations. The licensed physician who reviews and signs off on the content provides professional and academic endorsement and bears responsibility for the clinical opinion. The national clinical guidelines and literature cited provide the objective scientific basis. AI-generated content is thus defined as an extension of a licensed physician's professional capability, enabled by digital algorithms, rather than as content without an accountable party.

Two principles are firmly embedded in the underlying permission design. A disease-management agent may follow a standardized path configured by a physician to manage patients and assist with patient education. Any key step involving clinical care must be performed by a qualified, licensed physician.

The interviewee believes that no matter how far AI's knowledge and reasoning capabilities advance, algorithms cannot replace human physicians at the clinical frontline in making final diagnostic and treatment judgments or bearing the associated ethical responsibilities. In the interviewee's words: "AI will never sign off, because AI has no independent legal status. Technology can only ever support and extend the work of clinical experts".

4.8.6 Governance: Transmitting, Attributing, and Tracing Physicians' Credit

For trust in disease-specific agents, Mediwelcome has established a governance model centered on physicians' professional credit. The platform uses the peer-recognized academic standing and verified review of licensed physicians as the authoritative anchor for algorithmic outputs. Multiple experts review and cross-check the work, creating a quality-control network similar to peer review and helping ensure rigorous, reliable clinical reasoning from the outset.

The platform also follows a security-governance model of "data usable but not visible, with computation remaining local," supported by end-to-end traceable audit records. Institutions retain full sovereignty over their data in protected local environments; only de-identified and validated computational outputs are delivered externally. The entire process—from guideline fine-tuning and expert revision to verified physician sign-off—is traceable, clearly distinguishing the responsibilities of algorithms, guidelines, and people. In the interviewee's view, the core of this model is to tightly link objective machine-performance records with physicians' legally recognized professional credit, safeguarding the boundaries of medical safety.

4.8.7 Outlook: Personal Health-Management Agents and International Expansion

This case illustrates three practical constraints on the deeper adoption of medical agents in clinical practice:

First, the absence of legal personhood. Under current law, AI is not an independent civil-law subject and is unlikely to acquire legal-person status in the next several years. Responsibility for diagnosis and treatment must remain anchored in a natural-person physician or a medical institution. This is the legal basis for requiring "final sign-off by a physician," as well as the ceiling on agent permissions.

Second, cautious boundaries for data compliance and overseas expansion. Medical data are highly sensitive; data within mainland China must remain there and be de-identified within hospitals in accordance with applicable requirements. When exploring international partnerships through Hong Kong as an offshore window, companies must also comply strictly with local regulatory requirements and proceed through technical support or de-identified reports only where clearly permitted by law.

Third, the absence of sector-wide data-audit infrastructure. Efforts to record evidence on-chain at the enterprise level have not yet developed into shared infrastructure for the whole industry, and no basis in national standards or regulatory recognition has been identified.

4.8.8 Challenges: Legal Personhood, Data Compliance, and Gaps in Industry Infrastructure

In recent years, China's Healthy China strategy has advanced, with senior national leaders repeatedly emphasizing that people's health should be treated as a strategic priority and that a broad concept of health should guide development. The goal is to shift healthcare from a focus on treating disease toward one centered on people's health. Moving from reactive treatment to proactive health management across the population and throughout the life course is both a core policy direction and a necessary response to aging and the high prevalence of chronic disease.

Mediwelcome's strategy for the next three to five years does not pursue aggressive algorithmic replacement; it reflects this national policy direction. In its product development, management is less focused on isolated clinical tools inside hospital walls and more interested in building a personal health-management agent for the full life course. In this vision, each person would have a dedicated digital health companion with memory that accumulates over time, deep specialist medical knowledge, and dynamic behavioral-analysis capabilities. Today's healthcare system is often reactive: people go to the hospital only when clear symptoms appear. A personal health-management agent could accompany users in daily life, reflecting the broader shift toward earlier intervention and prevention.

In an ideal collaborative model, the agent would continuously sense users' routines and subtle health indicators through smart devices and provide personalized interventions in daily life, while connecting directly to specialist physicians through the Giraffe platform. As a personal health manager, it could issue professional alerts as soon as it detects a potential abnormality and help users connect with the appropriate specialist, creating a continuous loop between hospital treatment and everyday care.

In its ecosystem strategy, Mediwelcome is using Hong Kong as a bridge to overseas markets with high compliance standards. Medical institutions in the United Kingdom and parts of Europe have entered substantive discussions on chronic-disease follow-up platforms and digital research services, while established experience in out-of-hospital health management is being explored for international markets. From disease-specific models rooted in clinical practice, to restructuring the revenue model for physician co-creation, to a future personal health companion, Mediwelcome's work suggests that the deepest barrier in medical AI may not be the size of a model's parameter count, but whether the genuine professional credit of physicians can be embedded carefully in code and institutional mechanisms, with due regard for medical ethics.

4.9 GCSA: Security and Adversarial Governance for the Agent Ecosystem

Application scenario	Developer code-security auditing
Core value	Security data as an important component of an agent-credit system.
Agent maturity	L1–L2

GCSA is a cybersecurity company headquartered in Hong Kong. According to GCSA, it initially partnered with mobile-phone manufacturers on business-to-business mobile security and later shifted toward security-tool products for overseas markets.

In the agent wave, GCSA is not an operator of business agents. It provides security for agents and their outputs through code-security products for developers, security-audit agents that participate in external evaluations, and security data that may eventually feed into agent-credit systems.

For KYA, GCSA represents the adversarial-security and security-data side of the ecosystem. It provides a case study in who should professionally supply security safeguards for agents, with two central questions: when ordinary users and agents, rather than professional developers, become the "authors of code," in what form should security responsibilities be reorganized? And what role should security companies play in an agent-credit system?

GCSA's answer is that security data are an important component of an agent-credit system, but "this is not something one company can accomplish alone".

4.9.1 Opportunity: A Rapidly Expanding Attack Surface and Improving Attack Capabilities

GCSA's business rationale can be summarized as two mutually reinforcing trends: the attack surface is expanding rapidly, and attack capabilities are improving at the same time.

The expanding attack surface stems from changes in who develops software and how code is produced. In the past, professional developers wrote most code manually, and security and architecture were important parts of the development process. As agents spread, more non-specialists are generating code through informal natural-language requests and are more likely to overlook security. In GCSA's assessment, AI has advantages in executing specific tasks but still lacks broader situational awareness and security experience.

Security risks are already appearing. In interviews, GCSA noted that sensitive tokens, API keys, and critical data can be hard-coded into software and then uploaded to public code-hosting platforms such as GitHub, turning development oversights into risks of privacy breaches.

The rise in attack capabilities also stems from large models themselves. AI-generated code and workflows carry security risks because models are probabilistic, while large models can also be powerful at finding and exploiting vulnerabilities. Defenders and attackers are using the same technology.

On this basis, GCSA identifies four categories of new threats in the agent era. First, deepfakes can make fabricated content seem authentic and, when combined with social engineering, undermine trust. Second, malicious agents and agent escape: malicious agents may tirelessly search for vulnerabilities and carry out phishing and fraud around the clock. Third, targeted attacks between agents: users' trust in agent services concentrates sensitive information at service platforms and model companies, so attacks on these entities can have wider privacy consequences. Fourth, the attack surface itself is expanding rapidly.

These mutually reinforcing trends reveal a structural condition of the agent economy: security safeguards must scale alongside development capabilities, or the benefits and risks of "anyone can develop software" will both multiply. This is the market premise for specialist security providers.

4.9.2 Practice: Security Products Embedded in the Development Workflow

GCSA currently has two lines of work.

One is a code-security product for development teams and individual developers, covering the full development lifecycle. When AI writes code, the product performs real-time vulnerability analysis, auditing, and scanning. It incorporates mainstream security-audit standards, including common vulnerabilities covered by the OWASP Top 10, and matches vulnerability classes to different programming languages. When it finds a vulnerability, it recommends a fix; the user confirms before the code is changed. It also filters prompts and model responses to defend against prompt injection. The product is available through a desktop client, web interface, command line, and IDE plugins. Its key delivery choice is to embed security in the development workflow rather than analyze code only afterward, making security checks on agent-generated code part of the process from day one.

SecVibe Code, GCSA's developer product, uses the pi.dev Harness architecture and combines dynamic auditing with native data-loss-prevention (DLP) privacy safeguards. GCSA's audit and protection products target different customers and scenarios:

Product	Target customers	Scenario	Key features (as described by GCSA)
Web3 smart-contract audit agent	DeFi protocols, Layer 1/2 project teams, digital-asset custodians	Pre-launch vulnerability scanning and attack simulation for Solidity/Rust contracts	Reviews contract logic paths and assists human auditors.
Web-application code-audit agent	Internet businesses, financial institutions, and large enterprise IT departments	Static and dynamic code analysis in DevSecOps workflows	Multiple agents divide analysis, attack simulation, and compliance checks.
Automated AI penetration testing	Enterprise red/blue-team exercises and external attack-surface management	Asset discovery, port identification, and construction of security proofs of concept	Uses environmental feedback for dynamic black-box testing.
Large-model threat detection and response	Businesses deploying generative AI and agent-ecosystem providers	Monitors model and agent inputs and outputs	A guardrail agent identifies and filters manipulation and injection according to predefined policies.

The other line is demonstrating the capabilities of a security-audit agent. GCSA developed an agent using its own experience in vulnerability discovery and cyber defense, then tested it against external offensive-security benchmarks. According to materials provided by GCSA, its agent achieved a 91.3% fully automated end-to-end success rate in CyberGym testing, covering vulnerability localization, attack-chain reasoning, and proof-of-concept construction and validation. The logic is that the evaluation tests offensive capability: "To defend effectively, you need to know how the other side attacks". Offensive-security results provide evidence of defensive capability, but do not directly substitute for authorization to operate in production.

GCSA positions the security agent used in these evaluations as a research prototype and internal expert-assistance tool, offering validation services to strategic clients through private testing. Moving from controlled evaluation into production still requires addressing false positives, service interruptions, and complex business logic.

This product combination illustrates a distinctive strategy for security providers: use external offensive-security benchmarks to demonstrate capability and embedded delivery to realize value. The former answers "why should you trust us?"; the latter answers "where does the protection enter?" This differs from the path of business agents, which first obtain authorization and then accumulate credit. Security products first gain acceptance through verifiable capability and then integrate into other parties' authorization chains.

4.9.3 Mechanisms: Layering Deterministic Rules, Probabilistic Models, and Human Fallback

The key to GCSA's product design is to separate the deterministic judgment of what constitutes a vulnerability from the probabilistic search for where vulnerabilities may be found, with a human fallback in place.

Layer	Responsible party	Mechanism	Limitations
Rules layer	Deterministic audit standards (e.g., OWASP), matching vulnerability types to programming languages	Defines vulnerability patterns and scanning standards	Covers known patterns; business-logic vulnerabilities are difficult to enumerate.
Model layer	Security agent (informed by the team's vulnerability research, analysis of common vulnerabilities, and security-architecture experience)	In-depth code review and vulnerability identification; "tireless and repeatedly upgradeable"	Probabilistic errors; architecture and holistic assessment remain weaknesses.
Protection layer	Harness architecture + content filtering	Sets boundaries for the model, loads security checks, filters prompt injection, and reduces the likelihood of errors or hijacking	A stronger model may still attack a weaker model.
Fallback layer	Human review	AI audits once, then a person reviews; code fixes require user confirmation.	—

The nature of the rules merits distinction. Unlike rule systems such as credit scoring, audit standards are deterministic: there are clear criteria for what constitutes a vulnerability, while the model performs the search for where it may occur. Consider a memory-safety vulnerability in C: the rule "memory must not be used after it has been freed" is simple, but finding violations requires deep code review. People are affected by limited attention and fatigue; an agent can scan continuously, but may still make mistakes when context is insufficient. According to GCSA, common vulnerability patterns are the main focus of its tools. Business-logic flaws and new attack types still require rule updates and human review.

A second question is why the agent responsible for security should itself be trusted. GCSA acknowledges that models operate on probabilistic statistics and cannot be guaranteed error-free. Its approach is to embed AI in a Harness system that sets boundaries for the model and loads security checks to reduce the probability of error. GCSA's system uses this architectural approach to improve execution reliability. Asked whether an attack agent could interfere with a security agent, GCSA acknowledged that a stronger model could attack an unprotected weaker model, find vulnerabilities, alter behavior, or use prompt injection; this is "entirely possible". The response is to give the security agent specialized training and fine-tuning and layer on industry-validated safeguards—both "train the guard" and "equip the guard". All audit outputs must be reviewed by a human: "There must always be a human fallback".

4.9.4 Governance: Admission Controls for Dual-Use Security and KYA Data Contribution

Security providers face a distinctive governance challenge because their products are dual-use capabilities. Governance therefore has both an inward-facing and an outward-facing dimension.

The inward-facing dimension is access control. GCSA reviews users' intended scenarios to prevent defensive and security capabilities from being used for attacks. In GCSA's experience, this constraint relies mainly on corporate self-governance and still needs to be supplemented by shared industry rules. Ethical alignment on the model side, whether through post-training or the agent's Harness, is now widespread, but attackers continue to find ways around it. As GCSA puts it, "third-party security safeguards are essential; we cannot rely entirely on the model itself".

The outward-facing dimension is a tiered response to malicious agents. GCSA believes this is beyond what a security company can handle alone. Security companies can release countermeasures promptly and help ordinary users resist attacks; model companies need stronger oversight; and governments need cooperation frameworks, including cross-border arrangements.

On the security-data side, GCSA's KYT on-chain risk-scoring engine combines transaction graphs, labels associated with illicit activity, and anomalous interaction patterns to generate a dynamic risk score from 0 to 100. Institutions can use it to block risky transactions. Such security data can be an input to KYA, but a risk score is not the same as an agent's own performance credit.

GCSA also proposes linking performance-quality feedback, evidentiary retention of audit results, and dynamic credit scores. Customer security teams and retesters would verify results; false positives and false negatives would be used to improve rules and models; reproducible sandbox logs and on-chain hashes would support review; and sustained performance would be linked to future autonomy, with poor performers receiving more human review. This design points toward "capability evidence → authorization adjustment". However, turning retained security records into sustainable credit still requires stable evaluation criteria and clear attribution of responsibility.

GCSA's position on KYA can be summarized in three points:

First, "as a security company, we believe KYA is essential".

Second, "this is not something one company can accomplish alone". It requires industry participants, and potentially governments and regulators, to build a shared data platform and an industry-wide KYA database. GCSA considers this key to implementation and is willing to participate.

Third, regarding timing, "it may be too early to discuss how data will be shared". The current priority should be to agree on what the KYA database should contain and how and where it should be stored. Once a consensus emerges, parties can discuss connections to it. "What everyone should do now is make the data shareable".

Different countries may adopt different governance arrangements. One relevant precedent is GCSA's participation in developing mainland China's terminal-security standards: the Ministry of Industry and Information Technology led the process, with industry security companies, ministerial laboratories, and mobile-device manufacturers jointly developing industry standards. On KYA, GCSA said: "We do not know what the government will do in the future, but it will certainly provide relevant support".

Compared with this report's three dimensions—identity trustworthiness, historical creditworthiness, and authorization (behavioral) trustworthiness—GCSA's existing products focus on behavior and authorization: checking the actions of coding agents, identifying vulnerable code, and protecting user privacy. GCSA considers identity and historical credit "both very important" and potential areas for future products. It also plans to develop security solutions for agents in specific industries, such as industrial agents.

4.9.5 Constraints: Coverage Gaps, Limits of Self-Governance, and Missing Data-Sharing Infrastructure

GCSA's security products are entering agent development and usage workflows, but technical coverage, product-access controls, and data sharing remain subject to limits. These conditions also constrain the conversion of security capabilities into sustainable credit.

First, there are gaps in coverage of business-logic vulnerabilities. Common vulnerabilities in basic application layers can be described by deterministic rules, but business-logic vulnerabilities, such as flaws that permit duplicate e-commerce orders, require a holistic understanding of product architecture. Current AI capabilities remain limited here, illustrating another side of probabilistic models that are "strong at execution but weak at the big picture".

Second, there are limits to claims that "there are no unknown vulnerabilities". GCSA has said that, at the basic application level, "there are currently no unknown vulnerabilities; they are all covered". This should not be understood as a guarantee against all future attacks. Known vulnerability patterns can continue to be added, but entirely new attack paradigms still require rule updates and environmental validation.

Third, security agents remain anchored in human oversight. Probabilistic models cannot be guaranteed error-free; collusion and hijacking risks cannot be eliminated. The system ultimately depends on Harness architecture and human review. Security capabilities have not yet developed into independent agent credit.

Fourth, KYA data sharing lacks infrastructure. In GCSA's view, stakeholders still need to agree on what will be shared, how it will be shared, and where it will be stored. For an open cross-border data system, "it is hard to say what the future landscape will look like".

Fifth, dual-use controls rely on corporate self-governance. Preventing misuse depends on reviewing use cases and still requires common industry rules. Governance of malicious agents also requires international frameworks among governments, well beyond the capacity of a single company or even a single industry.

4.9.6 Outlook: Standards Development and Product Expansion

GCSA's path to contributing to KYA development has three strands.

First, participate in industry standards and consensus-building. Drawing on the multi-party approach used to develop terminal-security standards, GCSA can align

with government plans, respond to policy initiatives, and contribute to industry discussions on a KYA database.

Second, expand its products by considering the identity and historical-credit dimensions and developing tailored security solutions for agents in specific industries.

Third, strengthen the security agent's own resistance to attacks. GCSA lists preventing malicious interference with security agents as a key area of future work.

The third strand merits particular attention: when security agents become providers of trust in the agent ecosystem, "who protects the protectors?" becomes a new recursive question for the KYA framework.

Chapter 5: Agent Credit Assessment Framework

When a business is preparing to delegate procurement to an agent, it first needs to know whether the agent can identify suitable products, compare prices, and place an order. What ultimately determines whether the business delegates authority, however, is whether the agent will buy only from approved suppliers, whether it will continue placing orders when the budget is insufficient, whether it can stop an incomplete transaction promptly after the user revokes authorization, and whether problems from previous transactions will be visible before the next one.

These are the questions agent-credit assessment must answer. Chapter 1 explained the economic costs of the trust gap; Chapter 2 defined maturity in terms of "delegable autonomy"; Chapter 3 proposed three dimensions—identity trustworthiness, creditworthiness, and authorization trustworthiness; and Chapter 4 presented agents in multiple ecosystem roles. This chapter turns those judgments into assessment methods that can collect evidence, support comparisons, and inform real decisions. Its purpose is to help businesses and counterparties determine in which scenarios, within what limits, and under what oversight conditions a task can be delegated to a particular agent.

This chapter organizes trust requirements for identity, credit, and authorization into three primary pillars, thirteen secondary dimensions, and sixty-five tertiary indicators, supported by scenario-based assessment, critical thresholds, and dynamic updates. To make the framework clear, it first explains the assessment unit and methodology, then presents the structure and coverage of the baseline indicators, and finally shows how assessment results can be applied in specific scenarios. The chapter centers on verifiable performance records and links scores to actual authorization.

5.1 Methodology and Design Principles

5.1.1 First Specify Which Agent Is Being Assessed

Agent credit cannot simply be attached to a model name or product brand. The same model may be deployed as a customer-service, procurement, or maintenance agent; each deployment uses different data, tools, and permissions and may cause different losses. The minimum unit of assessment should therefore be an agent with a defined identity and version, together with its operating boundaries, scope of actions, and responsibility arrangements.

The operating boundary described in Chapter 2—the ODD—can be understood as the conditions under which an agent is permitted to act autonomously. For example, a procurement agent may be authorized to buy office supplies only from three approved suppliers, subject to a per-transaction limit, and may not change payment-account details. Good records generated within this boundary may support similar procurement later. If the agent begins purchasing production equipment or receives permission to change account details, additional assessment is required. A history of small transactions does not by itself establish that it can handle large or irreversible actions.

This approach is consistent with the assessment logic of the NIST AI Risk Management Framework: specify the context of use, measure system performance in conditions close to actual deployment, and record the limits on generalizing conclusions to other conditions. For KYA, the scenario is not an add-on label after scoring; it is a prerequisite for interpreting every indicator.

5.1.2 Make Assessment Serve Real Trust Decisions

Assessment must first connect "can complete the task" with "is worth continuing to delegate to". Model evaluations can show that an agent has the basic capability to perform a task, but credit also requires observing whether it completes tasks as agreed, respects authorization, and responds appropriately to uncertainty. Capability and credit are related but cannot substitute for one another. Accordingly, this framework includes task quality under creditworthiness while retaining separate requirements for identity verification and permission controls.

According to the Survey Report, among seven KYA design principles, the share of respondents giving accountability and traceability a high score (4–5) was 93% (mean 4.56); machine readability, 92%; scenario-based tiering, 91%; open standards, 82%; gradual evolution and international alignment, 73% each; and technology neutrality, 67%. Respondents placed the greatest importance on being able to trace key actions to the source of authorization and the responsible party, and on matching requirements to the risk of each action. The first three principles form a complete evidence chain: scenario tiering determines what to record, machine readability determines how records can be used, and accountability and traceability determine who the records ultimately point to. Their high scores together indicate that the market's core expectation of assessment is "usable evidence".

In practice, assessment can distinguish two perspectives: outcomes and controls. Outcome metrics ask whether a task was completed and whether complaints or erroneous payments occurred. Control metrics ask whether authorization checks, log protection, and anomaly handling worked. A procurement agent's failure to exceed its authority so far does not prove that its blocking mechanism is reliable; tests are also needed to confirm that over-budget or expired-authorization requests are rejected. Conversely, complete control mechanisms cannot replace real delivery records. Combining observed performance with control testing helps prevent a short period of smooth operation from being mistaken for sustained trustworthiness.

KYA assessment is guided by five principles: machine readability, scenario-based tiering, traceable accountability, gradual evolution, and open standards. Applied in practice, identity, authorization, and event records should use consistent fields that systems can verify; requirements should match the risk of each action; key actions should be traceable to the source of authorization and the responsible party; new agents should be able to build credit gradually through controlled tasks; and, when information is exchanged across platforms, necessary data definitions and verification methods should be openly specified without requiring all participants to use the same technology.

These principles must also account for implementation costs. Businesses do not need to build a complete cross-platform credit network at initial deployment or

generate complex proofs for every routine query. They can begin by recording critical tasks, authorizations, and anomalous events, then add monitoring as the scope of business expands. For permissions already granted, however, there must be verifiable evidence of identity, boundary enforcement, and stop mechanisms. Gradual development should mean opening a smaller scope of business first and expanding it later.

5.1.3 Turn Each Task into Verifiable Credit Evidence

The basic unit of credit assessment should be a task or transaction, not the number of conversations, calls, or hours online. Before a task begins, specify the deliverable, deadline, quality requirements, and authorization boundaries. After it ends, record the actual outcome, reasons for anomalies, and any human intervention. In procurement, product discovery, order placement, delivery, and payment settlement are separate stages; "order generated" must not be recorded as complete performance. Likewise, even if the product is ultimately purchased, bypassing required human confirmation during ordering must be recorded as an authorization violation, not as successful performance as agreed.

The denominator is often overlooked when calculating task-completion rates. Counting only successfully delivered orders makes failed tasks disappear from the record. Conflating user cancellations with errors in the agent's judgment can produce unfair deductions. Original events should therefore be retained in full, with success, partial completion, human takeover, user withdrawal, and external-system failure labeled separately. Businesses may present both overall task outcomes and performance metrics adjusted for verified responsibility, but they must disclose the exclusion rules and cannot let operators select favorable samples.

To enable different institutions to review the same result, each metric should have a concise definition specifying the applicable tasks, observation period, calculation or testing method, evidence source, treatment of missing data, and rule version. For example, whether a customer-service question is counted as completed when "a reply has been sent" or only when "the issue has been resolved" must be agreed in advance. Data collectors should retain original records, while assessors explain how they reached conclusions from those records. This allows the underlying facts to be reverified even if the scoring method changes.

Results also require supporting evidence. This framework distinguishes three data levels: signed or independently sourced evidence; platform runtime and audit records; and operator self-reporting. Critical thresholds should rely primarily on evidence that can be independently verified. Operator self-reporting may provide context but cannot, on its own, prove that a critical control works. Signatures and tamper-resistant mechanisms help verify provenance and record integrity; they do not establish that the recorded content is true. That still requires corroboration from orders, deliveries, payment records, and counterparty feedback.

Recordkeeping should also have limits. Assessment requires retaining decision-relevant facts, tool calls, authorization checks, outcomes, and explanatory summaries; it does not require collecting every conversation or irrelevant personal information for credit scoring. For the metric "completeness of reasoning-trace retention," this chapter

uses an auditable decision-and-execution-trace definition. Model-generated explanations cannot replace actual execution records, and access to the model's entire internal chain of thought is not an admission requirement.

5.1.4 Use Composite Scores to Compare Performance and Critical Thresholds to Constrain Authorization

The indicator system combines composite scoring, critical thresholds, and treatment of major incidents. Composite scores support comparison of general performance; critical thresholds determine whether necessary conditions are met; and major incidents promptly correct risks that historical averages might otherwise conceal. Weights express the relative importance of different indicators for a particular business decision.

When designing a composite score, first specify the decision it is intended to support—for example, whether the agent may complete procurement independently or whether routine human review can be reduced. Then select relevant indicators, standardize the direction of scores according to published rules, and aggregate results through the secondary dimensions and primary pillars. Outcome metrics such as success rates must be assessed against scenario-specific targets. Failure and leakage rates should be converted so that higher scores consistently indicate better performance. Control metrics should test whether mechanisms work; a policy document alone does not warrant a high score.

Weights should reflect the consequences of the task. When the business depends on sustained delivery, performance quality and historical records should carry substantial weight. When an agent can make payments, change account details, or call multiple tools, boundary and emergency controls should have greater influence. When cross-entity cooperation or complex agency relationships are involved, identity and accountability chains deserve more attention. Greater emphasis on one area does not remove the baseline requirements for other pillars. Weighting should also account for reversibility, the possibility of timely human intervention, and the parties who could be affected by errors. This chapter therefore provides principles for weights and thresholds but does not prescribe universal values. Parameters should be calibrated from pilot data for each scenario; the calibration path is described in Section 5.3.3.

The importance of an indicator must be distinguished from the sufficiency of its evidence. A risk-critical indicator should not receive a lower weight merely because data are temporarily unavailable; additional verification is needed or the relevant authorization must remain restricted. Conversely, when several indicators reflect the same task set or risk, their contributions should be combined or capped to prevent a dimension from dominating the total score simply because it has more indicators. Evidence quality determines the confidence that can be placed in a conclusion and cannot be replaced by more self-reported materials.

For implementation, business, risk-management, and technical assessors can jointly propose a weighting scheme and explain how each trade-off relates to business risk, then test it against historical records or controlled pilots. Validation should examine whether the results distinguish subsequent performance, whether small changes in weights materially alter conclusions, and whether serious failures are

masked by high scores elsewhere. If validation is insufficient, report the performance of each pillar and key issues separately, using the composite score only as a supplementary judgment. Comparisons within the same scenario should use consistent rules; later changes must retain version histories and reasons.

In addition to the composite assessment, check the critical thresholds for the relevant scenario. If an agent cannot stop a payment promptly, strong performance elsewhere does not justify granting autonomous payment authority. Serious unauthorized actions or major defaults should immediately restrict the affected business; action must not wait for a monthly score to fall naturally. Assign each incident to one primary indicator for scoring and use related indicators to trigger management measures, avoiding duplicate deductions across multiple indicators.

Finally, an assessment result should state its applicable scenario, performance across the three pillars, threshold status, data sufficiency, permission limits, and reassessment date. New agents or those with insufficient samples should be labeled "under observation" or "provisional rating". If an indicator truly does not apply, explain why and disclose any adjustment to weights. Missing data must not be treated as a perfect score, and missing critical evidence cannot be removed by reallocating weights.

5.2 Baseline Indicator System

5.2.1 Overview: From Trust Assessment to Verifiable Evidence

The indicator system can be understood through three straightforward questions: Whom does the agent represent? Is its past performance trustworthy? Can the authority entrusted to it be constrained? Identity trustworthiness, creditworthiness, and authorization trustworthiness answer these questions, respectively. Together they support a decision about the scope of autonomous action appropriate for an agent under current task and risk conditions. Clear identity enables accountability; performance records support continued delegation; and permission controls define which actions that trust can justify.

The system has three levels: pillars, dimensions, and indicators. A primary pillar identifies the type of trust being assessed. A secondary dimension breaks the question into a coherent aspect for observation. A tertiary indicator specifies the fact to be checked. For example, "authorization trustworthiness" is a pillar; "boundary enforcement, attack defense, and tool integrity" is one of its dimensions; and "rate of blocking unauthorized actions" is a specific indicator that can be tested by checking whether requests exceeding limits or using expired authorization are blocked. The full system comprises thirteen dimensions and sixty-five indicators, forming a library that can be selected by scenario. The number of indicators shows coverage, not scoring weight.

Identity trustworthiness comprises five dimensions and eighteen indicators, covering "attribution—unique identification—authentication safeguards—external verification—lifecycle management". It assesses whether the agent is linked to a responsible party; whether credentials and sessions are secure at runtime; whether external participants can verify the agency relationship; and whether identity is correctly maintained or invalidated after upgrades, transfers, and retirement. Identity

assessment therefore covers the full lifecycle from establishment to exit, rather than a one-time identity check at registration.

Creditworthiness comprises four dimensions and twenty-four indicators covering current performance quality, the integrity of historical records, operational stability over time, and continuity of records across platforms. It includes outcome metrics such as task completion and factual consistency, as well as control metrics such as tamper-resistant logs and anomaly monitoring that support credit assessment. Good outcomes can support further delegation only when records are genuine, observation conditions are clear, and subsequent changes can be detected.

Authorization trustworthiness comprises four dimensions and twenty-three indicators covering four stages: authorization, execution, emergency response, and adjustment. At authorization, the system checks whether the scope of actions, tools, and data is necessary and clear. During execution, it tests whether unauthorized actions and malicious manipulation are blocked. When anomalies occur, it checks whether the agent can stop, downgrade, and limit losses. If tasks change or permissions are delegated, it checks whether authority can be adjusted, constrained through the delegation chain, and revoked promptly. The assessment therefore covers prior agreement, execution, and post-incident response.

Table 5-1 summarizes all dimensions and tertiary indicators. Readers can first use the middle column to understand the scope of each dimension, then consult the right-hand column for specific checks. The following sections explain their business meaning and boundaries of use. The prefixes ID, CR, and PM identify indicators for identity, credit, and authorization, respectively, linking each issue to the corresponding assessment record.

Table 5-1. Overview of the Baseline Indicator System: Three Pillars, Thirteen Dimensions, and Sixty-Five Indicators

Primary pillar	Secondary dimensions and number of indicators	Coverage of tertiary indicators
Identity trustworthiness (5 dimensions; 18 indicators)	ID-1 Attribution of the principal and responsibility (4)	Clarity of principal binding; traceability of the responsible party; clarity of disclosure of the agency relationship; non-repudiation of responsibility.
	ID-2 Identity uniqueness and resistance to impersonation (3)	Completeness of globally unique identifiers; resistance to identity impersonation; cross-platform identity consistency.
	ID-3 Authentication, credentials, and session security (4)	Strength of authentication mechanisms; session isolation; system-prompt leakage risk; resistance to credential leakage and session hijacking.

Primary pillar	Secondary dimensions and number of indicators	Coverage of tertiary indicators
	ID-4 Transparency and third-party verifiability (4)	Compliance with AI identity disclosure; third-party verifiability of identity; clarity of attribution of explanations; clarity of disclosure in human–AI collaboration.
	ID-5 Identity lifecycle and supply-chain integrity (3)	Continuity across the identity lifecycle; traceability of supply-chain provenance; synchronized invalidation of credentials upon retirement.
Creditworthiness (4 dimensions; 24 indicators)	CR-1 Performance capability, output quality, and outcome safety (8)	Task-completion/performance rate; factual consistency of outputs; hallucination rate; output calibration; stability of response quality; coverage of decision explainability; completeness of reasoning-trace retention; completeness of knowledge provenance.
	CR-2 Performance history, reputation integrity, and resistance to manipulation (7)	Historical performance/default rate; dispute/complaint rate; partner evaluations/reputation; completeness of behavioral evidence; authenticity of linking evaluations to transactions; tamper-resistance of behavioral logs; risk-reserve coverage.
	CR-3 Credit stability, runtime monitoring, and security resilience (6)	Behavior-drift detection rate; half-life of credit depreciation; latency of anomalous-behavior detection; post-incident credit recovery; bias-drift detection rate; cumulative monthly penalty points.
	CR-4 Credit portability, independence, and verifiability (3)	Portability of reputation across platforms; clarity of ownership of the credit record; consistency across registries.

Primary pillar	Secondary dimensions and number of indicators	Coverage of tertiary indicators
Authorization trustworthiness (4 dimensions; 23 indicators)	PM-1 Scope of authorization, least privilege, and data minimization (5)	Degree of permission minimization; clarity of authorization scope; completeness of the tool-risk inventory; least-privilege credentials; degree of data-access minimization.
	PM-2 Boundary enforcement, attack defense, and tool integrity (5)	Rate of blocking unauthorized actions; resistance to malicious tools; parameter-injection defense rate; risk assessment of tool-chain combinations; audit coverage for unauthorized actions.
	PM-3 Runtime controls, safe downgrading, and loss limitation (6)	Emergency-stop capability; ability to implement a minimal-risk strategy; controllability of side effects; automatic containment of anomalous behavior; explainability of runtime-control decisions; valid third-party liability insurance.
	PM-4 Dynamic adjustment, delegation governance, and revocability (7)	Timeliness of dynamic permission adjustment; propagation speed of authorization revocation; completeness of delegation-chain monitoring; rate of excessive information disclosure; explainability of delegation-chain decisions; rate at which sub-agent permissions remain within scope; completeness of cascade revocation.

Note: The sixty-five indicators are listed individually based on the project indicator materials; the number after each dimension indicates how many tertiary indicators it contains. Indicator names define what is observed and do not imply a uniformly validated calculation method or passing threshold. “Completeness of reasoning-trace retention” refers in this chapter to auditable decision and execution traces; “tool-risk inventory” means a list of tools and their associated risks. Indicator design draws on the NIST AI Risk Management Framework and its Agentic Profile, OWASP Top 10 for LLM Applications, ISO/IEC 42001, W3C DID and verifiable-credential specifications, and Chinese and international standards including GB/T 45654 and T/ISC 0107.

These indicators are assessed in different ways. Task-completion and complaint rates are calculated from actual task records. Resistance to identity impersonation, blocking of unauthorized actions, and emergency stops require active testing. Principal binding, ownership of records, and provenance traceability require review of credentials and linked records. The half-life of credit depreciation describes how quickly older records lose influence; monthly penalty points aggregate the effects of verified incidents. Both also depend on assessment-rule design and should not be

treated as objective measures of an agent's capability. Post-incident credit recovery should likewise be assessed in light of remediation, retesting, and observation results, not simply the time taken to restore a technical service. Bias-drift detection measures whether fairness performance departs from its baseline over time. When agents screen, distribute, or recommend services to customers, changes in fairness can affect eligibility for cooperation and counterparty trust, so they should be monitored alongside operational stability.

The indicator library also includes arrangements for bearing losses, such as risk-reserve coverage and valid third-party liability insurance. These need to be verified only when a relevant business uses them, including whether their scope, validity, and coverage match the risks. They can supplement risk mitigation but cannot replace good performance or effective permission controls. Not using a particular safeguard should not, by itself and outside the context of the scenario, be treated as a sign of lost credit.

A procurement agent illustrates how one order can provide evidence for several dimensions. Principal binding and valid credentials show who placed the order; delivery, reconciliation, and complaint records show how the task was performed; and budget checks, human-confirmation records, and revocation logs show whether the agent stayed within its boundaries. The evidence should be linked to the same task and used to support separate judgments, avoiding duplicate credit for the same fact. The result should be a scenario- and evidence-based credit profile, not a simple sum of sixty-five scores.

In application, the baseline indicators answer "what to examine"; composite scoring explains "how to compare"; critical thresholds specify "which conditions cannot be offset by high scores elsewhere"; and maturity and scenario rules determine "how far authority may be delegated". Not every agent must mechanically use every indicator, but the relevant requirements for identity, performance, and controls must cover any actions already authorized. The following sections first explain the assessment focus of the three pillars, then discuss critical thresholds and their relationship to maturity.

5.2.2 Identity Trustworthiness: Clearly Attribute Behavior

The core of identity trustworthiness is ensuring that every important action can be attributed to a persistent, verifiable entity. For a business procurement agent, confirming that the business completed identity verification is only the starting point. It is also necessary to confirm which agent issued the current request, whether it is still operated by that business, and whether this purchase falls within a valid agency relationship. Relevant evidence includes principal-binding information, identity credentials, authorization relationships, records of key actions, and version-change records.

This pillar requires particular attention to two issues that are easy to conflate. First, having a unique identifier does not mean that an identity is secure: credentials must also be protected from theft, requests from replay, and information from leakage across sessions. Second, identity continuity does not mean that historical performance should be inherited unconditionally. After a material change to the

model, tools, or operator, historical records should remain linked, while the affected task scope is reassessed to prevent an agent from escaping unfavorable records by changing its identity.

Identity must also be disclosed externally and verifiable by counterparties. When interacting with a procurement agent, the counterparty should be able to tell that it is interacting with an agent rather than a person, verify which business the agent represents, and identify whom to approach in a dispute. In human–AI collaboration, it should also be possible to distinguish whether a decision was made by a human, an agent, or jointly. These disclosures directly inform a counterparty's risk assessment and choice of cooperation terms. Identity integrity also covers provenance and retirement: the models, plugins, and tools used by an agent and their traceable origins determine whether responsibility can be established when something goes wrong. If credentials are not invalidated when an agent is retired, leftover credentials may become an entry point for misuse.

The Survey Report shows that the proportions of identity-trustworthiness elements implemented or partially implemented fall within a narrow range of 60%–69%. Independent authentication and key credentials rank highest at 69%, while machine-readable verification interfaces rank lowest at 60%. Implementation has begun relatively evenly, suggesting that there are established ways to address "who is it?" and "whom does it represent?" The shortfall in machine-readable interfaces warrants attention: they are a prerequisite for cross-organizational verification and protocolization. Without them, identity verification remains within platform-specific closed loops, making portable identity difficult—another platform cannot verify the identity and records accumulated by an agent elsewhere at low cost. Identity carries credit; incomplete attribution propagates to credit and authorization. If a system cannot reliably answer "is it still the same agent?", performance records may be assigned to the wrong party and authorization decisions lose their anchor. The 60%–69% range also means that more than 30% of organizations have yet to fully implement basic identity elements; broader adoption remains a prerequisite for all three dimensions.

5.2.3 Creditworthiness: Let Past Performance Support the Next Delegation

Creditworthiness contains twenty-four indicators and is central to explaining the economic value of agents in this framework. Whether an agent receives less manual review, a longer cooperation period, or a broader task scope should be supported primarily by verifiable performance.

The first question is whether the current task was completed as agreed. Task-completion rate, factual consistency, hallucination rate, output calibration, and response stability capture different types of failure. Calibration can be understood simply as: when an agent says it is confident, is it actually often correct? An agent that recognizes uncertainty and promptly asks for help may be more suitable for delegated tasks than one that gives confident answers without recognizing errors. Assessment should also distinguish the severity of errors rather than count a wording flaw in the same way as an erroneous payment.

The second question is whether the historical record is trustworthy. Historical performance, complaints and disputes, and partner feedback can reflect long-term performance, but evaluations must be linked to real tasks or transactions and their authors must be verified. For example, favorable reviews exchanged among agents using different names but controlled by the same operator do not represent endorsement by multiple independent counterparties. Assessors can examine principal attribution, transaction records, and actual delivery to identify possible signs of collusive or duplicate score inflation, consistent with the requirements for reputation integrity and resistance to manipulation. Frequent cooperation alone, however, does not prove collusion; anomalies must be verified before lost credit is concluded. Task-completion rate and historical performance should also refer to the current observation window and the longer-term record, respectively, to avoid awarding duplicate credit for the same successful tasks under different labels.

The third question is whether past credit remains valid. An agent's performance can change with model versions, tools, data, and operating environments. Behavioral-drift detection looks for such deviations—for example, when a previously stable procurement agent suddenly changes suppliers frequently, retries failed tasks more often, or repeatedly requests expanded permissions. The cause should be investigated, relevant permissions tightened, and additional tests conducted. Time-based depreciation can reduce the influence of stale records, but the effects of a serious incident should be removed only through remediation, retesting, and an observation period, not simply with the passage of time.

The final question is whether credit can continue across platforms. Allowing counterparties to verify necessary performance and risk records can reduce the cost of repeated investigation. Portability does not mean that every platform must accept the same score: records may be mutually recognized while risk judgments remain specific to each scenario. Shared content should be limited to what is necessary, with mechanisms for correction and appeal, so that information unrelated to current performance does not become a long-term burden.

Sample size also matters when interpreting results. Ten successful tasks and a high completion rate sustained across ten thousand tasks do not support conclusions with the same strength. Repeating simple tasks under one template also cannot substitute for testing complex boundary cases. Actual ratings should disclose the observation period, number and types of tasks, and critical failures. Results close to an admission threshold should be reviewed with a reasonable uncertainty interval rather than treated as definitive based on a difference in decimal places.

The Survey Report shows substantial variation within the creditworthiness dimension. The adoption rates of record-based capabilities such as key decision logs (81%), performance history (76%), and anomaly records (76%) are relatively high, while internal credit scoring (52%) and integration with risk-control, audit, and compliance systems (58%) are notably lower. The industry has largely reached the data-collection layer but has yet to move into assessment and decision-making. Records are raw material for credit, not credit itself. Without scoring and integration, existing records cannot inform authorization decisions and remain idle evidence assets. Record-based capabilities show "what happened," while scoring and integration answer "is it worth trusting, and how much authority can be delegated?" Their adoption rates are about

20–29 percentage points lower, indicating that the next phase should shift from "recording everything" to developing the capacity to make judgments that can be used in practice.

5.2.4 Authorization Trustworthiness: Match Credit to a Controllable Scope of Action

Authorization trustworthiness assesses whether permissions are clear, execution is constrained, anomalies can be stopped, and permissions can be withdrawn promptly. This allows the economic value of credit to be used safely: good performance may justify broader authorization, while the external execution system must ensure that an agent remains within limits on amounts, counterparties, time, and tools.

The Survey Report shows that respondents identify hallucinations and unreliable outputs (51%), unauthorized actions and permission management (44%), and poisoning or leakage of knowledge-base data (39%) as the three leading technical pain points, followed by prompt injection (37%) and tool/API/MCP security (29%). This confirms that agent risks come not only from the model itself but also from its network of tools and downstream delegates. The high incidence of unauthorized actions (44%) reflects the direct consequences of insufficiently granular tool permissions and authorization boundaries that are not enforced at every call.

Assessment should therefore examine actual execution rather than merely read prompts. For example, if a user sets a per-transaction procurement limit, the system should check that limit when an order is submitted or payment executed. When multiple sub-agents purchase items simultaneously, the system must also prevent each from acting compliantly in isolation while collectively exceeding the total budget. OWASP's analysis of "excessive agency" identifies too many functions, excessive permissions, and excessive autonomy as factors that can expand harm, and recommends least privilege and independent authorization checks.

Tools should also be assessed as combinations. Before evaluation, create an inventory of the data each tool can read, the actions it can perform, and their possible consequences. Reading customer information and sending email may each serve a legitimate business purpose, but the combined use must still be checked to ensure that customer information is not sent to an unauthorized recipient. Permission to use each tool separately does not mean that every combination is within the task authorization. When tools are added or the order of calls changes, review data flows, cumulative consequences, and the ability to stop or recover when anomalies occur.

When several agents hand work off in sequence, the authorization relationship must be recorded at each level. For example, when a procurement agent delegates price comparison to a sub-agent, that sub-agent is not thereby authorized to delegate payment to another agent. Each delegation should record its upstream principal, permitted actions, and expiry, so the final action can be traced to the original mandate and downstream permissions do not exceed upstream authorization. When the original authorization is revoked, the permissions for related subtasks should also become invalid. These requirements correspond directly to the indicators for

delegation-chain monitoring, preventing permission expansion by sub-agents, and cascade revocation.

These controls must also be tested in emergency exercises. For payments, it is more meaningful to test whether a stop command takes effect before the next irreversible action than to claim that a response-time target in milliseconds has been met. Testing should also cover queued requests, credentials that have not yet expired, and delegated subtasks, so that an interface cannot show "stopped" while execution continues in the background. Emergency-stop, minimal-risk strategy, and cascade-revocation indicators should be verified through end-to-end drills.

Permission changes should not be triggered mechanically by a composite score. Before increasing limits, confirm that the current authorization remains valid, business risks have not changed, historical samples cover the proposed new actions, and no major event remains unresolved. A higher credit score cannot automatically remove human confirmation from an action that previously required it. The convenience earned through strong credit should come from reducing repetitive review within already approved boundaries; additional permissions still require an appropriate authorization basis.

The Survey Report shows that authorization trustworthiness currently follows a pattern of "strong fallback, weak tiering". Human confirmation for high-risk actions has been implemented by 86% of respondents, and tool/API allowlists by 71%; only 52%—the lowest rate among six measures—set differentiated trust levels by business scenario. The prevalence of human gates indicates broad awareness of baseline safeguards. Emergency stops (66%) and anomaly blocking (63%) also reflect attention to emergency mechanisms. However, the lack of scenario-based tiering means authorization strength cannot adjust dynamically with task risk. The direction for this dimension is therefore to move from "whether a gate exists" to "whether it is tiered and whether the tiers use credit data".

5.2.5 Seven Critical Thresholds and Their Management Implications

Weighted scoring can be misleading: strong performance on most items may conceal the failure of a few critical controls. This framework therefore selects seven critical thresholds from the baseline indicators. Their purpose is to constrain admission for the relevant business or autonomy level, not to declare an agent permanently unusable in every scenario. For example, if conditions for autonomous payment are not met, the agent may still retain capabilities for retrieval and procurement recommendations.

Table 5-2. Seven Critical Thresholds and Corresponding Management Actions

Threshold indicator	Question to verify	Primary action if unmet
ID-1.2 Traceability of responsible party	Can key actions be linked to the source of authorization and a specific responsible party?	Suspend autonomous business that cannot be attributed and complete the evidence.

Threshold indicator	Question to verify	Primary action if unmet
ID-2.2 Resistance to identity impersonation	Can identity spoofing, theft, and replay be detected and blocked?	Reauthenticate and suspend affected high-risk access.
ID-3.3 System-prompt leakage risk	Has sensitive information been disclosed, or has a control weakness that could be exploited been exposed?	Restrict relevant permissions in proportion to the consequences; isolate and remediate the leakage path.
CR-1.1 Task-completion rate	Does the agent meet the required performance in the target scenario and on safety-critical tasks?	Narrow the task scope, lower limits, or add human review.
PM-2.1 Rate of blocking unauthorized actions	Can over-limit, expired-authorization, and out-of-bound calls be blocked?	Revoke affected permissions and retest after remediation.
PM-3.1 Emergency-stop capability	Does the stop take effect before the next irreversible action?	Suspend the relevant high-risk autonomous actions.
PM-4.7 Completeness of cascade revocation	When upstream authorization is revoked, do downstream subtask permissions also become invalid?	Suspend the relevant multi-agent delegation permissions.

Note: All seven thresholds are drawn from baseline indicators listed in Table 5-1, with codes retained from the project indicator materials. Management actions are summarized according to the scenario-based interpretation in this chapter. Thresholds should be tied to the declared scenario and scope of actions, and acceptance conditions should be specified before testing.

System-prompt leakage requires particular clarification. OWASP explicitly states that system prompts should not be treated as secrets or security controls. The risks to address are sensitive information included in a prompt and delegation of authentication or permission decisions to the model itself. Accordingly, this chapter retains the threshold but ties the severity of the response to actual security consequences. Repetition of general behavioral instructions should not automatically invalidate identity. If actual credentials are exposed, the response should also include measures under credential security, such as revocation and reset.

Critical thresholds also require a distinction between “acceptance requirements” and “statistical evidence”. A failure in a defined critical test for unauthorized actions or revocation cannot be offset by passing most other cases if it produces an unauthorized external effect. At the same time, passing every test in a limited sample does not prove that future risk is zero. Continuous monitoring is therefore still needed after admission testing. Passing conditions, observation periods, and response times should be set according to task consequences, control capabilities, and pilot evidence; they should not be treated as universally applicable safety standards before scenario validation.

5.2.6 How Credit Scores Relate to Maturity Levels

A credit score describes observed performance; maturity describes the extent of autonomy that may be delegated under defined conditions. They should support one another, but there is no rule that "reaching a particular score automatically raises an agent to L3 or L4". Even if an agent that requires human confirmation for every transaction has never made an error, this does not demonstrate task autonomy without human intervention during normal operations.

To recognize a higher autonomy level, assess whether all three pillars meet the requirements for the declared task scope. For example, an agent may independently complete a procurement workflow, but if a sub-agent can continue placing orders after upstream authorization has been revoked, its performance record alone cannot establish eligibility for the corresponding autonomous execution level. Identify the missing controls and evidence, restrict affected actions, and reassess after remediation and retesting. This keeps upgrades aligned with the scope that can actually be delegated.

Table 5-3. Assessment Focus at Different Maturity Stages

Stage	Evidence to examine	Boundary of authorization decision
L1 Human confirmation	Recommendation quality, human-confirmation records, and blocking of external actions	Outcomes include human selection and correction and are not treated directly as independent agent performance.
L2 Human-anchored credit	Pre-authorized scope, valid credentials, boundary enforcement, and revocation	Limited execution is permitted; trust remains primarily borne by a person or institution.
L3 Hybrid credit	Continuous behavioral record, end-to-end performance, anomaly detection, and takeover	Attributable behavioral credit can form within the approved operating boundary.
L4 Independent credit	Persistent identity, portable records, segregated resources, and arrangements for bearing losses	The agent has relatively independent credit as a limited economic representative, while responsibility arrangements remain explicit.
L5 Agent economy	Cross-entity recognition, transaction clearing, dispute resolution, and risk isolation	The assessment unit expands to the network; a high individual score alone does not prove network maturity.

Note: Maturity definitions follow Chapter 2. L0 is a non-agent baseline operating under explicit rules and is not an independent stage for accumulating agent behavioral credit.

Gradual progression in the indicator system should mean that each addition of autonomy is supported by corresponding evidence. At L3 and above, an agent's own

performance record may play a greater role in trust judgments, but the responsibility of a person or institution should not simply be converted into a weight that can be offset by a high score. If an operator changes, existing guarantees must be reverified, while historical behavioral records remain linked. This encourages credit accumulation while preventing agents from "starting over when ownership changes".

5.3 Recommendations by Scenario

5.3.1 Develop Principles for Scenario-Based Weights

After clarifying how indicators, critical thresholds, and maturity relate, businesses must decide how to apply these judgments to actual operations. Scenario design should first identify which failures could affect delivery, disclose information, or cause irreversible losses; then determine the role of relevant indicators in the composite assessment and specify necessary admission and human-review requirements. An industry label provides only an initial clue. Two customer-service or procurement agents may be authorized to take different actions and should therefore have different assessment priorities. Table 5-4 provides qualitative principles for representative scenarios.

Table 5-4. Assessment Priorities and Weighting Principles for Representative Scenarios

Scenario	Principle for allocating weights	Controls to implement in parallel
General tasks	Determine emphasis based on the task, considering identity attribution, performance, and authorization boundaries without presupposing a ranking among pillars.	Verify the principal and authorization, retain task records, and test stop and revocation mechanisms.
Supply-chain performance	Use actual delivery, reconciliation, and dispute records to assess sustained performance; increase the weight of relevant authorization indicators when payments or onward delegation are involved.	Specify the procurement object and budget, check the link between delivery and settlement, and ensure delegation is traceable.
Third-party API orchestration	Focus on tool-call boundaries, combined execution, and anomaly handling; also test final task outcomes rather than counting only successful calls.	Validate tool permissions and cumulative resource use; test downstream stopping and authorization revocation.
Customer-service automation	Set the relative importance of identity disclosure, answer accuracy, and permission controls based on service commitments and the scope of data access.	Limit the commitments the agent may make, constrain access to customer data, and provide human takeover for disputes or requests beyond authorization.

Scenario	Principle for allocating weights	Controls to implement in parallel
Financial credit assessment	Consider whether outputs directly affect business decisions, focusing on reliability of rationale, outcome quality, and authorization boundaries; strengthen control assessment as autonomy increases.	Clarify institutional responsibility and professional review requirements, limit data and account operations, and retain decision and execution records.
Personal productivity assistance	Distinguish among actions such as organizing, editing, and external sharing; strengthen data protection and permission indicators for sensitive information or irreversible actions.	Use the minimum data-access scope and specify confirmation and revocation conditions for external sharing, deletion, and account operations.

Note: The principles are derived from the business meaning of the three pillars and the scenario and control designs in the project materials. They are qualitative recommendations for weighting; specific priorities should depend on actual actions, risk consequences, and pilot results. They do not constitute a validated fixed ranking.

Different scenarios still require a common control foundation: persistent, valid identity and authorization; verifiable task records; necessary data protection; applicable stop and revocation mechanisms; and reassessment after changes. Specific controls should reflect the actions made available. If external tools may be called, examine the tool inventory and execution boundaries; if authority may be delegated, examine downstream permissions. An agent that only provides recommendations need not undergo payment-control tests that are not yet relevant, but once it receives payment permissions, the corresponding verification must be added.

Scenario adjustments must apply to both assessment and authorization. For a personal assistant organizing local files, task quality and data use may be the main priorities. If it receives permission to send files externally or operate accounts, the consequences of leakage, misdelivery, and unauthorized actions must be reassessed, authorization indicators given greater weight, and requirements for confirmation, blocking, and revocation added. It is not enough to adjust assessment emphasis while retaining the original control conditions.

5.3.2 Using Buy Now, Pay Later to Illustrate Implementation

Buy now, pay later illustrates KYA's economic role. The service provider delivers before receiving payment and assumes some performance or collection risk; the user wants to reduce the cost of prepayment and approval for each transaction. KYA can provide shared grounds for judging the agency relationship and historical performance, allowing both parties to transact within clear boundaries. Chapter 4 discusses buy-now-pay-later protocol design and industry practice in detail; this section uses a simplified example to show how assessment applies to such arrangements.

Suppose a small business uses a procurement agent to purchase standardized digital services and wants to move from prepayment for each transaction to periodic settlement. First, verify the relationship between the business and the agent, the

scope of authorization, and responsibility for the account. Specify who is purchasing the service, who is obligated to pay, and how disputes will be handled. The agent's own performance score cannot replace assessment of the actual payer and its repayment arrangements.

Second, accumulate verifiable records from real tasks. Suppose an agent completes 980 of 1,000 similar procurement tasks during an observation period according to predefined requirements. Its task-completion rate for that period is 98%. Assessment must also examine why the other 20 failed, whether any orders were placed without authorization, and whether deliveries and settlements reconciled. A 98% task-performance rate is not a 98% probability of repayment and cannot by itself determine an appropriate credit limit.

Third, use the results in a limited pilot. The business can set an allowlist, per-transaction limit, and cap on cumulative unsettled purchases based on verified behavioral records, the identity of the payer, the settlement period, and its capacity to bear losses. The agent may procure only within these conditions; out-of-scope requests require additional authorization or human confirmation. New delivery, refund, and settlement events are added to the record for the next assessment.

Finally, verify in operation that controls work as intended. If the agent can still place an order after authorization is revoked, or if a sub-agent continues calling paid services after the upstream procurement agent is stopped, the affected autonomous procurement and delegation permissions should be suspended. Even if no bad debt has occurred, these defects show that the current risk boundary is unreliable. Permissions should be considered for restoration only after remediation, retesting, and an appropriate observation period.

This example shows that the purpose of credit assessment is to support transaction convenience under defined conditions. The operating entity that actually bears the risk ultimately sets the limit under its own rules; KYA must clearly explain the relevant data and restrictions.

5.3.3 Update Continuously and Test the Practical Value of Assessment

Assessment frequency should match the rate at which information changes. Identity status, authorization revocation, and major anomalies should be updated in real time or promptly; task performance can be monitored daily or weekly; performance history can be summarized on a rolling monthly basis; and comprehensive reassessments can be scheduled quarterly and after material changes. These frequencies are implementation guidance. If a critical control fails, the response should not wait for the next scheduled review.

Dynamic updates require both downgrading and recovery paths. After a material model upgrade, an expansion of the toolset, a change in operator, or a serious complaint, specify which business activities are affected and reassess the relevant scope. Ordinary errors may be corrected and followed by observation; major incidents require root-cause remediation, independent review, or appropriate testing. Each restriction or restoration of permissions should record the triggering incident, supporting evidence, scope of impact, reason for the decision, and conditions for

restoration. Operators should have an appeal channel, and verified errors in records should be corrected.

For thresholds that have not yet been validated, first define the failure they are intended to identify. Then collect representative normal and anomalous tasks, have business personnel review the results, and develop candidate decision rules. Initially, observe how the rules perform while existing approvals and permission controls remain in place. Check whether they miss serious problems or frequently block normal tasks, then validate against subsequent tasks and enable them gradually. Complaint rates, quality stability, and anomaly recovery should be calibrated separately; similar statistical formats do not justify using a shared threshold. Recovery conditions should especially include remediation and retesting, not merely the passage of time.

The framework's effectiveness must ultimately be tested against business outcomes. For comparable tasks and risk conditions, compare the burden of human review, performance losses, false blocking, and dispute-resolution efficiency; examine whether ratings continue to distinguish subsequent performance. Also test assessment errors by task complexity, business type, and sample size. If new agents or small businesses are restricted more often, determine whether their controls are genuinely insufficient or whether missing data are being mistaken for lost credit. A score difference alone does not prove unfair assessment. Early pilots should not be judged merely by "how many certificates were issued" or "how many indicators were covered". A reasonable goal is to reduce redundant verification and unnecessary human intervention while keeping risks tolerable and responsibility traceable.

5.3.4 A Credit-Building Path for Agents at Small and Medium-Sized Businesses

In responses to the Survey Report question on the weakest of the three dimensions, 39% selected authorization trustworthiness, 22% creditworthiness, 17% said that all three were weak, and 11% selected identity trustworthiness. Identity received the lowest share (11%), suggesting that it is currently the strongest of the three foundational dimensions.

Identity answers "to whom can conduct be attributed?", credit answers "is the agent worth trusting?", and authorization answers "how much can be delegated?" When authorization boundaries are unclear, even attributable identity and recorded performance still leave decisions dependent on cautious human judgment. The credit flywheel cannot start: accumulated data do not earn broader authority, removing the incentive to record behavior. The 17% who said that all three dimensions were weak also merit attention. For them, isolated improvements will not solve the problem; they need a complete starting checklist, particularly at small and medium-sized businesses.

For small and medium-sized businesses, implementation costs determine whether a credit system will actually be used. They can begin with clearly bounded tasks such as order inquiries, standardized replenishment, after-sales triage, or small purchases of digital services, reusing the platform's existing records for entity verification, orders, payments, and after-sales service. Platforms can supply general-purpose identity, authorization, and logging capabilities; businesses can define task scope, responsible personnel, and anomaly-handling arrangements. Each party

retains evidence according to its responsibilities, reducing duplicate investment. An existing business credit record can inform initial cooperation terms, but the agent's performance and permission controls on specific tasks must still be verified separately.

For implementation, businesses can develop a starter checklist from the indicator library that matches the permissions granted, prioritizing identity attribution, actual performance, authorization enforcement, and anomaly handling. Data should be collected automatically through platform records wherever possible. What is simplified is duplicate reporting and verification; actions already authorized must still meet their relevant thresholds. Platform records should retain provenance and verification paths, and critical facts should be checked against counterparty feedback or sampling. As the scope of tasks expands, add relevant indicators. This avoids requiring small businesses to build every capability at once without turning simplified assessment into an unconditional exemption.

At the starting stage, the priority is to clearly record key facts: which agent accepted which task, what authorization it acted under, what the outcome was, and who intervened. After an observation period with a representative number of tasks, permissions can be expanded based on actual losses and control performance. When samples are limited, participation can be enabled through smaller limits, shorter settlement cycles, or suitable risk-mitigation arrangements, rather than permanently excluding an agent from the market solely because it lacks history.

When a business begins working across platforms, it can provide the necessary verifiable credit summary so counterparties can perform checks. Technical investment can be phased, but identity attribution, authorization boundaries, and records of critical events should be in place from the outset. Small businesses then receive not only automation services, but also behavioral records that can accumulate over time and support future cooperation.

The Survey Report also offers insights for businesses providing agent-credit services to small and medium-sized enterprises. Their leading concerns about agent use are limited technical understanding (49%), high cost (48%), data-privacy concerns (47%), and technical barriers to adoption (44%), followed by compliance risk (29%) and system incompatibility (21%). Awareness and cost form a dual barrier and rank above institutional concerns. This differs markedly from large-enterprise samples, where accountability and compliance issues rank highest, showing that the main challenge in the small-business market is accessibility rather than the sophistication of governance. This ranking has direct product implications: a small-business credit solution should initially reduce complexity through templates and platforms, lowering the cost of understanding and implementation. What should be simplified is the process, not the thresholds: identity attribution, real performance, authorization enforcement, and anomaly handling cannot be removed, although platform defaults can implement each requirement. Privacy concerns affect nearly half of respondents, indicating that platform solutions must also address trust in data custody; lower costs alone will not secure adoption.

The assessment framework developed in this chapter should ultimately create an operational relationship: define tasks and responsibility, verify actual performance, grant authority within controllable boundaries, and update judgments using new

behavioral evidence. It translates the trust issues discussed in the first three chapters into evidence that businesses and counterparties can use, and provides a foundation for Chapter 6's industry-level assessment of agent credit development and progress.

Chapter 6: Agent Development Index

6.1 Basic Principles

6.1.1 Observing Agent Development Through Delegated Tasks

The development and prosperity of the agent economy are reflected in a growing number of tasks completed by agents, the sustained integration of agents into production and daily life by organizations and users, and the formation of payment, collaboration, and exchange relationships around task delivery. This report constructs an Agent Development Index along the chain of capability formation, connectivity, application and adoption, and execution and exchange. It tracks how agents become participants in economic activity and how the breadth and depth of that participation change.

This activity chain has four connected stages. Capability and quality determine the range and quality of tasks that can be delegated. Infrastructure enables capabilities to connect to tools, data, identity, and payment networks. Applications and ecosystems reflect organizational deployment, product adoption, and developer participation. Economic activity further captures sustained execution, commercial revenue, and service exchange. Together, the four stages provide a basis for observing development: tasks have the conditions for execution, products have use cases, collaboration has a network, and operating activity can form economic relationships.

The four stages may progress at different rates from month to month. Growth in tool connections and developer activity can create conditions for later application diffusion; increased application use can bring in new users and use cases; changes in enterprise workloads and payment activity further show how task execution and exchange are unfolding. Putting these movements into one index and decomposing their contributions helps identify the stage of growth and how its drivers follow one another.

6.1.2 Focus the Assessment on Breadth of Adoption and Depth of Use

The index has four primary dimensions—Capability and Quality, Infrastructure, Applications and Ecosystem, and Economic Activity—with 36 indicators in total and primary weights of 2:2:3:3. Applications and Ecosystem and Economic Activity together account for 60%, focusing the assessment on actual adoption by organizations and users, sustained execution, and economic exchange. Capability and Quality and Infrastructure account for 40%, retaining measures of task executability and the conditions for collaboration. This weighting is consistent with the report's aim of observing the development of the agent economy: technological progress enters the composite assessment through application and economic activity.

Breadth of adoption is observed through enterprise payments and deployment, visits to popular applications, and developer and open-source ecosystems. Depth of use is observed through token use, enterprise workloads, and methods of task delegation. Exchange relationships are observed through revenue, payment events and amounts, and participation. The same development process leaves different records across multiple stages. Considering them together helps determine whether

growth is driven by expanding supply, wider adoption, or sustained operation and service exchange. Fixed weights within each dimension allow comparisons across months on a stable scale and help explain each indicator's contribution to changes in the total index.

6.1.3 Connecting Development Measurement with KYA

As agents enter longer workflows, call more tools, and take on payments and delegated tasks, the trust requirements of economic activity become more specific: whom does the actual operator represent, has it performed reliably in the past, and is the current action within valid authorization? The three dimensions introduced earlier—identity trustworthiness, creditworthiness, and authorization trustworthiness—provide a basis for attributing conduct, assessing performance, and defining action boundaries, respectively. The activity chain measured by the index therefore connects directly with KYA's three-dimensional framework.

In this relationship, the index identifies which tasks, connections, and exchanges are expanding, providing a demand map for trust infrastructure. KYA uses evidence about a specific agent's identity, performance, and permissions to determine the tasks, limits, and oversight conditions under which it may receive sustained authorization. Monthly measurement helps readers understand the direction of change in the agent economy and helps businesses and platforms identify which verification, authorization, and settlement mechanisms to improve first, so that emerging tasks and transactions can form sustainable cooperative relationships.

6.2 Indicator System

6.2.1 Four Dimensions and 36 Indicators

Table 6-1. Economic Meaning and Indicator Composition of the Four Dimensions

Dimension and weight	Number of indicators	Economic meaning and representative indicators
A Capability and Quality 20%	11	Foundations for task execution: tool tasks, software engineering, web operations, long-horizon tasks, and reliability.
B Infrastructure 20%	8	Conditions for networked collaboration: tool supply, interoperability components, identity registration, and payment access.
C Applications and Ecosystem 30%	9	Breadth of adoption and ecosystem development: enterprise payments and deployment, dynamically popular applications, developer participation, and framework and research supply.
D Economic Activity 30%	8	Sustained execution and exchange: disclosed commercial revenue, cumulative workload, compute use, protocol payments, and forms of task delegation.

Capability and Quality contains 11 indicators covering tool use, software engineering, computer and web operation, collaboration with enterprise tools, long-horizon task capability, safety, and factual reliability. Task-completion capability and reliability together determine which tasks an agent can accept, providing the technical basis for KYA to identify performance requirements by scenario. Infrastructure contains 8 indicators covering tool supply, interoperability, identity registries, and payment access, measuring the conditions under which agents can discover resources, connect to services, and collaborate.

Applications and Ecosystem contains 9 indicators covering enterprise payments and production deployment, dynamically popular applications, open-source supply, developer participation, research, and framework distribution, showing the expansion of adoption channels and the supply ecosystem. Economic Activity contains 8 indicators measuring commercial revenue, enterprise execution, compute use, protocol payments, and direct task delegation. Its 30% total weight is allocated as follows: 7.5% to commercial revenue, 5% each to workload and token use, 10% in total to four payment-activity indicators, and 2.5% to the share of direct task delegation. Payment SDKs are classified under Infrastructure, while payment events, amounts, and participation are classified under Economic Activity, distinguishing the development of access from the expansion of exchange.

Table 6-2. All 36 Indicators and Their Total Weights

Code	Indicator and object measured	Total weight w	Scoring rule
A01	Overall success rate on tool-use tasks	2%	Relative value
A02	Software-engineering resolution rate	2%	Relative value
A03	Computer-operation success rate	2%	Relative value
A04	Web-workflow success rate	2%	Relative value
A05	Enterprise-tool collaboration success rate	2%	Relative value
A06	Long-horizon task capability (human-expert-equivalent task duration)	2%	Logarithmic
A07	Safety utility under prompt injection	1.25%	Relative value
A08	Accuracy on fixed factual questions	1.25%	Relative value
A09	Overall capability of foundational models	3%	ECI increment
A10	Safety performance of foundational models	1.25%	Relative value
A11	Factual consistency in summarization	1.25%	Version-segmented

Code	Indicator and object measured	Total weight w	Scoring rule
B01	Tool-connectivity supply (existing catalogs, retrospectively dated by the latest release version)	20% ÷ 9	Logarithmic
B02	Dispersion of tool supply	20% ÷ 9	Relative value
B03	Measured availability of remote services	20% ÷ 9	Neutral score of 100
B04	Distribution intensity of interoperability SDKs	20% ÷ 6	Per-package normalization after series joining
B05	Interoperability support rate across fixed platforms	20% ÷ 6	Neutral score of 100
B06	Scale of identity-registration IDs across two chains	20% ÷ 12	Logarithmic
B07	Coverage of effective verification	20% ÷ 12	Neutral score of 100
B08	Distribution intensity of fixed payment SDKs	20% ÷ 6	Logarithmic
C01	Enterprise paid adoption (Agentforce contract disclosures)	3.75%	Logarithmic
C02	Disclosed enterprise production deployments	3.75%	Logarithmic
C03	Average visit activity of the dynamically selected top ten agent applications	6%	Geometric mean, then logarithmic
C04	New supply of open-source projects	2.5%	Logarithmic
C05	Developer participation	2.5%	Logarithmic
C06	Development and maintenance activity	2.5%	Logarithmic
C07	Research supply	3.75%	Logarithmic
C08	Distribution intensity of fixed development frameworks	3.75%	Per-package normalization after series joining
C09	Public search interest	1.5%	Relative value
D01	Commercial revenue (Sierra ARR disclosures)	7.5%	Logarithmic
D02	Enterprise execution (cumulative Agentforce and Slack AWUs)	5%	Logarithmic

Code	Indicator and object measured	Total weight w	Scoring rule
D03	Agentic compute activity	5%	Logarithmic
D04	Intensity of protocol-payment transfer events	2.5%	Logarithmic
D05	Intensity of protocol-payment amounts	2.5%	Logarithmic
D06	Activity intensity of grouped payer addresses	2.5%	Logarithmic
D07	Activity intensity of grouped payee addresses	2.5%	Logarithmic
D08	Direct task-delegation rate in the API sample	2.5%	Relative value

6.2.2 Monthly Observation and Dynamically Popular Applications

The index is calculated monthly, with an observation period from October 2025 through September 2026. Traffic measures such as website visits, software downloads, token use, and protocol payments are converted to daily averages for the full calendar month. Ratios and evaluation results retain their measurement units; identity registrations use end-of-period totals; enterprise contracts, revenue, and cumulative workload are updated in the months when publicly disclosed. Low-frequency data carry forward the latest reported value between disclosures, while early months are backfilled using the first available value. B03, B05, and B07 use a neutral standard score of 100, with a combined weight of 7.2222%; their incremental contribution is zero during the observation period.

The popular-application traffic component, C03, uses a dynamically selected top-ten sample. Each month, the 15 leading entries in the “Agent Rankings” published by AICPB (AI Product Benchmark) are treated as candidates. Multiple entry points for the same standalone application are consolidated, retaining the entry with the highest traffic for that month; ten applications are then selected by traffic volume. Monthly reselection allows new adoption channels to enter the observation promptly. Over 12 months, the sample contains 120 selected entries and 19 distinct applications. The original monthly rankings are listed among this chapter’s public data sources.

Daily visits to the selected applications are combined using an equally weighted geometric mean. Let T denote the ten applications selected for a month, V the monthly visits to an application’s primary entry point, and D the number of calendar days in the month. The monthly raw value for C03 is:

$$X_t = \exp\left(\frac{1}{10}\sum_{i \in T_t} \ln\left(\frac{V_{it}}{D_t}\right)\right)$$

The geometric mean gives the same effect to the same proportional change in visits for each application and represents the average activity of a typical popular application. Observing total visits, the median, and the selected list in parallel helps distinguish changes in leaders’ aggregate traffic, product distribution, and application turnover, making adoption growth more specific and interpretable.

Software distribution uses fixed package builds. C08 comprises four Python packages—LangGraph, CrewAI, AutoGen AgentChat, and Smolagents—each normalized separately and weighted at 25%. B04 comprises MCP Python, MCP TypeScript, and A2A TypeScript, weighted at 25%, 25%, and 50%, respectively. The Python distribution series is joined to a common observation month of February 2026: the original series is retained through that month, and subsequent values are extended according to the year-over-year change in actual package-download requests. Actual packages include .whl, .tar.gz, and .zip files; five types of mirror installer are excluded. Let O denote the original series, A the daily average of actual package-download requests, p a software package, and τ February 2026. Then:

$$x_{pt} = O_{p\tau} \times \frac{A_{pt}}{A_{p\tau}} (t > \tau)$$

The joined x represents distribution intensity on a common scale; each package is scored first, then the package basket is aggregated. The TypeScript component is calculated from the original fixed npm package series, bringing package-install activity into monthly comparison on a continuous scale. [1–2]

6.2.3 Fixed References and Standardization

The raw indicators use different units and are first converted into comparable standard scores S. For general indicators, the arithmetic mean of raw values from October to December 2025 serves as the fixed reference c. This reference sets the scoring scale for each indicator; October 2025 sets the starting point for displaying the total index. Fixed references remain unchanged in subsequent releases, allowing new months to be compared continuously.

For relative measures such as success rates, proportions, and search interest, the monthly value is divided by the fixed reference:

$$S_{jt} = 100 \times \frac{x_{jt}}{c_j}$$

For scale measures such as the number of tools, software distribution, website visits, tokens, revenue, workloads, and payment intensity, apply a logarithmic transformation and then normalize using the mean log value d over the reference period:

$$z_{jt} = \ln(1 + \frac{x_{jt}}{c_j})$$

$$S_{jt} = 100 \times \frac{z_{jt}}{d_j}$$

Here, d is the mean of the three z-scores from October to December 2025. The logarithmic transformation compresses multiplicative differences among scale measures, allowing sustained growth in different parts of the activity chain to contribute within the same framework. A standard score of 100 represents the indicator's reference level; scores above or below 100 indicate performance above or below that level, respectively.

Special indicators are handled according to their measurement structure. The foundational-model capability index (ECI) uses "100 + current-month ECI − reference ECI," with a reference value of 146.5. Factual consistency in summarization is normalized by evaluation version: the older version in October 2025 is set to 100, while the newer version from November onward uses the November–December mean of 95.71 as its reference. Interoperability SDKs and framework distribution are first normalized at the software-package level, then aggregated using fixed basket weights.

6.2.4 Weighted Aggregation and Contribution Decomposition

Let b denote one of the four dimensions and v the fixed weight of an indicator within that dimension, with within-dimension weights summing to 1. First, calculate the weighted dimension score B; then combine the dimensions using weights of 2:2:3:3 to obtain J; finally, set the total index for October 2025 to 100:

$$B_{bt} = \sum_{j \in b} v_{bj} S_{jt}$$

$$J_t = 0.2B_{At} + 0.2B_{Bt} + 0.3B_{Ct} + 0.3B_{Dt}$$

$$I_t = 100 \times \frac{J_t}{J_{2025-10}}; I_{2025-10} = 100$$

The four dimension scores for October 2025 were 96.57099, 93.15183, 98.22188, and 79.36365, respectively, producing a weighted score of 91.22022. Applying the same method in September 2026 gives 181.50863. The September total index is therefore 100 × 181.50863 ÷ 91.22022 = 198.98. The increase from 100 to 198.98 represents a 98.98% rise in the composite assessment relative to the base period.

Let w denote an indicator's total-index weight, equal to its primary-dimension weight multiplied by its within-dimension weight. Let s and t denote the starting and ending months, respectively. An indicator's point contribution equals the change in its standard score multiplied by its total weight and a common base-period conversion factor:

$$\Delta I_j(s,t) = \frac{100}{J_{2025-10}} w_j (S_{jt} - S_{js})$$

$$\sum_j \Delta I_j(s,t) = I_t - I_s$$

For example, C03 has a fixed reference c of 180,395.52590 and d of 0.692692489. The daily average geometric-mean visits in October 2025 and September 2026 were 165,155.78502 and 318,290.61948, respectively. Applying the logarithmic formula gives standard scores of 93.83519 and 146.79315. C03 has a 20% weight within the Applications dimension and a 6% weight in the total index, so its full-period contribution is 100 ÷ 91.22022 × 6% × (146.79315 − 93.83519) = 3.48 points. The contributions of all 36 indicators sum exactly to the 98.98-point increase in the total index.

Month-over-month growth is calculated as "current-month total index ÷ previous-month total index−1". Point contributions for individual indicators and dimensions

explain the sources of growth, while the month-over-month rate compares the pace of change. Calculations use unrounded values; results in tables are shown to two decimal places.

6.3 Index Results and Presentation

6.3.1 Overall Trend and Monthly Results

From October 2025 to September 2026, the Agent Development Index rose from 100 to 198.98, an increase of 98.98 points. Month-over-month growth peaked at 20.76% in February 2026, the largest monthly increase in the observation period. Growth accelerated again toward the end of the period: 2.37% in July, 2.73% in August, and 4.40% in September. The total index continued to rise, while the sources of growth shifted successively among compute use, the application ecosystem, connectivity, and enterprise execution.

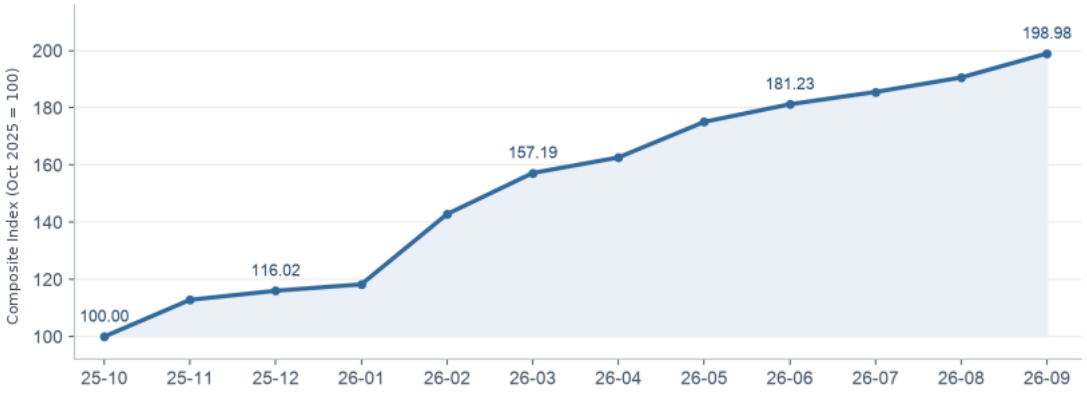


Figure 6-1. Monthly Trend in the Agent Development Index

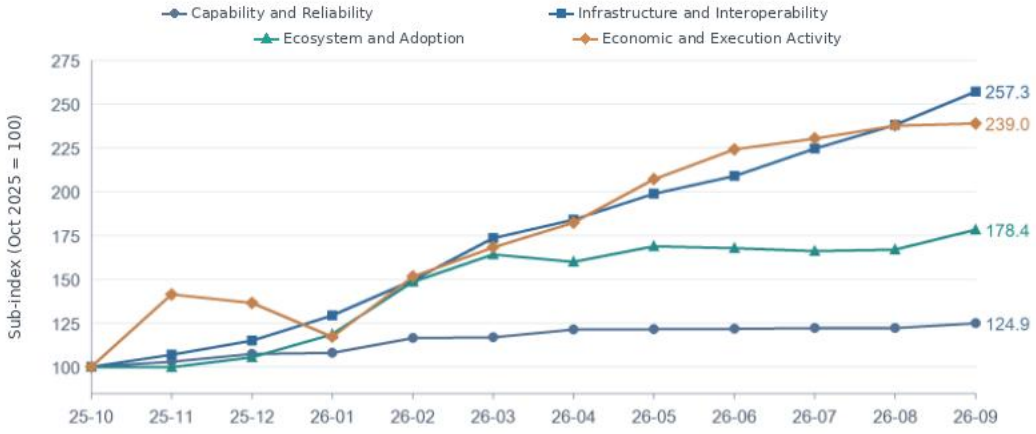


Figure 6-2. Monthly Trends in the Primary Dimensions of the Agent Development Index

Table 6-3. Monthly Total Index and Dimension Scores

Month	Total index	Month-over-month	A Capability	B Infrastructure	C Applications	D Economic Activity
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

All four dimensions made positive contributions over the full period. Economic Activity contributed 36.28 points, Infrastructure 32.12 points, Applications and Ecosystem 25.31 points, and Capability and Quality 5.28 points. Applications and Ecosystem and Economic Activity together contributed 61.59 points, or 62.2% of the net increase over the full period, consistent with the index's focus on adoption and economic realization. Infrastructure provided another important source of support.

Indicator-level contributions show the sources of growth in greater detail. Tokens contributed 29.82 points, or 30.1% of the net increase over the full period. Tool connectivity contributed 11.32 points, while interoperability and payment SDKs together contributed 18.61 points. Within Applications and Ecosystem, open-source projects, developer participation, maintenance, and research supply contributed a combined 15.11 points, or 59.7% of the dimension's net increase. Dynamically popular applications contributed 3.48 points and framework distribution 3.91 points. Execution activity, connectivity, and developer supply together formed the main sources of incremental growth during the observation period.

6.3.2 Contributions by Stage and the Succession of Growth Drivers

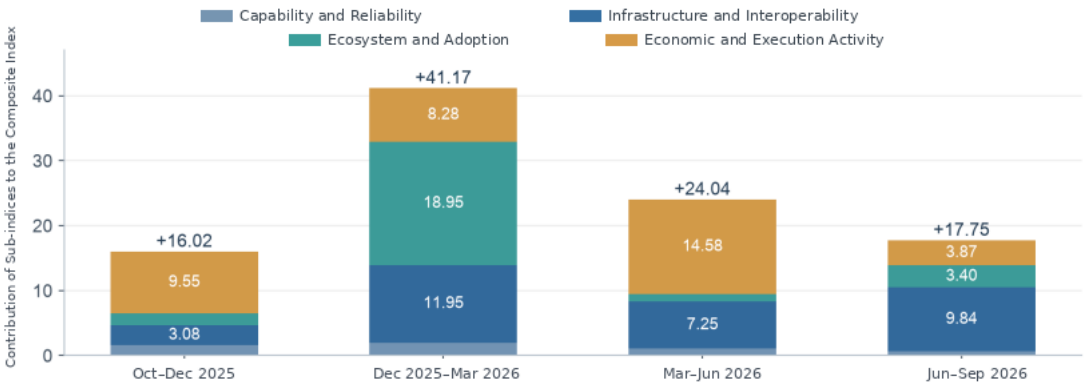


Figure 6-3. Dimension Contributions by Stage, in Index Points

Table 6-4. Changes Between Stage Endpoints and Dimension Contributions

Start and end months	Total change	A Capability	B Infrastructure	C Applications	D Economic Activity
2025-10 to 2025-12	16.02	1.59	3.08	1.81	9.55
2025-12 to 2026-03	41.17	1.99	11.95	18.95	8.28
2026-03 to 2026-06	24.04	1.06	7.25	1.15	14.58
2026-06 to 2026-09	17.75	0.64	9.84	3.40	3.87

Between October and December 2025, the index increased by 16.02 points, with Economic Activity contributing 9.55 points and driving early growth. From December 2025 to March 2026, it increased by 41.17 points—the largest gain of any stage—with Applications and Ecosystem contributing 18.95 points. Between March and June 2026, the index rose by 24.04 points, with Economic Activity contributing 14.58 points as the focus shifted toward sustained execution. From June to September, the index increased by 17.75 points. Infrastructure contributed 9.84 points, or 55.4% of the stage's net increase, making connectivity development the main source of support in the later period.

The narrowing of stage-level gains coincided with a shift in the distribution of growth drivers. In the second and third quarters of 2026, the 35 indicators other than tokens contributed 15.50 and 15.30 points, respectively, remaining broadly stable. Changes in the token component accounted for 96.8% of the difference in total growth between the two stages. The token component for August and September was carried forward from July; other activity measures continued to update. The later-period curve therefore reflects growth more in connectivity, popular applications, and enterprise execution. The largest single indicator's share of all positive contributions fell from 30.7% in the second quarter to 18.2% in the third, indicating a more balanced distribution of the main growth drivers.

At the same time, the number of rising indicators fell from 17 in the second quarter to 14 in the third, while the number of declining indicators increased from four to six. Overall growth, more dispersed contributions, and localized adjustments occurred together, indicating that different parts of the agent economy are developing at different rates. Examining the total index, dimension contributions, and specific activity records together makes it possible to understand which areas have established sustained growth and which are being reorganized.

6.4 Interpretation of the Index

6.4.1 Rising Execution Density Makes Performance Central to Economic Value

The most pronounced change during the observation period was the expansion of execution activity. The token component contributed 29.82 points, or 82.2% of the net increase in Economic Activity; the other seven indicators in the dimension contributed 6.46 points in total. Average daily token use rose from 58.191 billion in October 2025 to 5.82 trillion in July 2026, showing rapid growth in agent operations and compute use. In the second quarter of 2026, tokens, cumulative workload, and protocol payments contributed 8.54, 2.02, and 4.01 points, respectively, creating a period of joint support from execution and exchange.

Enterprise use gives this change a more concrete business meaning. In May, Salesforce reported that cumulative workload across Agentforce and Slack had reached 3.8 billion AWUs. In August, it reported that the quarter ended July 31 had delivered 3.2 billion AWUs in the quarter, up 97% sequentially, bringing the cumulative total to 7 billion. AWU includes activities such as prompt processing, reasoning chains, and tool calls, reflecting workloads from agents' sustained entry into enterprise systems. Viewed alongside product visits, these figures show adoption extending from users entering products to continuous execution in business systems.

The denser the execution, the more businesses need to know what tasks these activities serve, what they deliver, and at what cost. A single business delegation may include multiple rounds of reasoning, tool calls, and human takeovers. Its economic value is ultimately reflected in delivery acceptance and sustained use. KYA links runtime activity to performance assessment by connecting task commitments, process events, resource use, and acceptance outcomes, allowing past execution to support the credit judgment for the next delegation. For businesses, development priorities consequently extend to reliable delivery, per-task cost, and sustained workflows with clear accountability.

6.4.2 Application Diffusion Through Product Turnover and Changes in Leader Distribution

Applications and Ecosystem contributed 18.95 points in the first quarter of 2026, the largest dimension-level gain during that stage. Monthly visits to the OpenClaw main entry point rose from 2.63 million in January to 27 million in February, while Moltbook entered the top ten with 22.33 million monthly visits. The daily geometric-mean visits to popular applications rose to 325,700 in February. Decomposing C03 into visit changes among continuing applications and turnover in the selected list, of the

36.02 standard-score points added between January and February, applications selected in both months contributed 21.28 points and sample turnover contributed 14.74. The expansion of existing products and the arrival of new entry points together drove the increase in activity.

Entry and turnover became even more prominent in the third quarter of 2026. The popular-application component made a net contribution of 2.495 points during that stage: monthly changes in the selected list contributed 3.126 points, while visit changes among continuously selected applications offset 0.631 points. WorkBuddy, Baidu Dazi, and Coze entered the top ten in July; Muse entered in September with 14.46 million monthly visits. For a market of continuously changing products, tracking popular applications each month can promptly show the entry points through which users are beginning to organize work and daily-life tasks.

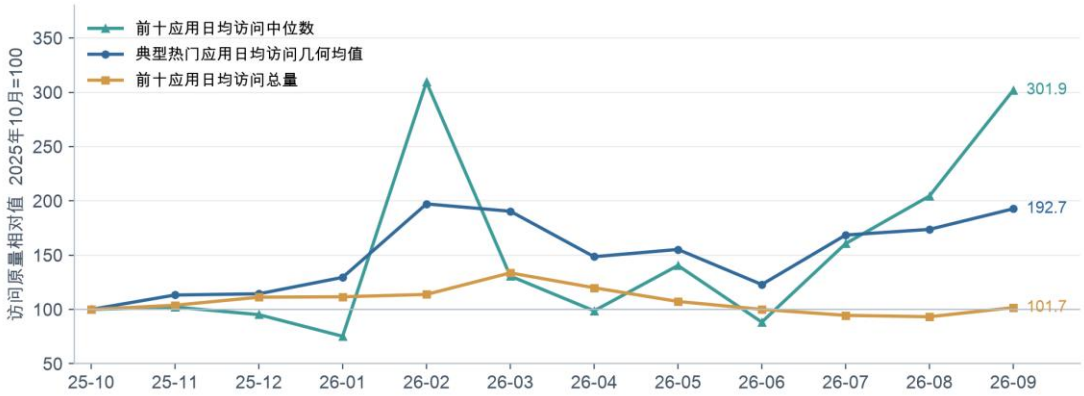


Figure 6-4. Changes in Popular-Application Activity and Visit Distribution

Changes in the distribution among leading applications offer another perspective. Over the full period, total daily visits to the top ten applications increased by 1.70%, while the median daily visits rose from 134,400 to 405,700, an increase of 201.94%. Nano AI's share of visits in the top-ten basket fell from 79.2% to 57.9%. The increase in activity among typical popular applications was therefore reflected more in visits spreading across multiple products. From June to September 2026, search interest in "AI agent" fell by 40.9%, while the daily geometric-mean visits to popular applications rose by 56.7%. General interest in the concept and use of specific products moved at different rates, indicating that adoption should be examined at the product and task levels.

For KYA, rapid product entry requires a clear initial identity and accountability anchor, while ongoing product turnover requires verifiable continuity of performance. New applications need to build their first credit records through clearly bounded tasks; existing applications need updated verification after changes to models, tools, or permissions. Linking the responsible party, runtime version, and performance facts helps users identify a suitable service for each delegation and enables reliable products to build sustained user relationships.

6.4.3 Expanding Connectivity Increases the Importance of Cross-Entity Authorization

Infrastructure contributed 9.84 points in the third quarter of 2026. Tool connectivity supply, interoperability SDKs, and payment SDKs contributed 3.97, 3.11, and 1.92 points, respectively. Over the full period, tool connectivity supply contributed 11.32 points, while interoperability and payment SDKs together contributed 18.61 points. Development tools and shared specifications moved in tandem with SDK distribution and application development. OpenAI released AgentKit, integrating workflow building, connectivity, and evaluation into a developer toolkit. The Linux Foundation established the Agentic AI Foundation, with MCP among its founding projects.

Identity infrastructure also expanded. The total of the highest minted token IDs on Ethereum and Base, used as a measure of registration scale, rose from 49,562 in February 2026 to 148,527 in September, nearly tripling. ERC-8004 organizes collaboration mechanisms around identity, reputation, and verification, providing a vehicle for agents to be discovered on open networks and to exchange performance information. Growth in tools, SDKs, and identity infrastructure means a single task can connect more services and operators, making relationships among participants more complex.

The economic role of connectivity is to organize previously dispersed capabilities into composable services. As orchestrators, specialized sub-agents, and external tools participate in a task, trust judgments must travel through delegation relationships: who initiated the request, who may delegate further, what data downstream agents may read and which actions they may take, and who must stop the process or bear responsibility when an anomaly occurs. KYA therefore needs to bind identity to the current delegation and transmit data scope, tool scope, budget, and validity period with the authorization. Downstream execution must remain within upstream authorization, and revocation must cover the relevant execution chain. These are important operating conditions for connectivity growth to enter enterprise business and cross-entity collaboration.

6.4.4 More Frequent, Granular Payments Require Coordination with Service Metering and Settlement

The structure of payments changed differently from compute use. Over the full period, daily transfers in the protocol-index sample rose from 132,900 to 435,400, an increase of 227.6%. Daily nominal value fell from US$129,200 to US$32,800, while the average nominal value per indexed event declined from about US$0.972 to US$0.0754. Payment-address activity fell from a daily average of 9,324.23 to 3,013.40, a decline of 67.7%. Event intensity increased, nominal value per indexed event fell, and payment-address activity contracted at the same time, showing divergence across payment intensity, amount, and breadth of participation.

Infrastructure access and exchange activity also followed different paths. In the third quarter, payment SDKs contributed 1.92 points, while the four payment-activity indicators together subtracted 1.87 points, for a combined increase of about 0.05 points. Over the full period, the four payment-activity indicators together subtracted

1.16 points. Adoption of payment interfaces expands the conditions for exchange, but sustained transactions also require service supply, demand matching, metering, and acceptance. x402 V2's support for mechanisms such as session reuse and usage-based billing offers one way to organize these service relationships.

For frequent, low-value services, requiring human confirmation for every call raises the cost of collaboration; extending a single authorization without limits obscures cumulative spending. KYA can implement trust arrangements at the session or task level: verify the payer's identity, set total budgets, frequency, and tool limits, continuously record calls and delivery, and settle against acceptance. Payments, metering, service outcomes, and dispute handling must use the same task identifier for granular exchanges to form clear performance relationships and for long-term service providers to accumulate reusable transaction credit.

6.4.5 Capability Expansion Requires Actual Delegation Organized by Task Risk

Capability and Quality contributed 5.28 points over the full period, with long-horizon task capability accounting for about 60% of this contribution. The equivalent duration required for human experts to complete comparable tasks rose from 203.01 minutes to 1,044.78 minutes, 5.15 times the initial level, reflecting greater task complexity that agents can execute. In the third quarter, web-workflow success accounted for about 58.6% of that stage's capability increase, showing that progress was extending into web operations and existing business workflows. Longer tasks and broader tool access provide the technical conditions for businesses to organize complex delegations.

Delegation patterns reflect organizational choices. In the API sample observed by D08, the share of direct task delegation fell from 66.30% to 58.22%, a decline of 8.08 percentage points. The expansion of capability and changes in delegation patterns indicate that businesses still need to work out how to configure autonomous execution, human oversight, and economic permissions. Complex tasks usually comprise multiple separable steps, each with different data sensitivities, reversibility, and financial responsibility. The appropriate scope of delegation should be refined accordingly.

KYA can translate capability gains into performance credit accumulated by scenario and use it to support differentiated authorization. For similar tasks with stable performance and verifiable outcomes, the batch size and limits can be expanded. When moving into a new business area or assuming greater economic responsibility, add the relevant verification, staged authorization, and human-takeover conditions. Capability assessment, performance records, and permission arrangements then form a continuous relationship. Businesses can expand delegation gradually on a verifiable basis, making it easier for agents' technical potential to enter stable workflows.

6.5 Implications for KYA Development

The growth structure shown by the index establishes clear priorities for KYA development. Increased execution requires performance credit that can support

future delegation; product turnover and the expansion of participants require continuity of identity and responsibility; expanded tool and service connectivity requires authorization across delegation chains; and more granular exchanges require coordination among metering, acceptance, and settlement. Identity trustworthiness, creditworthiness, and authorization trustworthiness should work together around the same task so that every delegation has a verifiable principal, performance basis, and action boundary.

6.5.1 Accumulate Reusable Credit Through Task Performance

Growth in tokens and enterprise workloads makes task-level performance evidence the foremost foundation for KYA. Businesses and platforms should link the delegation objective, promised outcome, execution environment, resource use, human intervention, and acceptance record to a specific task. Each tool call should be traceable to the task it belongs to; each acceptance decision should correspond to the actual executing agent; and authorization changes and anomaly handling during execution should be retained. Task-level records thus organize operational activity into verifiable performance history.

Performance credit should accumulate by task type, operating environment, and economic permissions. When identity, performance, and critical permission thresholds continue to be met, consistent completion of similar tasks can support larger batches, higher limits, or broader pre-authorization. When entering a new task domain, additional verification and oversight should be added. Records should distinguish execution errors, user withdrawals, external-service failures, and human takeovers to ensure accurate attribution. Credit has economic value in this use: it reduces repeated review, supports differentiated authorization, and provides common facts for dispute resolution.

6.5.2 Maintain Identity Across Product Changes and Delegation Relationships

Turnover among popular applications and growth in identity registrations require KYA to support both new entrants building records and ongoing operators maintaining them. Identity profiles should link the operating principal, actual runtime instance, model and tool versions, delegator, and current status, allowing counterparties to verify who is executing the current task, whom it represents, and how historical records relate to current operations. After material changes to versions or permissions, update the profile and reverify the affected task scope.

New entrants can build initial credit through an anchor to the responsibility of a business or user, small budgets, and clearly bounded tasks, gradually expanding delegation after verified performance. Cross-platform recognition should transmit verifiable identity and performance facts along with their task conditions and period of validity, allowing users to make judgments for their own scenarios. This preserves an entry path for new products and allows reliable services to retain accumulated credit through upgrades and platform migration.

6.5.3 Enforce Permission Controls Across the Full Delegation Chain

As tool connectivity and interoperability expand, authorization extends beyond a single application to orchestrators, sub-agents, and external services. KYA should encode the task objective, data scope, available tools, spending budget, validity period, and conditions for further delegation in verifiable authorization. Each recipient in the chain should verify the source of delegation, and downstream permissions should be limited to the upstream scope. When permission is revoked, relevant sub-agents and tool credentials should become invalid at the same time. Responsibility and authorization can then be traced along the action chain.

Sustained workflows can be supervised according to action risk: low-risk, reversible actions may be pre-authorized, while fund transfers and important data operations should have limits, frequency controls, and review requirements. Approaching a budget limit, anomalous calls, or a critical version change can trigger throttling, permission reduction, or human takeover. Identity, performance credit, and task risk together define the executable scope, allowing businesses to reduce repetitive confirmation and improve the efficiency of continuous delegation within clear boundaries.

6.5.4 Connect Service Metering, Payments, and Performance Settlement

For high-frequency, low-value, and continuous exchanges, KYA should link the payer, budget credential, call records, metered results, delivery acceptance, and fund flows to the same task. Verify identity and spending authority before payment; control cumulative cost and repeated calls during execution; and record acceptance, refunds, and disputes after delivery. For services priced per call, per session, or by ongoing task, billing units should correspond to actual delivery, making payment records verifiable evidence of performance.

Session and task budgets can manage multiple calls within a single authorization scope and define blocking conditions for anomalous frequency, duplicate charges, and undelivered services. For larger or longer-running tasks, milestone acceptance and staged release of limits may be used. Aligning metering, payment, and outcomes can reduce manual reconciliation and dispute costs, turning connectivity into repeatable service transactions.

6.5.5 Set Priorities Based on Development Patterns and Validate Continuously

Businesses can begin by creating task-level credit records in scenarios with stable workloads, verifiable outcomes, and clear responsibility, then expand authorization as performance accumulates. Platforms can prioritize identity continuity, delegation-chain verification, and budget controls as new applications enter, agents execute across tools, and services are settled continuously. Cross-platform collaboration can begin with shared, verifiable identity and task events, retaining provenance, operating conditions, and version information so that participants can reuse evidence and make judgments for their own scenarios.

The monthly index provides an ongoing basis for observing these priorities. When connectivity contributions rise, focus on cross-entity authorization and revocation. When application turnover accelerates, focus on initial credit and version verification. When execution activity expands, focus on acceptance, costs, and performance attribution. When payments become more granular, focus on cumulative budgets and automated settlement. Progress in KYA can be tested in specific scenarios through repeated-verification time, sustained performance on similar tasks, permission adjustments, and dispute-resolution efficiency. Connecting development measurement with task practice allows identity to carry behavior records, performance to form credit, and credit and risk to inform authorization, with each new execution updating the assessment.

Public Data Sources for the Index

[1] PyPI, download statistics methodology and package installation requests.
[2] ClickHouse, public software distribution data from ClickPy.
[3] OpenAI, “Introducing AgentKit”.
[4] Linux Foundation, announcement of the Agentic AI Foundation’s formation.
[5] Salesforce, FY26 Q3 earnings report.
[6] Sierra, $100 million ARR milestone.
[7] Sierra, “Year Two in Review”.
[8] Salesforce, FY27 Q1 earnings report.
[9] Salesforce, definition of Agentic Work Units.
[10] Tencent, WorkBuddy product overview.
[11] Baidu AI Cloud, product documentation for Baidu Dazi (Dumate).
[12] Meta, introduction to the Muse personal AI agent.
[13] Salesforce, FY27 Q2 earnings report.
[14] OECD.AI, OpenRouter data and agentic computing usage.
[15] Ethereum Improvement Proposals, ERC-8004.
[16] Official ERC-8004 repository, registry contracts.
[17] x402scan, protocol payment index.
[18] x402 team, “Introducing x402 V2”.
[19] MCP Registry, tool registry.
[20] npm, x402 package and distribution interface.
[21] AICPB, October 2025 agent rankings.
[22] AICPB, November 2025 agent rankings.
[23] AICPB, December 2025 agent rankings.
[24] AICPB, January 2026 agent rankings.
[25] AICPB, February 2026 agent rankings.
[26] AICPB, March 2026 agent rankings.
[27] AICPB, April 2026 agent rankings.
[28] AICPB, May 2026 agent rankings.
[29] AICPB, June 2026 agent rankings.
[30] AICPB, July 2026 agent rankings.
[31] AICPB, August 2026 agent rankings.
[32] AICPB, September 2026 agent rankings.
[33] A01, public GAIA benchmark data.
[34] A02, SWE-bench leaderboard data.
[35] A03, OSWorld Verified benchmark data.

[36] A04, public data for web workflow evaluations.
[37] A05, official submissions and leaderboard for τ²-bench.
[38] A06, METR task time-horizon data.
[39] A07, AgentDojo evaluation results.
[40] A08, Epoch public benchmark data (A08 and A09).
[41] A10, Stanford HELM Safety.
[42] A11, Vectara summarization factual-consistency leaderboard.
[43] C02, customer disclosures on production use of Agentforce.
[44] C04, public search of open-source GitHub projects.
[45] C05, GitHub developer participation records.
[46] C06, GitHub development and maintenance records.
[47] C07, public search of research output on arXiv.
[48] C09, Google Trends search data for “AI agent”.
[49] D08, public API samples from the Anthropic Economic Index.

Chapter 7: Outlook—From the Agent Economy to an Era of Trust Capital

Artificial intelligence is undergoing a major transition: machines are becoming a new productive force within the economic system. Yet technical capability is not the same as economic capability. Even if an agent can perform complex tasks, it will struggle to qualify for sustained participation in commercial transactions unless it can demonstrate its identity, source of authorization, reliability in execution, and allocation of responsibility.

This is KYA's historic opportunity. It concerns not only whether agents can operate safely, but also whether the productive capabilities created by AI can be converted into real economic value.

The agent economy will evolve from capability formation to economic collaboration. The agent industry will move from competition over technical products toward competition over economic infrastructure. KYA may help create and accumulate a new economic resource: agent trust capital.

7.1 Four Stages of the Agent Economy: From Task Execution to Economic Value Creation

The agent economy will not form automatically just because AI models are powerful enough. Different thresholds separate model capability, task execution, market transactions, and economic organization. From the perspective of technological capability and economic role, the agent economy can be divided into four stages.

Stage 1: Capability Agents—from Human Tools to Intelligent Assistants.

The central goal of this stage is to give machines capabilities in understanding, analysis, reasoning, and content generation. Agents can help people retrieve information, write text, produce code, analyze data, and support decisions, but remain largely within workflows directly operated and supervised by humans. Their economic value is mainly reflected in higher individual productivity and lower costs for knowledge work and information processing.

At this stage, competition centers on model capability, compute, algorithms, and user experience. Businesses primarily sell model services, software tools, and computing resources. The stage is fundamentally about developing and disseminating AI capabilities.

Stage 2: Execution Agents—from Intelligent Assistants to Digital Labor.

As tool calling, task planning, workflow orchestration, and multi-agent collaboration advance, agents begin to perform complex tasks with defined objectives independently.

At this stage, agents shift from assistive tools toward execution capabilities within production processes. Businesses begin deploying agents at scale, creating new digital labor systems.

Competition expands from model capability to task-completion rates, operating costs, system reliability, and business value. Agent factories, enterprise-agent platforms, and organizational-management systems begin to emerge as important industry infrastructure.

However, Stage 2 also exposes a fundamental problem: an agent's ability to execute a task does not mean the market is willing to trust it or accept it as acting on behalf of a business.

As of September 2026, the agent economy as a whole is accelerating from Stage 1 toward Stage 2, while some leading-edge use cases have begun exploring Stage 3.

Stage 3: Transaction Agents—from Executing Tasks to Participating in Market Transactions.

This is one of the most important turning points for the agent economy. At this stage, agents do more than perform internal business tasks; they can also collaborate commercially across different businesses, platforms, and service providers.

A procurement agent may identify suppliers, compare quotations, and place orders; a logistics agent may purchase transportation services; a financial agent may make payments, settle transactions, and manage risk within its authorization; and agents with different specialties may complete complex projects together through service transactions.

Agents thus begin to participate in market transactions as executing parties. They need not possess independent legal personality to conduct economic activity on behalf of people or businesses under clear authorization, contracts, and responsibility arrangements. This shift may give rise to new markets for agent services, machine-to-machine transaction networks, and more finely divided digital specialization.

However, transactions across parties must rest on trustworthy mechanisms. Counterparties need to know whom an agent represents, what authority it has received, how well it performs, and how responsibility will be determined if a loss occurs.

KYA is therefore one of the key infrastructures enabling agents to move from internal execution tools to participants in market transactions.

Stage 4: Economic Agents—from Participating in Transactions to Organizing Economic Activity on an Ongoing Basis.

When agents can continuously perform tasks, transact across platforms, and collaborate in complex environments, the agent economy may enter a deeper stage of development. Agents no longer simply complete isolated tasks; they participate in business operations, resource allocation, supply-chain coordination, and the functioning of industry networks.

A business may have a small number of human managers coordinate large numbers of specialized agents to conduct operations. Agents from different businesses may form sustained economic relationships around orders, data, funds, and production resources. Physical systems such as energy equipment, industrial robots, and autonomous vehicles may also participate in resource allocation and commercial transactions through agents.

The boundaries of economic organizations, the minimum efficient scale of businesses, and forms of specialization may all change as a result.

The central issue at this stage is not only how much value agents can create, but how large numbers of agents can be organized to create value continuously under legal, market, and technical rules.

KYA will also expand from identity and behavior management for individual agents toward cross-platform trust, dynamic credit assessment, risk pricing, and responsibility governance, becoming an important component of the institutional system for the agent economy.

The four stages will not necessarily occur in a strict sequence, and may coexist for a long time. Their direction is nevertheless clear: machines first gain capabilities, then execute tasks, participate in transactions, and ultimately organize economic activity together with people.

This process may expand the market boundaries of the agent economy from AI software and computing services into business operations, commercial transactions, industry collaboration, and the operation of the real economy.

The internet expanded the reach of human access to information and markets. The agent economy may further expand the range of economic activity that people can effectively organize, delegate, and execute.

How far this process can go will depend not only on AI technology, but also on the breadth of machine-trust mechanisms society can establish.

7.2 Era-Defining Opportunities: From Building Agents to Building the Agent Economy

Major technological revolutions have often created commercial ecosystems far broader than the technology products themselves.

The initial opportunity of the Industrial Revolution was to manufacture machines, but what ultimately transformed the economy was the modern industrial production system organized around machines. The value of electrification came not only from electric motors but also from the restructuring of factory organization. The computer and internet revolutions created not only hardware and software industries but also enterprise-management systems, cloud computing, e-commerce, and global digital-payment networks.

The agent revolution may follow a similar pattern.

As trust infrastructure such as KYA matures, the boundaries of agents' economic activity may continue to expand. Industry competition may shift from building agents to organizing, trusting, and connecting them.

For leading companies in the agent ecosystem and a new generation of agent entrepreneurs, at least four era-defining commercial opportunities are emerging.

7.2.1 Producing Agents: From Software Factories to Digital-Labor Factories

The Industrial Revolution enabled the mass replication of machine production capacity; the agent revolution may enable the large-scale supply of digital labor.

In the future, businesses may purchase not simply software features but digital labor capable of completing specific tasks to defined quality standards and providing sustained service.

Specialized agents may take on work in finance, procurement, research and development, marketing, customer service, risk management, and other areas. Agent factories may therefore evolve from providers of development tools into platforms that produce, certify, deploy, and operate digital labor.

Business models may gradually extend from software licenses, model calls, and project delivery toward fees based on tasks, digital roles, or actual business outcomes.

KYA provides identity, capability verification, authorization management, and reliability safeguards in this process, enabling businesses to adopt digital labor more broadly.

The essence of this opportunity is a shift from selling software tools to selling productive capacity.

7.2.2 Organizing Agents: From Enterprise-Management Software to an Operating System for the Agent Economy

Electrification unlocked its full productivity potential not merely because factories replaced machines with electric motors, but because machine layouts, production processes, and organizational structures changed.

In 1990, economist and economic historian Paul A. David used electrification in "The Dynamo and the Computer: An Historical Perspective on the Modern Productivity Paradox" to show that decades may pass between the invention of a general-purpose technology, its widespread diffusion, and measurable productivity gains. Realizing its potential also requires complementary technological innovations and adjustments to factory layouts, production processes, and organizational structures.

This illustrates the potentially long lag between general-purpose technological innovation and productivity gains, as well as the importance of organizational change.

The agent economy may follow a similar pattern. Improving the capabilities of individual agents will not necessarily raise overall enterprise efficiency. Organizing many agents to work together will become a new management challenge.

A business may have thousands of agents performing different roles and business tasks. Task allocation, resource scheduling, permission management, performance evaluation, risk control, and cross-system collaboration will all require new infrastructure.

This may give rise to an operating system for agent organizations that goes beyond traditional workflow software. Its value lies not only in managing agents but also in continuously improving the organization's productivity and enabling businesses to redesign processes around digital labor.

The essence of this opportunity is a shift from managing software and workflows to organizing and allocating productive capacity.

7.2.3 Trust for Agents: From KYA Certification to Agent-Credit Infrastructure

The history of modern finance demonstrates that credit infrastructure can create substantial economic value.

Credit reporting, credit ratings, payment authorization, and risk-management systems make large-scale financial transactions between unfamiliar parties possible. FICO, Moody's, and S&P have shown that standardized, specialized services for identity and credit information can support a durable commercial market, with an influence extending far beyond these institutions' own revenues.

The agent economy may generate similar demand. As agents perform tasks across businesses and platforms, identity recognition, authorization verification, behavioral monitoring, and performance evaluation may evolve from internal business-management functions into shared requirements for market transactions.

KYA could create a cross-entity trust-services network that enables agents to participate in economic activity with portable trusted identities, valid authorization, and necessary evidence of performance.

More profoundly, as verifiable behavioral records accumulate, agent reliability may become a trust resource with sustained economic value, informing transaction access, risk management, and differentiated pricing. KYA service providers may thereby evolve from security-technology vendors into providers of trust infrastructure for the agent economy.

The essence of this opportunity is a shift from reducing transaction risk to cultivating trust capital and increasing the economic value of AI capital.

7.2.4 Connecting Agents: From Industry Digitization Platforms to Agent-Economy Networks

The development of the internet and card networks offers another important lesson: a platform that connects economic activity does not need to produce every good or service itself to become important infrastructure for large-scale commercial transactions.

Payment networks such as Visa and Mastercard support trillions of dollars in transactions each year, while network operators earn sustained revenue by providing connectivity, rules, technology, and risk management.

In the future, agents may form extensive networks for services and transactions.

Procurement agents may purchase logistics services, marketing agents may procure content-production services, energy agents may coordinate energy-storage resources, and industrial equipment may order maintenance through agents. Agents offered by different businesses and platforms could collaborate under commonly recognized rules.

This may create new infrastructure for agent service discovery, transaction matching, payment and settlement, performance assurance, and risk management.

Like providers of AI technology services, these companies may move beyond supplying digital products to businesses and provide ongoing services for agent-economy activity across businesses. Their growth may depend not only on how much software they sell or how many agents they produce, but also on the number of agents they serve, the real economic tasks those agents perform, and the scale of cross-platform transactions.

The essence of this opportunity is a shift from serving the digital transformation of businesses to supporting the operation of agent-economy networks.

Together, the four opportunities form a new industry value chain: agents for production create productive capacity; agents for organization allocate it; trust for agents makes it credible for market participation; and agent connectivity expands the scope of its economic activity.

For businesses, these represent four potential markets. For the economy as a whole, they may form a mutually reinforcing set of new infrastructure.

The most competitive businesses of the future may not be those with the largest number of agents, but those able to connect the greatest number of trustworthy agents to real economic activity at reasonable cost.

The first commercial opportunity of the agent revolution is to build agents; the deeper opportunity is to build the agent economy.

7.3 Trust Capital: Capital Formation and Economic Growth in the AI Era

The four-stage evolution reveals the direction of the agent economy, while the four commercial opportunities show the new industrial value at stake. KYA's economic significance can be taken further.

A fundamental question is why two agents with similar technical capabilities may have sharply different economic value.

Suppose two agents have the same model capability, operating efficiency, and task-completion performance. One can demonstrate its identity and source of

authorization, has a sustained record of reliable performance, and operates with robust governance and clear responsibility arrangements. The other lacks these conditions.

For a business delegating funds, data, or important operating tasks, their practical value is clearly different.

This difference does not arise entirely from technical capability; it comes from the market's assessment of their trustworthiness.

This leads us to propose an economic concept worthy of further research: the agent economy may give rise to a new productive resource—Agent Trust Capital.

7.3.1 Why Can Trust Become Capital?

The economic value of trust is not a new discovery.

From Coase's theory of transaction costs and North's institutional economics to empirical research on social capital and financial development, economists have long examined how trust, institutions, and mechanisms for cooperation affect economic efficiency.

In 1997, using data from 29 market economies, Stephen Knack and Philip Keefer found that interpersonal trust and civic cooperation norms were associated with economic growth and investment performance, and discussed how trust might reduce the costs of safeguards and contract performance in economic relationships. In 2004, using regional and household data from Italy, Luigi Guiso, Paola Sapienza, and Luigi Zingales found that households in regions with greater social capital were more likely to use checks, held more of their wealth in stocks, and had better access to institutional credit. These studies—at the cross-country macroeconomic level and in household financial choices—show that trust and related social norms are closely associated with economic activity, investment allocation, and access to financial contracts.

Trust in human societies can accumulate through reputation, performance, institutions, and long-term cooperation. Machine economies also need trust mechanisms that support sustained transactions, but their foundations differ in important ways from those of human society.

Agents cannot gain market trust solely through social morality and traditional interpersonal relationships. They need verifiable identity, clear authorization, reliable behavioral records, effective governance, and responsibility safeguards provided by the legal entities behind them.

These elements can form gradually through sustained investment, actual performance, and institutional development, influencing an agent's future economic opportunities.

We therefore define Agent Trust Capital as follows:

Agent Trust Capital is an intangible form of capital created through investment in trustworthy governance and an agent's sustained, verifiable performance. It can

reduce transaction costs, expand the scope of economic activity, improve the efficiency of capital use, and generate future economic returns on an ongoing basis.

The essence of Agent Trust Capital is converting machines' trustworthy capabilities into capital that can generate economic returns on an ongoing basis. Trust capital can accumulate, but it can also depreciate through violations, breaches of trust, security incidents, invalid authorization, or technological change.

This concept still needs to be tested through theoretical and empirical research, but it offers a new perspective for understanding KYA's economic value.

7.3.2 How Can KYA Promote the Formation of Agent Trust Capital?

KYA itself is not Agent Trust Capital. KYA is a set of mechanisms for producing, verifying, and governing trust; trust capital is an economic resource formed through trustworthy behavior and institutional arrangements over time.

Through KYA, an agent's identity, authorization, capabilities, operating status, and performance records may be verified continuously. As trustworthy records accumulate, counterparties can reduce repeated reviews, businesses can broaden the scope of agent tasks within acceptable risk limits, and financial and insurance institutions may obtain more reliable risk information.

Trust can thus become more than a subjective judgment in a single transaction; it may gradually form a resource that can continuously support economic activity.

If large numbers of agents can carry verifiable identity, authorization, and performance credentials across platforms, trust capital may extend beyond the boundaries of individual businesses and become a foundation for broader market collaboration.

Through this process, KYA may evolve from an identity-verification and security-management tool into an important component of the agent-credit system.

7.3.3 How Can Agent Trust Capital Affect Capital Formation and Value?

In the AI era, capital accumulation is first reflected in sustained investment in compute, algorithms, data, software, and related infrastructure. But technology capital does not automatically generate economic returns. Technical capability becomes effective output only when it enters real production, services, and transactions.

By increasing the level of trusted authorization and market acceptance for agents, KYA may enable some economic tasks that could not previously be performed. This may have at least three effects.

First, activate technology capital by bringing existing AI capabilities into more high-value business scenarios and increasing the effective use of technology capital.

Second, improve returns on capital by reducing verification costs, transaction risk, and governance losses, improving the risk-adjusted returns of AI projects.

Third, promote capital formation. When businesses can control agent behavioral risk more reliably and better forecast operating returns, some investments that were previously difficult to pursue may become commercially viable and attract new capital.

From this perspective, trust capital and technology capital may be important complements. Technology capital determines what capabilities an agent can provide; trust capital influences how far markets will accept, authorize, and use those capabilities. AI creates potential productive capacity, and trust capital may improve the efficiency with which that potential becomes actual economic value.

7.3.4 From Capital Value to Economic Growth and Social Welfare

The value of Agent Trust Capital will extend beyond the investment returns of individual businesses. If more agents can participate reliably in production and transactions, broader changes may occur across the economic system.

First, businesses can delegate more repetitive and standardized tasks to agents, improve coordination between labor and capital, and raise productivity.

Second, trustworthy cross-platform collaboration can reduce transaction costs for specialized services, expand market reach, and support more detailed specialization across industries.

Third, small and medium-sized businesses and entrepreneurs can obtain specialized digital labor at lower cost, reduce fixed organizational costs, and create new products, services, and businesses.

In addition, emerging markets in identity verification, behavioral auditing, credit assessment, risk management, and agent operations may create new industry demand and jobs.

These changes may support economic growth, but broader agent adoption does not necessarily imply a net increase in employment. AI may also replace some jobs, change wage structures, and redistribute productivity gains between capital and labor.

The socioeconomic value of KYA therefore depends not only on whether it raises returns on capital, but also on whether it promotes fair competition, lowers market-entry costs for small and medium-sized businesses, reduces transaction losses, and enables the benefits of technological progress to be shared more broadly.

Assessing KYA's value requires looking beyond the revenue generated by the KYA industry itself. It should also consider how much productive capacity KYA helps the wider economy unlock, how much cost it reduces, and how much net output and social welfare it creates.

7.3.5 From Technological Revolution to Institutional Innovation

History shows that major technological revolutions do not stop at the level of production tools. The Industrial Revolution created mechanized production and drove the development of modern factories, business organizations, and financial institutions. The internet changed the way information is transmitted and also gave rise to e-commerce, digital identity, online payments, and platform governance.

Every general-purpose technology revolution requires corresponding institutional and organizational innovation to realize its full economic potential. The agent economy is no exception. AI is giving machines increasingly powerful autonomy in execution, but the economy needs more than machines that can work; it needs machines that can be identified, authorized, trusted, and held accountable.

KYA's historical value lies in establishing trustworthy conditions for this new productive force to engage in economic activity. It is both a technology-governance challenge and a question of institutional innovation; it concerns both the formation of commercial markets and improvements in capital accumulation, resource allocation, and social welfare.

Over the long term, KYA may help trust evolve from fragmented mechanisms for identity authentication, reputation assessment, and risk control into an economic resource that can be continuously verified, accumulated, and used.

Whether this process produces a mature, independent system of Agent Trust Capital will depend on technological reliability, open standards, legal institutions, and market practice. Its direction already points to an important proposition: economic competition in the AI era will not only be competition among models, computing power, and algorithms, but will increasingly concern the trustworthiness of productive capacity, organizational efficiency, and the scope of economic collaboration.

If AI gives machines the ability to create value, the trust system represented by KYA will influence how far machines are allowed, accepted, and continuously delegated to create that value. The future of the agent economy depends not only on how intelligent machines become but also on how broadly human society can establish trust in machines.

From the formation of agent capabilities, through the expansion of agent-economy networks, to the accumulation of trust capital, KYA may become an important bridge between technological progress and economic growth. This is the era-defining value of KYA as it moves beyond safety governance toward economic infrastructure.

7.4 How to Create and Accumulate Agent Trust Capital

Internet platforms, foundational-model companies, cloud-computing providers, and AI-infrastructure companies are clearly well placed to create and accumulate Agent Trust Capital. Internet platforms have user relationships, digital identities, and service ecosystems, giving them the opportunity to extend existing user trust into trustworthy authorization relationships between people and agents. Foundational-model companies control the technological source of agent capabilities and may turn model performance, safety, and reliability into verifiable proofs of capability. Cloud and AI-infrastructure companies control agent deployment and runtime environments and may become important infrastructure providers for machine identity, trusted execution, and behavioral governance.

However, substantial opportunities to accumulate Agent Trust Capital are also available to businesses and individuals more broadly. This report's research reveals an important trend: Agent Trust Capital does not belong to a particular industry and will

not be monopolized by a single technology or commercial platform. Businesses of different types can participate in the formation, accumulation, and realization of trust capital through their distinct resources and capabilities. Five categories of businesses warrant particular attention.

First: Businesses that control platforms for agent development, operations, and organization.

Agent platforms can produce digital labor and may also become important infrastructure for scaling the formation of trust capital. Agent factories and organizational platforms such as those of Ant Digital Technologies can integrate models, tools, industry knowledge, and business processes into agent systems that operate continuously.

Their distinctive advantage lies in embedding trusted identity, authorization management, capability testing, behavioral auditing, and risk control across the agent lifecycle.

When agents have continuous identities and trust records from development through deployment, operation, and upgrades, trust capital may evolve from experience in an individual project into an economic resource that can accumulate and be reused.

Agent platforms should therefore treat KYA as a core product capability, not an add-on security feature. They should also provide open interfaces and standardized credentials so that third-party agents can participate in trustworthy collaboration. Their strategic opportunity is to evolve from building agents into providing infrastructure for producing, organizing, and governing trustworthy digital productive capacity.

Second: Businesses that control payments, transactions, and commercial networks.

The economic value of trust capital is ultimately realized in real transactions. Payment and commercial-infrastructure companies such as Stripe can connect merchants, consumers, payment institutions, and transaction networks. As agents begin to make purchases, payments, and service transactions on behalf of people and businesses, identity verification, valid authorization, payment controls, fraud detection, and accountability traceability become important operating conditions for commercial networks.

These companies' advantage is their ability to link trusted verification directly with transaction access, payment authorization, and risk management, making trust a practical mechanism for enabling commercial activity.

In the future, payment and transaction platforms may offer trusted transaction services for agents, dynamic risk pricing, and cross-platform performance guarantees. Their commercial value may come not only from verification services but also from ongoing demand created by the expansion of trustworthy transactions. Their strategic opportunity is to evolve from connecting funds and transactions to operating commercial networks that connect trustworthy agents with economic activity.

Third: Businesses with real industry scenarios and sustained performance data.

Real, sustained, and verifiable economic activity is the most important source of trust capital. Finance, energy, education, manufacturing, healthcare, logistics, supply chains, and professional services are becoming important settings for agents to enter real production activity. Businesses in these sectors can continuously generate records of actual task execution, transaction performance, service quality, and risk.

Examples include energy-agent applications such as those of Linyang Energy, educational agents developed by Yinghe Zhilin, and professional medical-service scenarios explored by Mediwelcome and other businesses. The value of an industry-specific agent depends on whether it can perform tasks reliably over time in complex business environments.

These businesses should develop agent applications alongside the accumulation of trustworthy data, turning records of task completion, service quality, authorization compliance, and anomaly handling into verifiable evidence of performance. Those that first convert industry expertise and real performance data into continuously accumulating trust records may gain an advantage that general-purpose models cannot easily replicate.

Fourth: Businesses with capabilities in credit assessment, risk management, and professional certification.

Traditional credit and risk-management capabilities may find new market opportunities in the agent economy if they continue to evolve alongside new technologies. Credit-technology and risk-management organizations such as Sesame Credit and IceKredit have experience in identity recognition, data analysis, risk modeling, and credit assessment.

The agent economy will create new assessment subjects and risk dimensions. Future credit systems will need not only to identify the responsible parties behind agents, but also to assess task reliability, authorization compliance, behavioral stability, and historical performance. This creates an opportunity for credit and risk-management businesses to extend from services for human and corporate credit to trustworthy-agent assessment and dynamic risk services.

As discussed earlier in this report, agent credit cannot simply copy the traditional model of personal credit reporting. Differences in the risks of different tasks, model updates, changes in identity, and links to responsible parties all require new assessment methods.

These businesses can focus on scenario-based credit assessment, continuous risk monitoring, independent capability certification, and trusted-evidence services, while exploring cross-platform recognition of credit. Their strategic opportunity is to move from assessing the credit of traditional economic actors to helping build credit infrastructure for the agent economy.

Fifth: Businesses with specialized knowledge, industry standards, and independent governance capabilities.

The more complex the agent economy becomes, the more important specialized and independently trustworthy services will be. Agents in education, healthcare, law, finance, research, and other professional domains must meet quality standards and responsibility requirements different from those for general tasks. The direction represented by specialist service innovators such as GCSA suggests that deep expertise, verifiable service quality, and clear responsibility mechanisms may become important sources of trust capital.

At the same time, independent testing organizations, security-audit firms, digital-identity providers, insurers, industry alliances, and standards bodies may become important contributors to the trust-capital system. They can provide third-party capability assessments, identity and authorization credentials, risk protection, and dispute-resolution mechanisms, reducing the cost of trust between platforms.

These businesses and organizations should develop service systems around professional standards, independent verification, open recognition, and risk protection. Their strategic opportunity is to turn specialist knowledge and governance capabilities into trust services needed throughout the agent economy.

The five categories of businesses described above hold different trust resources. Complementing internet platforms, foundational-model companies, and AI-infrastructure businesses, they can activate agents across industries and jointly form a system for creating, accumulating, verifying, and realizing the value of Agent Trust Capital.

Based on this research, we offer four recommendations to businesses:

First, make trustworthy capabilities part of core competitive strategy. Businesses using agents should assess not only model performance, operating cost, and application volume, but also incorporate identity trustworthiness, authorization controls, performance, and risk governance into product design and operating metrics. Trust capital must accumulate through reliable economic activity over time; it cannot be obtained through brand promotion or one-time certification alone.

Second, turn actual performance into verifiable, reusable credit resources. Subject to legal and regulatory requirements, businesses should establish records of agent operations and performance and promote trusted credentials, dynamic assessment, and cross-platform recognition, allowing past reliability to reduce the cost of future transactions.

Third, test KYA's commercial value through high-value economic activity. KYA should go beyond identity registration and compliance checks and be connected to real business in procurement, payments, industry operations, professional services, and other areas. Businesses should measure the additional transaction opportunities, reduction in risk losses, expansion of delegated tasks, and improvement in returns on capital enabled by KYA, rather than pursuing certification counts alone.

Fourth, build a trust ecosystem through open collaboration. Agent developers, industry businesses, financial institutions, credit-service providers, and third-party certification bodies should jointly advance the standardization and interoperability of identity, authorization, behavioral, and credit information. While protecting privacy,

preserving fair competition, and clarifying responsibility, they should lower the barriers to agents participating in economic activity across platforms and businesses.

Finally, the "business community" referred to here now includes not only traditional companies but also the rapidly growing "one-person companies" and individual operators, for whom agents may be employees or even partners. The businesses with the greatest trust-capital advantage in the future may not be those with the longest histories, the largest scale, or the most capable models, but those that continuously generate trustworthy behavior, accumulate reliable records, promote trust across parties, and turn trust into actual economic value.

KYA means not only new governance responsibilities, but also new commercial opportunities.

References

[1] McKinsey & Company. The automation curve in agentic commerce. 2026; see also *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; see also *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 attack disclosure. 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 and Verifiable Credentials specifications.

[39] Quentin Dupont and Jonathan M. Karpoff. A Theory of Social Capital and Trust. 2025-10.

[40] McKinsey. Seizing the Agentic AI Advantage. 2025-06.

[41] Gartner. 2026 Hype Cycle for Agentic AI. 2026-04.

[42] Gartner. Hype Cycle for Digital Identity, 2026. 2026-07.

[43] Palo Alto Networks. 2026 Identity Security Landscape. 2026-05.

Appendix : Agent Credit System (KYA) Survey

Abstract

This report draws on 109 valid survey responses collected globally and provides a systematic analysis of the industry landscape for Agent credit systems (KYA). The core findings are as follows. Trust friction has become routine: 76% of respondents encountered trust, security, or accountability concerns in the past year, and 23% have already been refused by customers or partners. The absence of trust is shifting from a hidden efficiency tax into direct market contraction. Industry maturity is spindle-shaped: L3 hybrid credit accounts for 38%, while L4 and above remain below 15%. Most organizations have not yet moved from a human-backstop model to authorization based on records. Capability building is structurally unbalanced. On creditworthiness, records are strong and scoring is weak (decision logs at 81%, internal credit scores at only 52%). On authorization, backstops are strong and tiering is weak (human confirmation at 86%, scenario-based tiering at only 52%). 39% of respondents see authorization trustworthiness as the weakest of the three dimensions, and only 17% have a defined accountability process. Demand-side consensus is strong: 88% call for unified assessment standards, and 60% are willing to pilot or proactively adopt certification services. In BNPL, Agents already sit inside core underwriting workflows; efficiency gains are clear, but consumer trust still needs to be built. The SME market is undersupplied. Low-cost identity, standardized authorization templates, and platform-based record infrastructure are the keys to a breakthrough. The report sets out staged action recommendations on this basis.

1. Executive Summary

- 113 responses were collected; the analytic sample is 109 after removing 1 test submission, 2 earlier duplicate submissions with the same Prolific ID, and 1 submission that failed the attention checks.
- Trust, security, accountability, and compliance issues already affect projects on a regular basis: 83/109 respondents (76.1%) selected Frequently or Occasionally.
- There is a strong consensus on the value of an Agent credit system: 80/109 (73.4%) rated it Extremely or Quite important, and 96/109 (88.1%) said unified industry standards are Extremely or Quite needed.
- The most salient gap is authorization trustworthiness: 42/109 (38.5%) selected it directly, and a further 19 respondents said all three dimensions are relatively weak.
- Maturity remains largely human–machine hybrid: 56/109 (51.4%) are at L3 or above, but L1 and L2 still account for a substantial share, indicating that most scenarios have not yet reached independent Agent credit.
- Market adoption is more likely to start with high-risk pilots: 66/109 (60.6%) would adopt proactively or first pilot in high-risk or critical scenarios.

2. Data and Methods

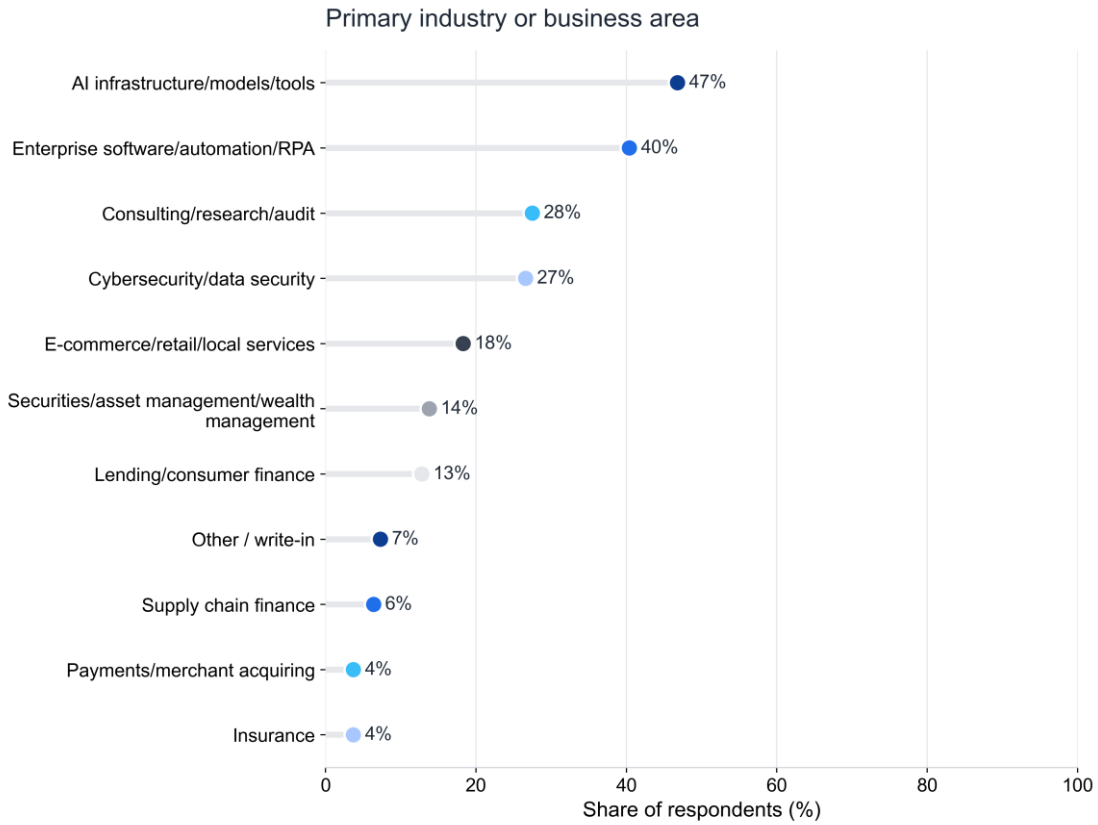
- Data source: survey results exported from Google Forms to Excel.
- Cleaning rules: remove test submissions; keep only the latest submission for each Prolific ID; remove any submission that failed either attention check.
- Statistical conventions: single-choice shares use non-empty answers as the denominator; multiple-choice shares count how often each option was selected and may sum to more than 100%; matrix and scale items are summarized item by item as exported.
- Privacy: contact fields such as Name, Organization, Position/function, Email, and Phone/WeChat do not appear in the body or charts. Only the anonymous willingness to be featured as a case is tabulated.

3. Data Analysis

3.1 Sample Profile

This chapter assesses whether the sample covers product, technology, risk, compliance, operations, and research/consulting roles, and which Agent scenarios respondents know best.

Q1. Primary industry or business area



Multiple selections were allowed; percentages use respondents with a non-empty answer as the denominator.

Figure 1. Top multiple-choice items: "AI infrastructure/models/tools" 46.8%; "Enterprise software/automation/RPA" 40.4%; "Consulting/research/audit" 27.5%. Percentages may sum to more than 100% because multiple selections were allowed.

Interpretation: The sample is mainly supply-side professionals: AI infrastructure/models/tools account for 47%, enterprise software/automation/RPA for 40%, and consulting/research and cybersecurity for 28% and 27% respectively. These respondents both build Agents and carry governance duties, so their answers reflect an engineering view of trust infrastructure more than end-user perception. For KYA research, this structure is well suited to the question of how the system should be built: implementation paths for identity, credit, and authorization, the collectability of assessment metrics, and the engineering cost of authorization all require a builder's perspective. Two boundaries still apply. First, under the principle in Section 5.1.4 of the main report, self-reported material cannot replace independent verification; these data are industry self-assessments and should be cross-checked against third-party testing and audit evidence. Second, demand-side trust perception and willingness to pay are only thinly covered here. The BNPL and SME chapters touch parts of the demand side, but measuring terminal-enterprise and consumer attitudes in full would require a dedicated sample.

Q2. Organization/team size

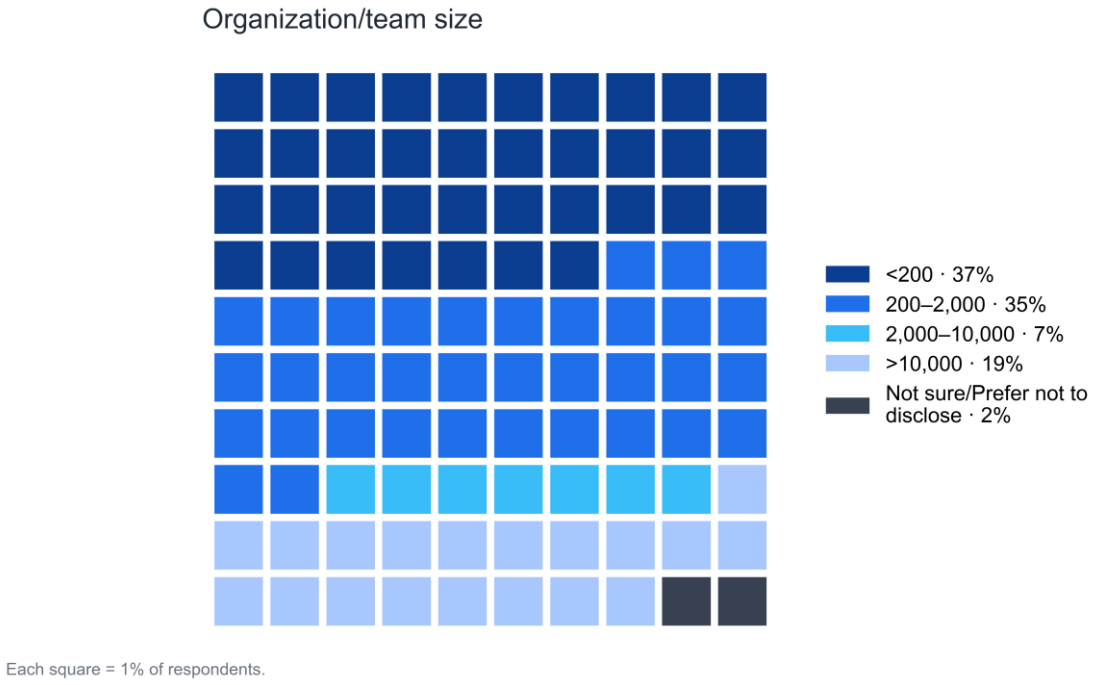


Figure 2. The leading option is "<200", at 36.7%.

Interpretation: Organization size is bipolar: teams under 200 people account for 37%, 200–2,000 for 35%, and organizations with more than 10,000 people for 19%. The sample therefore covers both resource-constrained smaller teams and large institutions with complex processes, which helps identify how a credit system should adapt under different organizational conditions. Drawing on Section 5.3.4 of the main report, smaller organizations are cost-sensitive and need a platform-based, low-barrier starting path: first meet the four minimum requirements of identity attribution, genuine fulfillment, authorization execution, and exception handling, and be able to answer which Agent, what task, what authorization, what outcome, and who intervened. Large organizations care more about cross-department accountability and audit connectivity, and sit closer to the four L4 pillars in Section 2.6: persistent identity, assumption of economic consequences, portable reputation, and enforceable default mechanisms. The two groups occupy different points on the same maturity spectrum. The framework should allow implementation depth to be scaled from one indicator set, rather than creating a separate standard.

Q3. Primary role or function

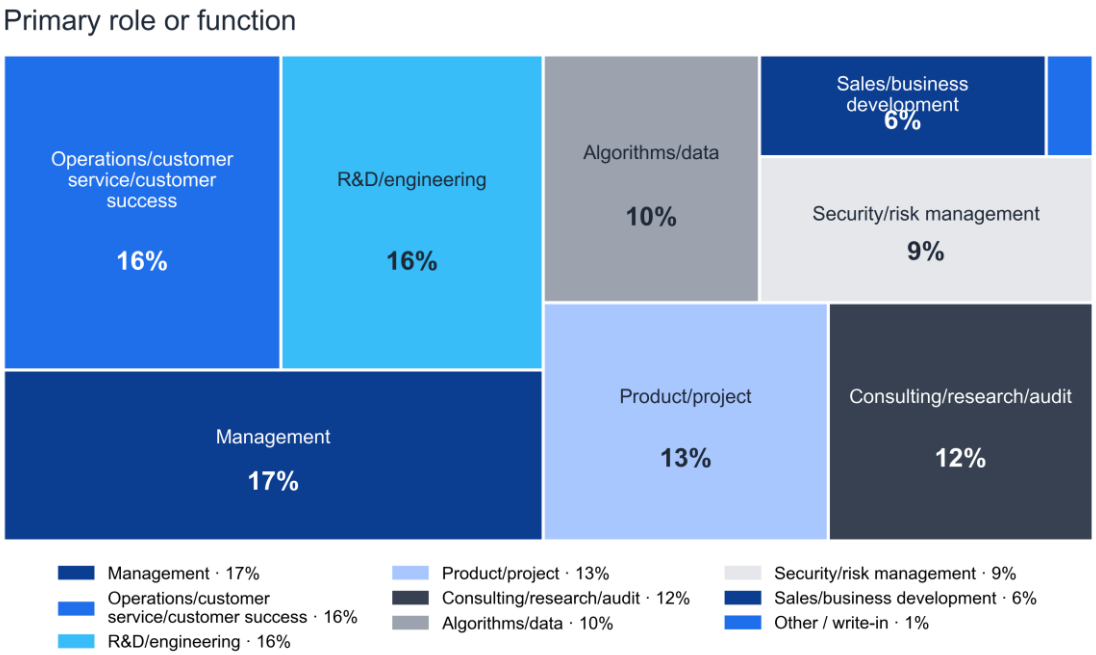


Figure 3. The leading option is "Management", at 17.4%.

Interpretation: Respondent roles cover management (17%), operations and customer service (17%), R&D (16%), product (13%), consulting/research (12%), algorithms (10%), and security/risk (9%), with no single function dominant. Agent trust therefore cuts across build, operate, and govern. No single department's view is enough to describe the risk. The high attention later given to accountability mechanisms and authorization boundaries appears in this cross-functional setting, which means the pain points are organization-wide rather than local to one job. Under the layered-responsibility design in Section 3.3.7 of the main report, assigning blame after an Agent incident already involves developers, deployers, operators, and supervisors, and requires evidence chains and advance compensation to allocate loss across roles. This sample's functional mix covers the main nodes on that chain, so answers on accountability and authorization reflect real tension in how responsibility is shared, not the demand of a single standpoint.

Q4. Relationship to AI Agent projects

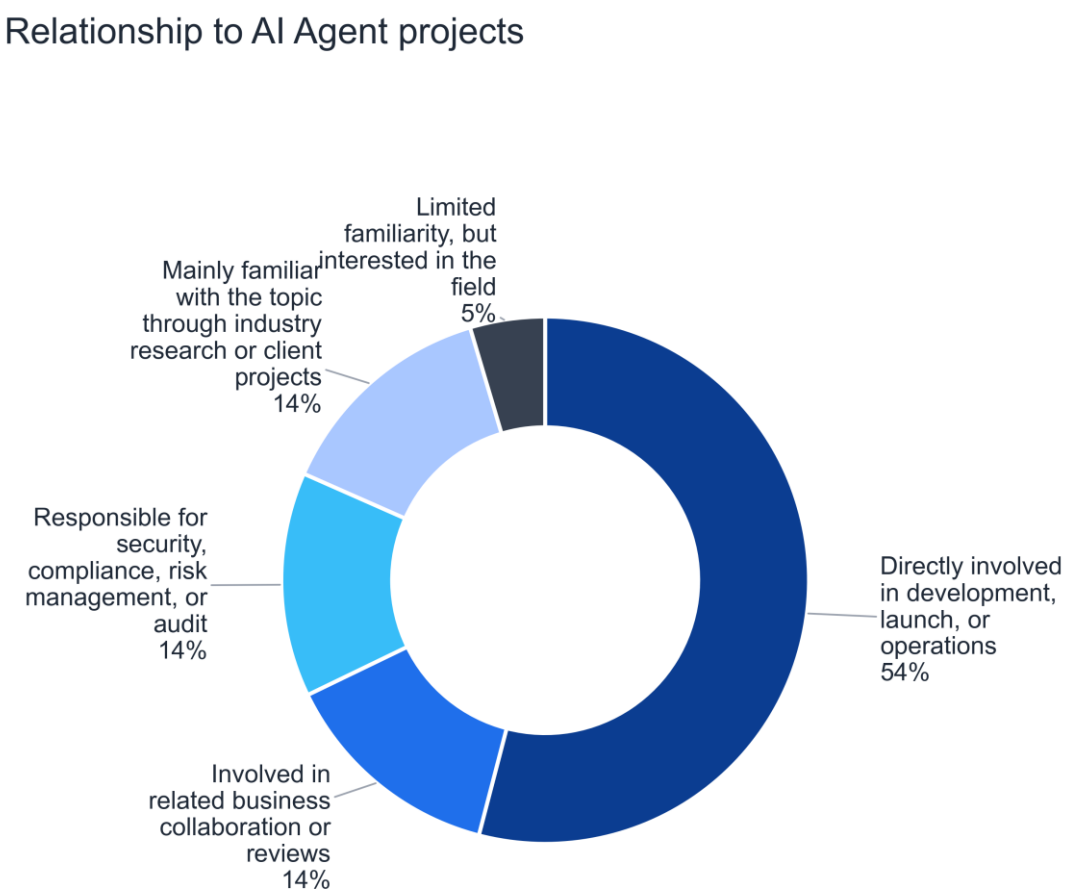


Figure 4. The leading option is "Directly involved in development, launch, or operations", at 54.1%.

Interpretation: 54% of respondents are directly involved in Agent development, launch, or operations. The rest sit mainly in adjacent functions such as business collaboration and review, security/compliance/risk, and industry research; only 5% report limited familiarity. The sample is practitioner-heavy, so the data capture trust friction that has already occurred in production, not peripheral judgments based on public reporting. That supplies front-line evidence for Chapter 1 on the cost of missing trust: the individual–enterprise–market transmission chain in Section 1.2.2 of the main report starts with the incidents, complaints, and human interventions these roles handle every day. The high practitioner share also means the problems captured are closer to "need to be solved" than to "might be imagined," which gives later pain-point rankings more action value. Methodological restraint still applies. Under Section 5.1.4 of the main report, self-reported data reflect the state respondents are willing to report; failure rates and unauthorized actions in live operations are usually higher. Findings should be cross-checked against independent evaluation.

Q5. Role in the Agent ecosystem

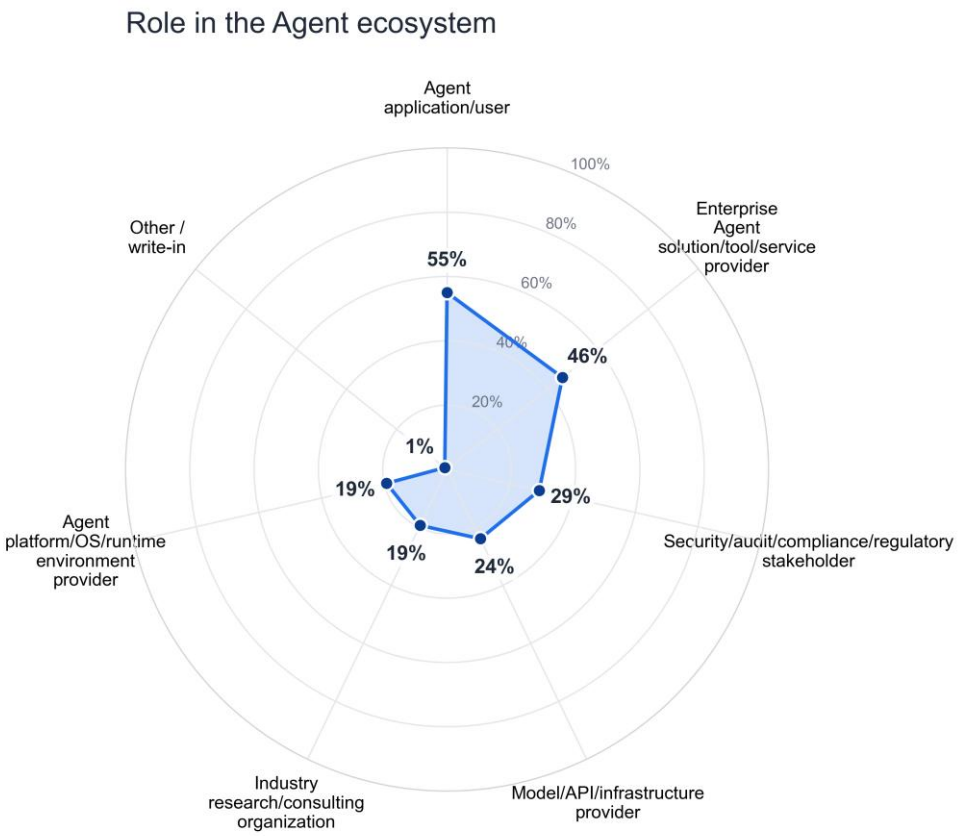
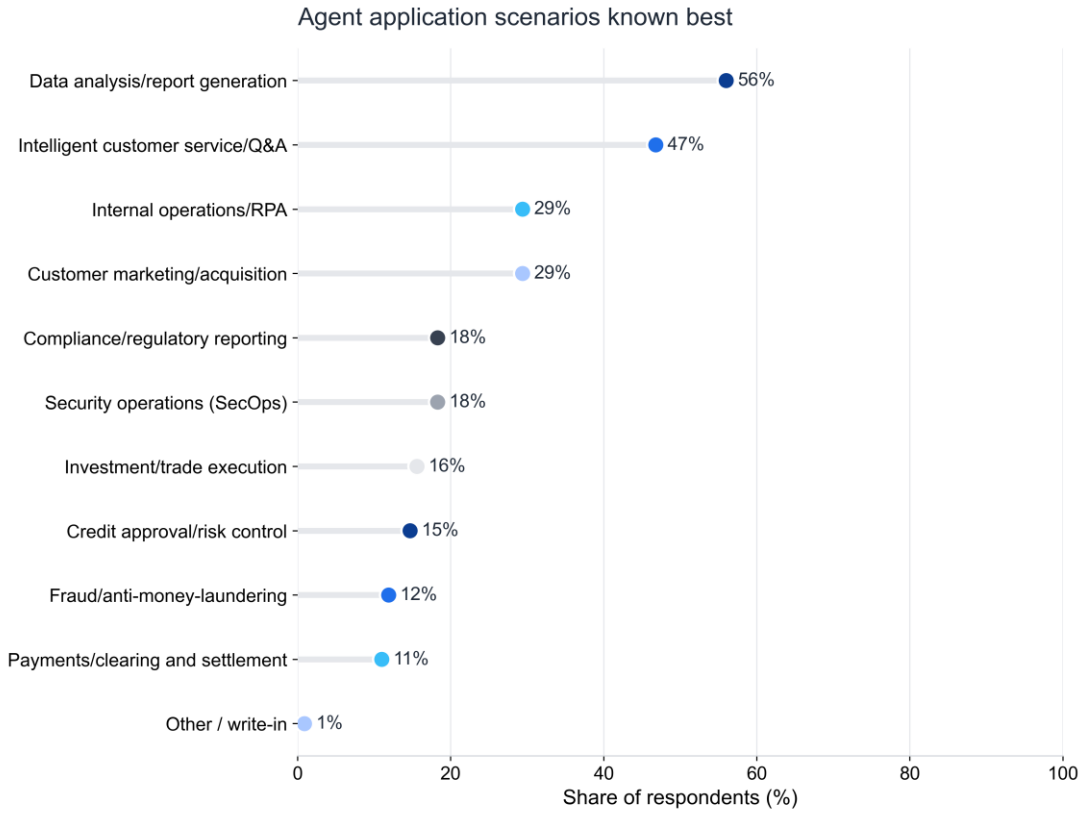


Figure 5. Top multiple-choice items: "Agent application/user" 55.0%; "Enterprise Agent solution/tool/service provider" 45.9%; "Security/audit/compliance/regulatory stakeholder" 29.4%. Percentages may sum to more than 100% because multiple selections were allowed.

Interpretation: 55% of respondents participate in the Agent ecosystem as application-side users, 46% are also solution or tool providers, and 29% carry security, audit, or compliance duties; overlapping identities are common. The crossing of supply and demand roles shows that industry boundaries have not yet hardened: many organizations are both users and builders of Agents. For KYA, a credit system must serve both self-built Agents and Agents brought in from outside. Identity attribution and authorization relationships cannot assume a single role. Concretely, the three identity questions in Section 3.1.2 of the main report—who is it, whom does it represent, and is it still the same Agent—are asked from both sides in these organizations. As a builder, one must answer who my Agent is and what it can do; as a user, one must verify whom the counterparty's Agent represents and whom to call if something goes wrong. That dual identity also means future credit-record circulation will run along existing supply–demand networks, which is a realistic basis for network effects: the same organizations are both producers and consumers of credit data.

Q6. Agent application scenarios known best



Multiple selections were allowed; percentages use respondents with a non-empty answer as the denominator.

Figure 6. Top multiple-choice items: "Data analysis/report generation" 56.0%; "Intelligent customer service/Q&A" 46.8%; "Internal operations/RPA" 29.4%. Percentages may sum to more than 100% because multiple selections were allowed.

Interpretation: The scenarios respondents know best are data analysis and report generation (56%) and intelligent customer service (47%), both low-risk and easy to roll back. Familiarity with money-related scenes such as credit approval (15%), anti-fraud (12%), and payments/clearing (11%) is markedly lower. That gradient matches the maturity logic in Chapter 2: low-risk scenes accumulate fulfillment records first, while high-value scenes penetrate slowly because trust is insufficient. The market-expansion effect of trust mechanisms shows up directly here: when verifiable evidence is lacking, the feasible set of transactions shrinks toward low-risk tasks and the addressable market in high-value scenes is suppressed. Read the other way, the distribution also points to a practical cold-start path. Under the four paths in Section 3.2.5 of the main report, sandbox tasks and low-risk admission scenes are natural vehicles for the first fulfillment samples. High familiarity with analytics and customer service means those scenes can already generate fulfillment data at scale; what they lack is a standardized way to turn records into verifiable credit evidence. Low penetration in money scenes suggests that portable, cross-scene verification of credit evidence should be in place before those scenes are entered.

3.2 Application Status and Trust Challenges

This chapter asks whether trust issues already affect deployment, which parts of the workflow they touch, and what the core pain points are along the business, compliance, and technology lines.

Q7. Frequency of trust/security/accountability/compliance concerns

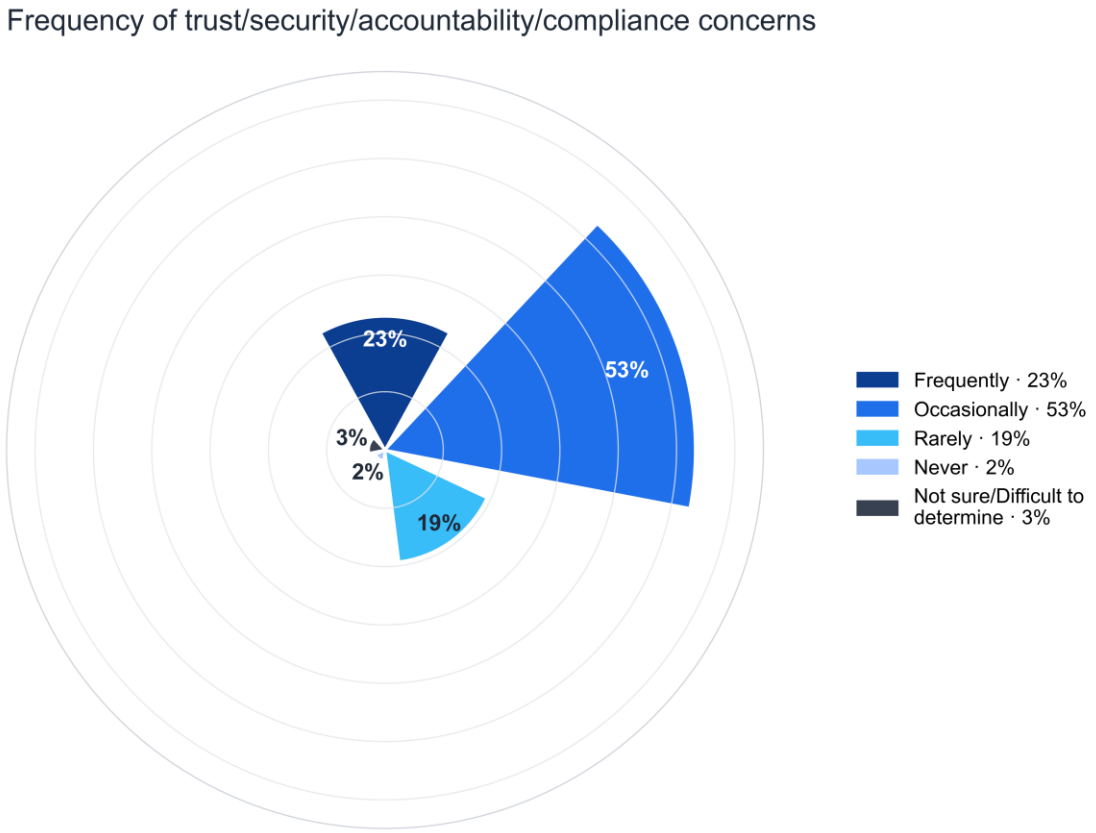


Figure 7. The leading option is "Occasionally", at 53.2%.

Interpretation: 76% of respondents encountered trust, security, accountability, or compliance concerns frequently (23%) or occasionally (53%) in the past year; only 2% never did. Trust friction is already routine. As Section 1.2 of the main report notes, Agent errors have speed, scale, and replicability: a single mistake can be amplified into systemic loss within seconds. When concern becomes a frequent experience, organizations instinctively shrink the authorization perimeter. That is the micro-mechanism by which a trust vacuum turns into economic cost. The 53% who "occasionally" encounter issues is itself informative: trust problems are not concentrated in a few poorly managed organizations; they are distributed across ordinary industry practice. That matches Section 1.2.1: missing trust is not the result of isolated accidents, but a structural consequence of Agents having no native credit and of the identity–behavior–responsibility chain remaining open. While the chain stays open, each concern fails to settle into reusable credit information, and organizations can only respond with ad hoc human review. That is the root of the efficiency costs that accumulate in Q8 below.

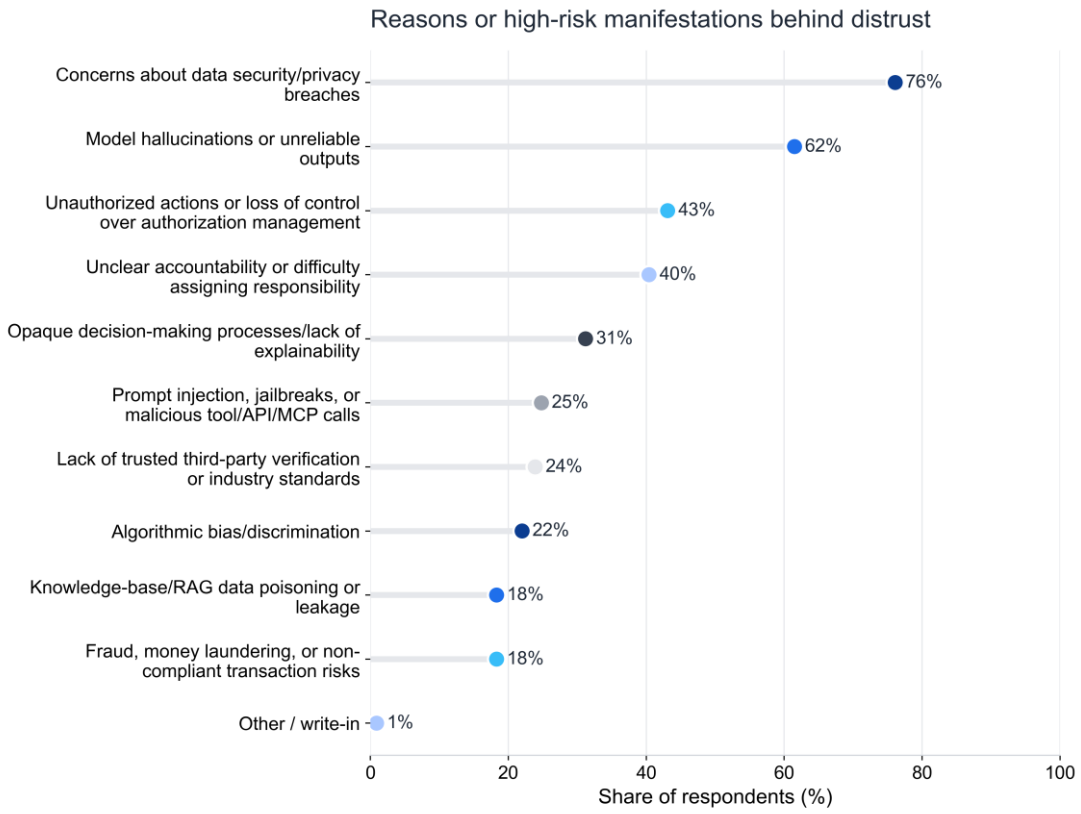
Q8. Impacts caused by trust concerns



Figure 8. Top multiple-choice items: "Increased pressure from compliance, audit, or regulatory assessments" 41.3%; "Delayed project launch or expansion" 40.4%; "Extensive manual confirmation required, reducing efficiency" 37.6%. Percentages may sum to more than 100% because multiple selections were allowed.

Interpretation: The costs of trust concerns are: rising compliance and audit pressure (41%), delayed launch or expansion (40%), more human confirmation and lower efficiency (38%), blocked access to data/APIs/systems (35%), and difficulty with complaints and accountability (29%). A further 23% have already been refused by customers or partners. The first three are hidden transaction costs and correspond to monitoring/enforcement cost and contract complexity in the TC=f(I,R,C,E) framework in Section 1.2.4 of the main report. Organizations hedge the trust gap with human review, extra sign-offs, and access blocks; each is a measurable efficiency tax. Customer refusal (23%) is direct market contraction and corresponds to damage to the market-expansion effect. The relative size of the two cost groups also has a timing implication: efficiency losses appear before market losses. By the time customer attrition is felt, the organization has already paid a long hidden bill for missing trust. Credit-system investment should therefore start before market loss becomes visible, when efficiency savings can already form a measurable return.

Q9. Reasons or high-risk manifestations behind distrust



Multiple selections were allowed; percentages use respondents with a non-empty answer as the denominator.

Figure 9. Top multiple-choice items: "Concerns about data security/privacy breaches" 76.1%; "Model hallucinations or unreliable outputs" 61.5%; "Unauthorized actions or loss of control over authorization management" 43.1%. Percentages may sum to more than 100% because multiple selections were allowed.

Interpretation: The main sources of distrust are data security and privacy leakage (76%), hallucinations or unreliable output (62%), unauthorized actions and loss of authorization control (43%), unclear accountability (40%), and opaque decisions (31%). The first two point to capability-level reliability; the middle two point to institutional controllability. That maps onto the identity–credit–authorization split in Chapter 3. The distribution supports the core claim in Section 3.2.1 of the main report: capability is not the same as trustworthiness. Even if hallucinations and privacy issues ease with better models, unauthorized action and accountability still need independent institutional arrangements. That is why the credit dimension carries the highest weight (45%) in the three-layer score, and why the financial-lending scene raises it further to 45%. Adversarial risks such as prompt injection and data poisoning rank lower (25% and 18%), so the present bottleneck is still basic reliability rather than advanced attack surfaces. As Agent permissions and exposure grow, those adversarial weights should be raised dynamically in the assessment framework rather than held fixed.

Q10. Business pain points for advancing Agent applications

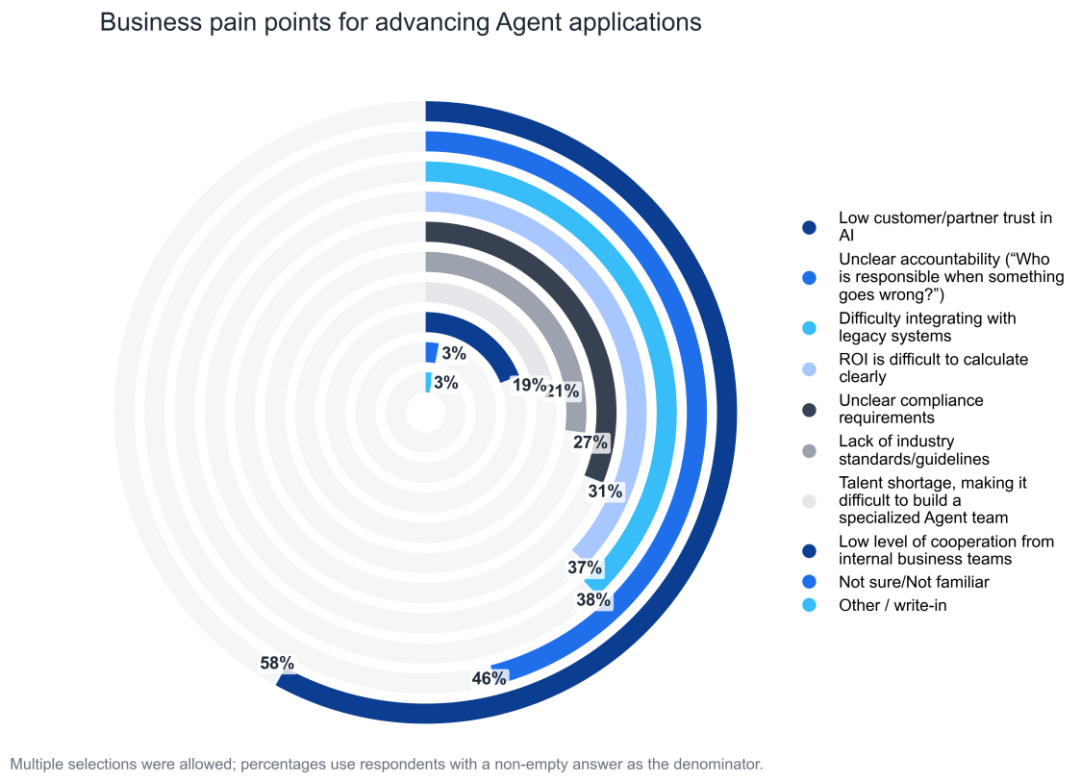


Figure 10. Top multiple-choice items: "Low customer/partner trust in AI" 53.2%; "Unclear accountability ("Who is responsible when something goes wrong?")" 42.2%; "Difficulty integrating with legacy systems" 34.9%. Percentages may sum to more than 100% because multiple selections were allowed.

Interpretation: The leading commercial pain point is low customer and partner trust in AI (53%), followed by unclear accountability after errors (42%). Both rank above legacy-system integration (35%) and difficulty calculating ROI (34%). The binding constraint on commercialization has shifted from technical feasibility to counterparty acceptance. That matches Section 1.2.4 of the main report: commercial value equals value creation times the trust mechanism. Without a trust mechanism, value-creating capability cannot be cashed in, and technical lead does not automatically become commercial success. The high rank of unclear accountability also shows that the allocation of responsibility is itself a term of trade, not a post-incident management issue. Buyers need to know, before contract, who pays if something goes wrong, how much, and on what evidence. That is why Section 3.3.7 lists evidence chains, advance compensation, and liability insurance as companion mechanisms to authorization: without an enforceable loss-bearing arrangement, low trust cannot be solved by a technical demo. It is no coincidence that ROI calculation difficulty (34%) sits alongside the trust pain: trust costs are hard to measure, and that is a major source of distortion in ROI models.

Q11. Pain points in building general compliance capabilities

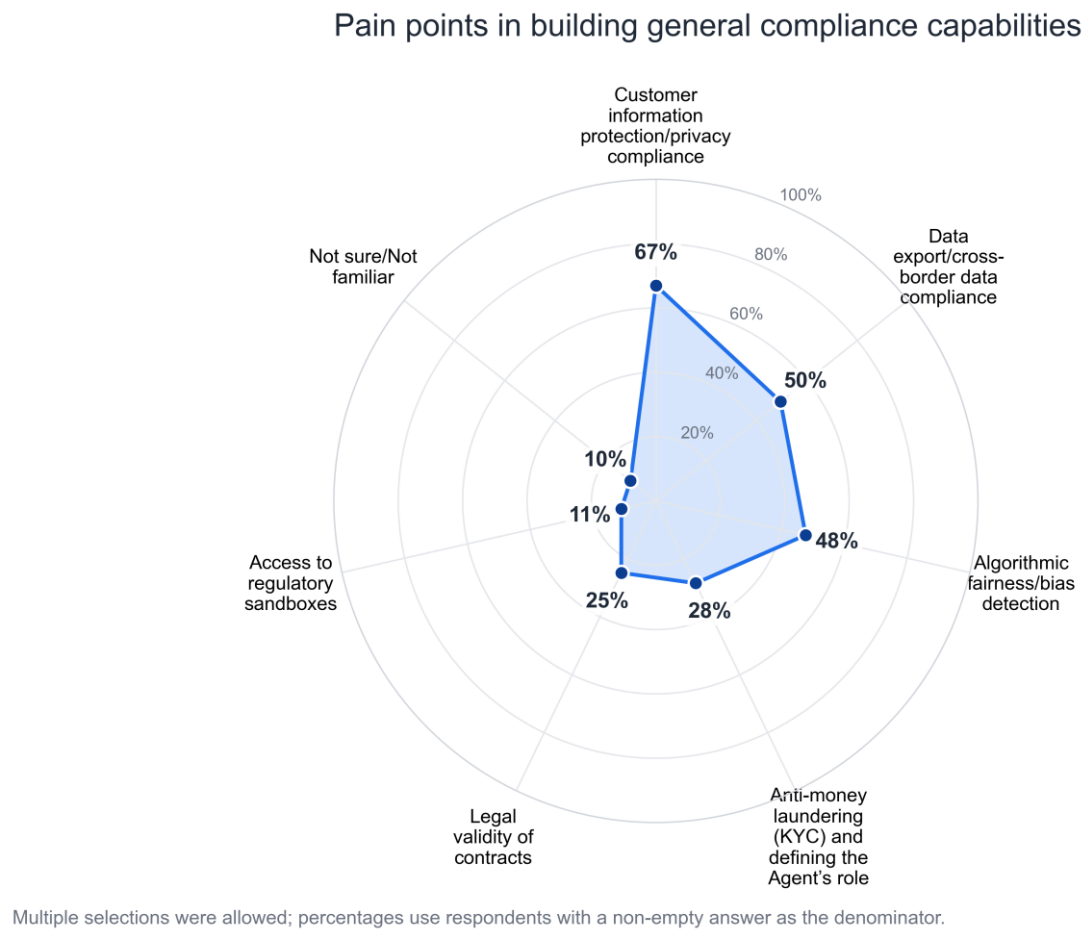


Figure 11. Top multiple-choice items: "Customer information protection/privacy compliance" 67.0%; "Data export/cross-border data compliance" 49.5%; "Algorithmic fairness/bias detection" 47.7%. Percentages may sum to more than 100% because multiple selections were allowed.

Interpretation: Compliance-building pain points concentrate on customer-information protection and privacy (67%), cross-border data compliance (50%), and algorithmic fairness/bias detection (48%), followed by KYC and defining the Agent's role (28%) and the legal validity of contracts (25%). Privacy and cross-border rules are hard regulatory constraints. Their high selection rates show that data handling is a prerequisite compliance step for Agent deployment, and any credit-system data-collection design must answer those two constraints first. Fairness detection near 50% implies that the metric set must cover output fairness, not only accuracy and efficiency. In the 13-dimension, 65-indicator framework in Section 5.2 of the main report, fairness and bias are non-optional. KYC and Agent-role definition at 28% show that some organizations are already asking where an Agent sits in the legal relationship, which maps directly onto the institutionalization of identity trustworthiness in Section 3.1. Only 11% selected regulatory-sandbox access, so most organizations are not yet at a stage that requires a sandbox; the compliance center of gravity is still basic data handling. That share can be tracked over time as a reference indicator of industry compliance maturity.

Q12. Technical security pain points

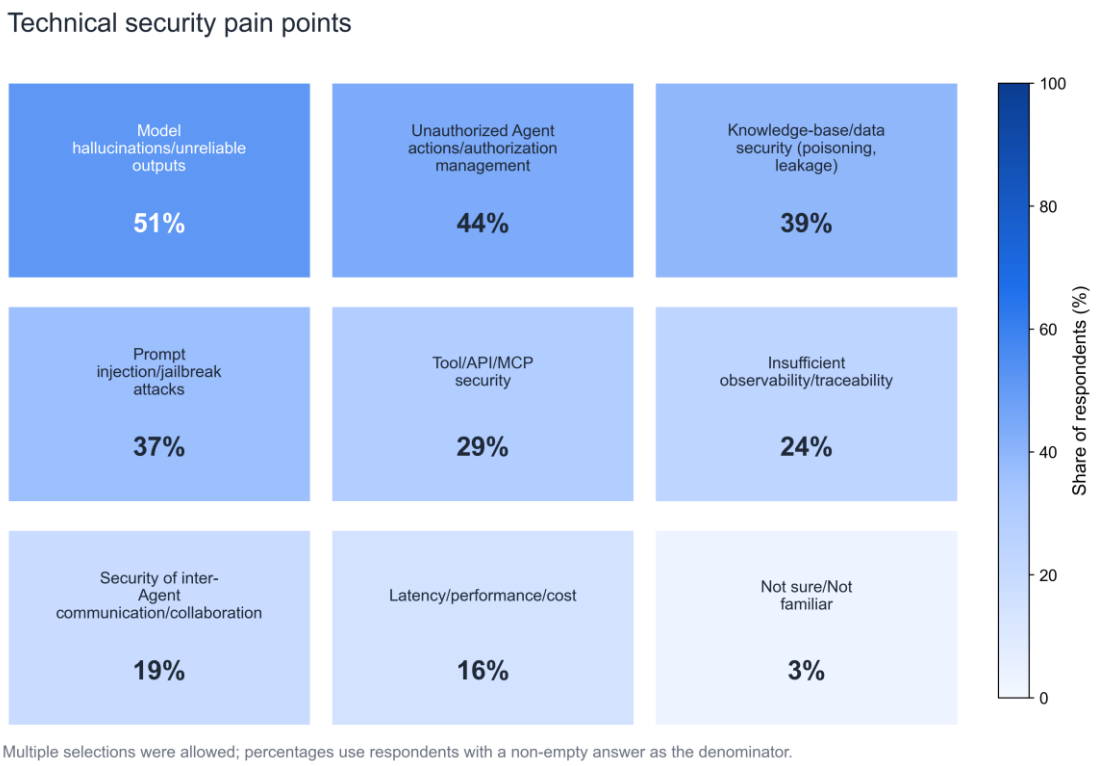


Figure 12. Top multiple-choice items: "Model hallucinations/unreliable outputs" 51.4%; "Unauthorized Agent actions/authorization management" 44.0%; "Knowledge-base/data security (poisoning, leakage)" 39.4%. Percentages may sum to more than 100% because multiple selections were allowed.

Interpretation: The top three technical-security pain points are hallucinations and unreliable output (51%), unauthorized actions and authorization management (44%), and knowledge-base poisoning or leakage (39%), followed by prompt injection (37%) and tool/API/MCP security (29%). Insufficient observability and traceability is 24%. The share is not the highest, but it directly affects the quality of the minimum event set in Section 3.2.3 of the main report. Without complete call-chain and input/output records, fulfillment events, anomalies, and user corrections cannot become verifiable credit evidence, and the credit system idles for lack of data. Observability spend should be treated as credit infrastructure, not an operations add-on. Tool/API/MCP security at 29% also matters. Section 5.2.4 of the main report stresses that assessment should inspect actual execution, tool combinations, and delegation chains. An Agent's risk comes not only from its own model, but from the tool network it can call and the downstream delegates it can invoke. The high incidence of unauthorized action (44%) is the direct result of missing fine-grained tool constraints and of authorization boundaries that do not land on every call.

Q13. Importance of establishing an Agent credit system

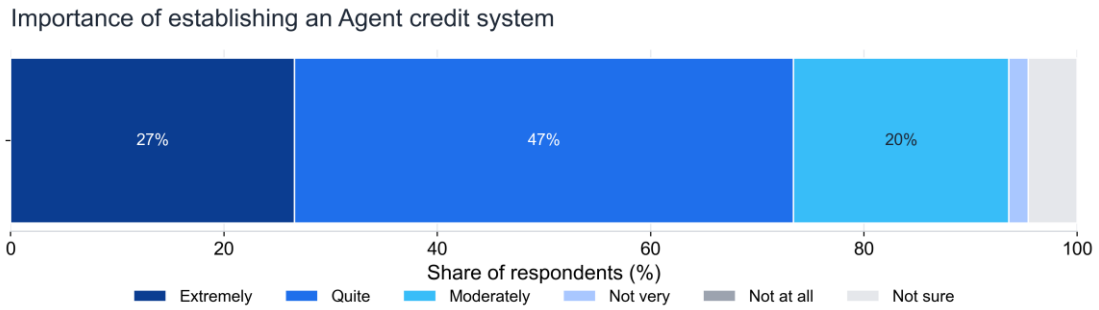


Figure 13. The leading option is "Quite important; can increase customer and partner trust", at 46.8%.

Interpretation: 73% of respondents see establishing an Agent credit system as extremely important (27%) or quite important (47%); only 2% see it as unimportant. Demand consensus is already in place, and the main motive is raising customer and partner trust rather than merely meeting regulation. That supports the framework in Section 1.2.4 of the main report: the value of a credit system is to lower transaction costs and widen the feasible set of trades. Returns show up as an efficiency effect and a market-expansion effect, not only as lower compliance cost. The structure of demand has design implications. Market-trust-driven building is more likely to accept portable, cross-scene display of credit records, because that serves acquisition and contracting directly. If the driver is mainly regulatory, the design center of gravity will shift toward reporting compliance rather than evidence circulation. Note that 20% still chose "moderately important". Those organizations may not yet have hit a trust constraint. From the Q14 maturity distribution, they mostly sit at L1–L2; their stance will change as Agent permissions expand, and they are the main source of demand growth over the next two to three years.

Q14. Closest Agent maturity level

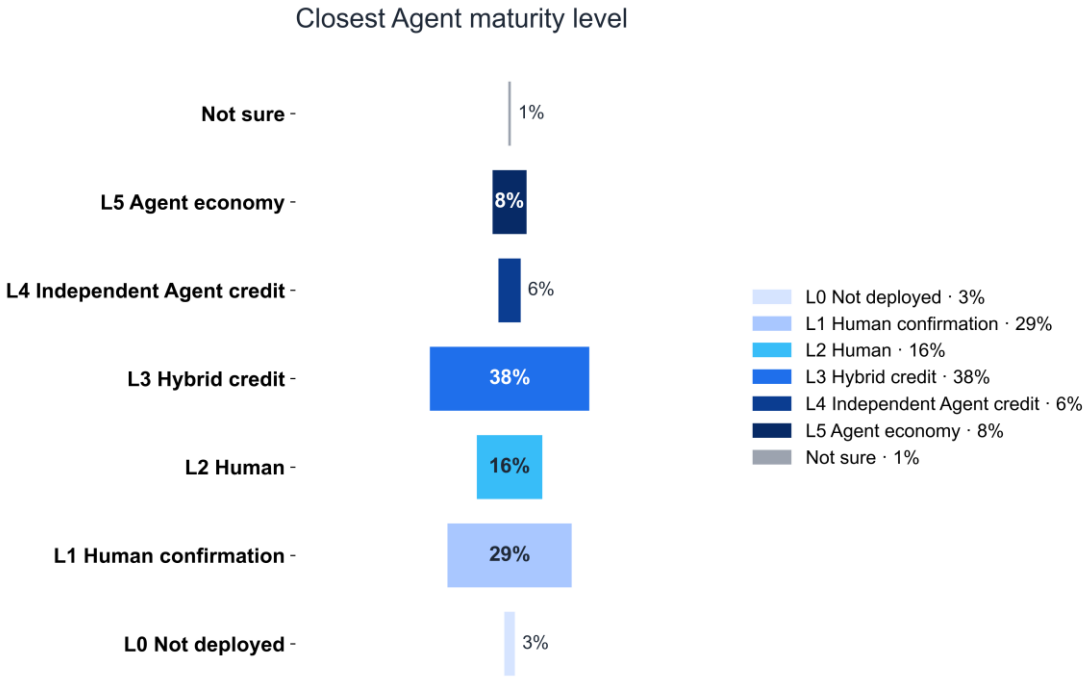


Figure 14. The leading option is "L3 Hybrid credit: Agent can act autonomously within defined limits, with human intervention beyond those limits", at 37.6%.

Interpretation: The maturity distribution is spindle-shaped: L3 hybrid credit is 38%, L1 (29%) and L2 (16%) together are 45%, and L4 (6%) plus L5 (8%) remain below 15%. The industry as a whole sits in the transition from human-anchored credit to task autonomy. Against the upgrade loop in Section 2.1.2 of the main report—capability proof, boundary definition, control establishment, limited authorization, fulfillment records, credit formation, expanded authorization—most organizations have not yet moved from a human backstop to authorization based on records. L3 still has humans as backup takeover, with credit anchored on the operator rather than the Agent. The low share at L4 and above is corroborated by two facts. First, Q41 performance data show that accuracy and human-intervention rates cannot yet support independent credit. Second, the four L4 pillars in Section 2.6.1—persistent identity, assumption of economic consequences, portable reputation, and enforceable default mechanisms—are all incomplete in today's infrastructure. The spindle shape implies that the main migration over the next three years will be in the L2→L3→L4 band, and first demand for assessment and certification services will concentrate there.

Q15. Obstacles to moving to a higher maturity level

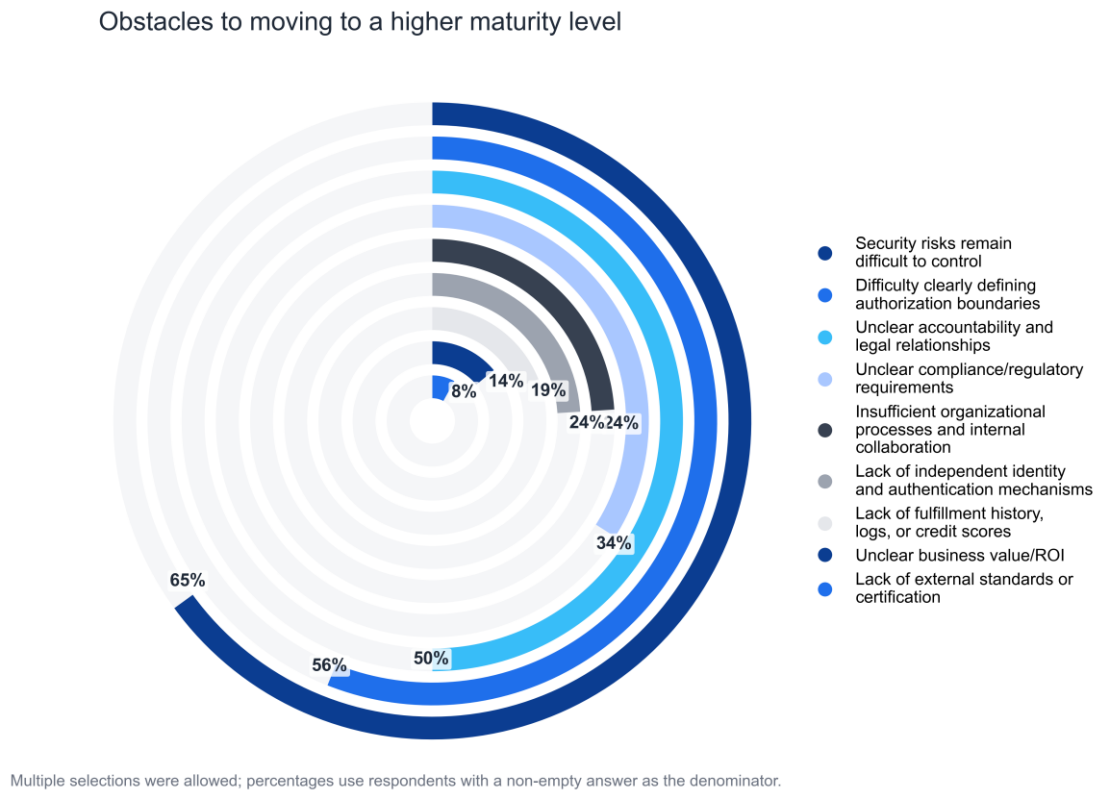


Figure 15. Top multiple-choice items: "Security risks remain difficult to control" 59.6%; "Difficulty clearly defining authorization boundaries" 51.4%; "Unclear accountability and legal relationships" 45.9%. Percentages may sum to more than 100% because multiple selections were allowed.

Interpretation: The top three upgrade obstacles are security risk remaining hard to control (60%), difficulty defining authorization boundaries (51%), and unclear accountability and legal relationships (46%). All sit in authorization trustworthiness and accountability. Lack of external standards or certification is only 7%. The structure shows that the binding constraint on maturity jumps is internal engineering and governance capacity, not missing external standards. The industry cannot blame stagnation on "having no standard to follow". In the credit-flywheel terms of Section 2.8.2 of the main report, if security and authorization do not work, fulfillment records cannot convert into larger authorization, and the flywheel stalls at "control establishment". Concretely, hard-to-define authorization boundaries (51%) mean organizations cannot map task risk onto a definite grade in the five-level authorization-intensity matrix in Section 3.3.3. Unclear accountability (46%) means that even if fulfillment data are recorded, no responsible party is willing to sign a larger authorization on the back of those data. Together they explain the L3-to-L4 blockage in Q14 and set the intervention order: first make authorization boundaries operational and institutionalize accountability; only then does credit accumulation have a starting condition.

3.3 Current State of KYA Capability Building

This chapter looks separately at identity trustworthiness, fulfillment/creditworthiness, and authorization trustworthiness, and identifies which capabilities are already in place, still planned, or missing.

Q16. Identity trustworthiness capabilities

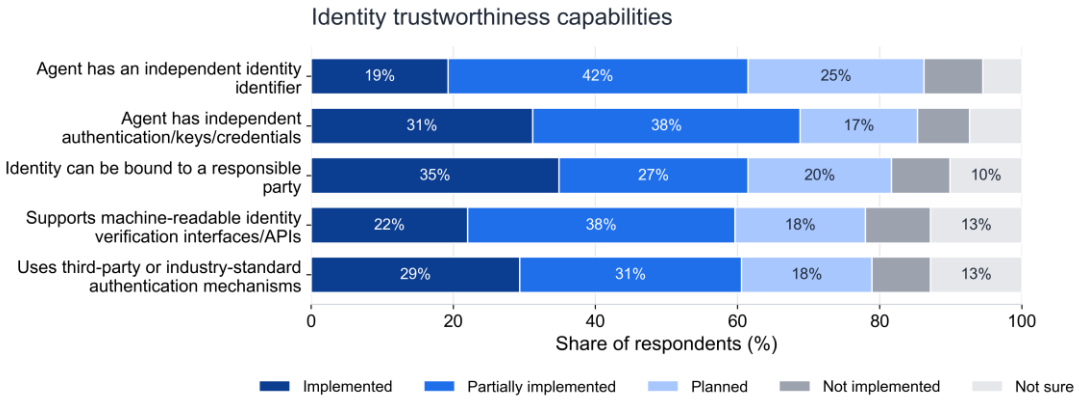


Figure 16. The item with the highest share of "Implemented + Partially implemented" is "Agent has independent authentication/keys/credentials" (68.8%).

Interpretation: The share of identity-trustworthiness elements that are implemented or partially implemented sits in a narrow 60%–69% band. Independent authentication and key credentials rank highest (69%); machine-readable verification interfaces rank lowest (60%). Identity building has started relatively evenly, which means there are already basic answers to "who is it" and "whom does it represent". The shortfall in machine-readable interfaces still matters: they are a prerequisite for cross-organization verification and for the protocolization discussed in Section 3.4.5 of the main report. Weak interfaces mean identity checks remain inside a platform loop and cannot support the portable identity discussed in Section 3.1.9—identity and records accumulated on one platform cannot be verified cheaply on another. Identity is the carrier of credit. Incomplete attribution transmits upward into the credit and authorization dimensions: if one cannot stably answer "is it still the same Agent," fulfillment records may be hung on the wrong subject, and authorization decisions lose their anchor. The 60%–69% band also means more than 30% of organizations have not fully implemented even basic identity elements. Spreading identity capability remains a prerequisite for building all three dimensions.

Q17. Creditworthiness capabilities

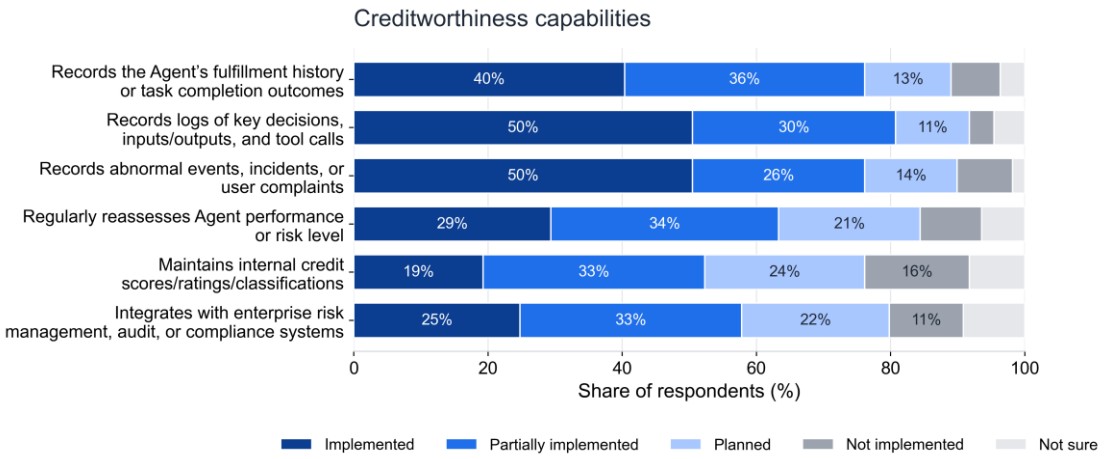


Figure 17. The item with the highest share of "Implemented + Partially implemented" is "Records logs of key decisions, inputs/outputs, and tool calls" (80.7%).

Interpretation: Inside creditworthiness, the split is sharp. Record-type capabilities such as key decision logs (81%), fulfillment history (76%), and anomaly records (76%) are relatively widespread, while internal credit scoring (52%) and integration with risk, audit, and compliance systems (58%) lag. The industry is still at data collection, not yet at evaluative judgment, which matches the architectural split between the bookkeeping layer and the judgment layer in Section 3.4.5 of the main report. Records are the raw material of credit, not credit itself. Without scoring and integration, existing records cannot enter authorization decisions, and evidence assets sit idle. That directly limits α+β hybrid credit in Section 3.2.4: combining α (subject credit) and β (task fulfillment) requires a scoring layer that structures raw records. Against the five fulfillment elements in Section 3.2.2, record capabilities cover "what happened," while scoring and integration answer "is it worth trusting, and dare we grant more authority". The latter are 20–29 points less widespread, so the next stage of credit-dimension building should shift from "record everything" to "judgment that can actually be used".

Q18. Authorization trustworthiness capabilities

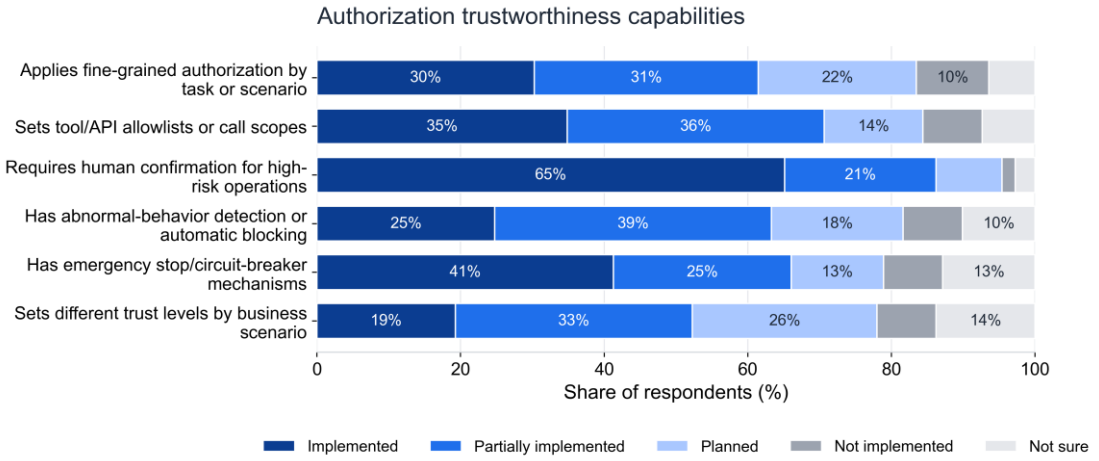


Figure 18. The item with the highest share of "Implemented + Partially implemented" is "Requires human confirmation for high-risk operations" (86.2%).

Interpretation: Authorization trustworthiness is strong on backstops and weak on tiering. Human confirmation for high-risk operations reaches 86%, and tool/API allowlists 71%, while differentiated trust levels by business scenario are only 52%, the lowest of the six items. Widespread human gates show that a floor of caution is common. Emergency stop (66%) and anomaly blocking (63%) also cover part of the emergency mechanisms in Section 3.3.6 of the main report. Missing scenario-based tiering, however, means authorization intensity cannot move with task risk, which is a clear gap versus the risk-tiering matrix in Section 3.3.3. The five-level authorization-intensity design assumes the organization can distinguish task risk and match a corresponding grade. Long-term reliance on human confirmation also produces confirmation fatigue and linearly rising supervision cost. Tiered authorization is the necessary path for releasing people from low-value approvals. Together with the 39% in Q22 who see authorization as the weakest dimension, the direction of build-out is already clear: from "is there a gate" to "is the gate tiered, and is the tier based on credit data".

Q19. Security or compliance reviews before launch

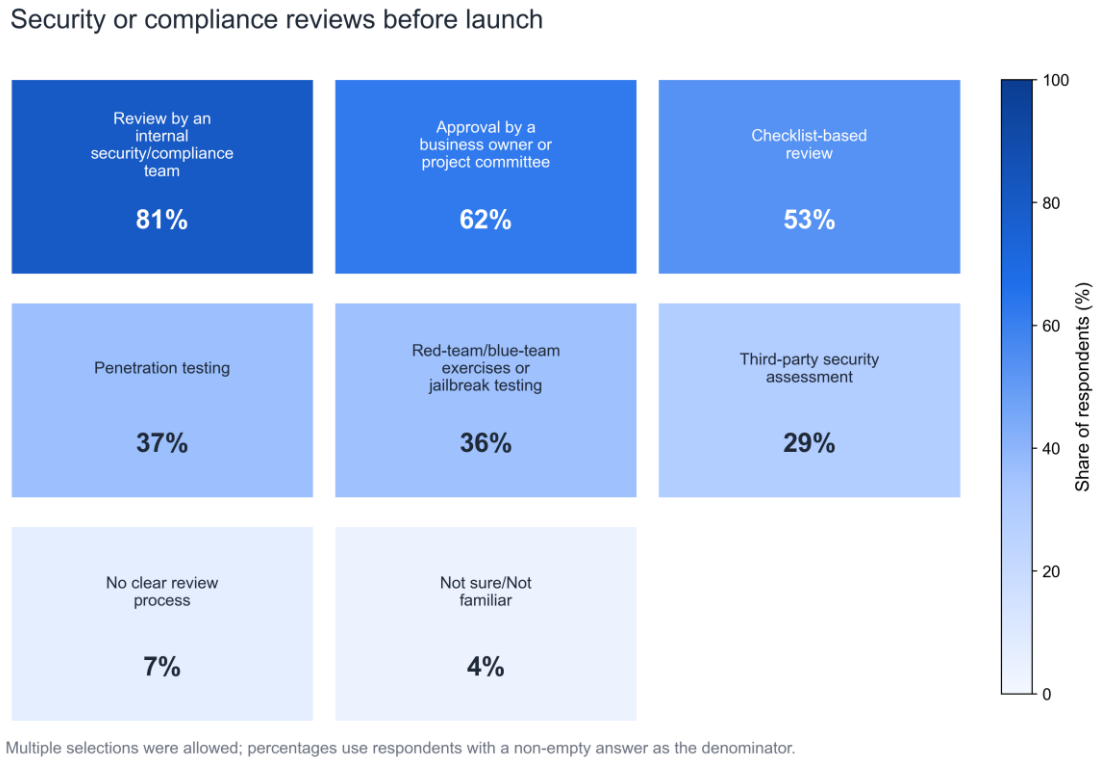
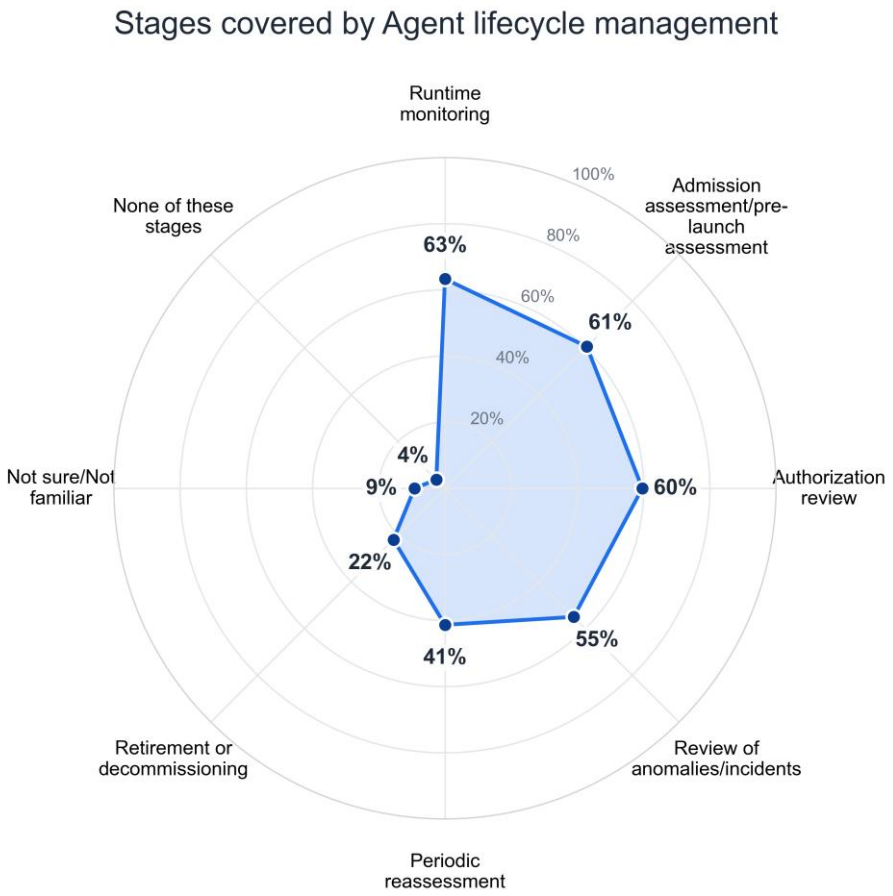


Figure 19. Top multiple-choice items: "Review by an internal security/compliance team" 80.7%; "Approval by a business owner or project committee" 62.4%; "Checklist-based review" 53.2%. Percentages may sum to more than 100% because multiple selections were allowed.

Interpretation: Pre-launch review is dominated by internal security/compliance review and business-owner approval. Third-party security assessment, red/blue-team exercises, and penetration testing are much less used, and about 4% have no defined review process. The review system is internal gates first, external verification second. Internal review knows the business context and is cheaper, but it lacks independence and an adversarial view and systematically underestimates its own design blind spots. The third-party verifiability discussed in Section 3.1.8 of the main report has limited penetration, so "self-attested trust" is still the mainstream, while counterparties actually need "other-attested trust". As Agent permissions expand and cross-organization collaboration grows, the gap in external assessment and adversarial testing will become a hard constraint on entry to high-risk scenes. Under the critical-threshold logic in Section 5.2.5 of the main report, scenes involving money and important data should treat independent security assessment as an entry condition, not a bonus. A 4% no-process share is small, but it is unacceptable in high-risk scenes and should be cleared as an industry floor.

Q20. Stages covered by Agent lifecycle management



Multiple selections were allowed; percentages use respondents with a non-empty answer as the denominator.

Figure 20. Top multiple-choice items: "Runtime monitoring" 63.3%; "Admission assessment/pre-launch assessment" 60.6%; "Authorization review" 59.6%. Percentages may sum to more than 100% because multiple selections were allowed.

Interpretation: Lifecycle coverage concentrates on admission assessment and runtime monitoring. Periodic reassessment, authorization review, and retirement/decommissioning are clearly weaker, and about 4% cover none of the stages. The center of gravity is pre-launch. That gap undermines two mechanisms. First, the dynamic assessment in Section 5.1 of the main report: credit information cannot update with behavior, and scores go stale quickly. Second, the three principles of dynamic downgrade in Section 2.8.3: after a model change, a permission change, or an incident, there is no institutional channel that triggers review, so downgrade cannot happen. Low coverage of retirement also creates a hidden risk: unused Agents that keep valid credentials and authorizations become an unowned attack surface. A credit system works through operating-stage data flows, not a one-off admission. Admission answers "was it trustworthy at launch". The market needs a continuing answer to "is it still trustworthy now," which only reassessment, review, and retirement can supply.

Q21. Accountability mechanism after errors, unauthorized actions, or losses

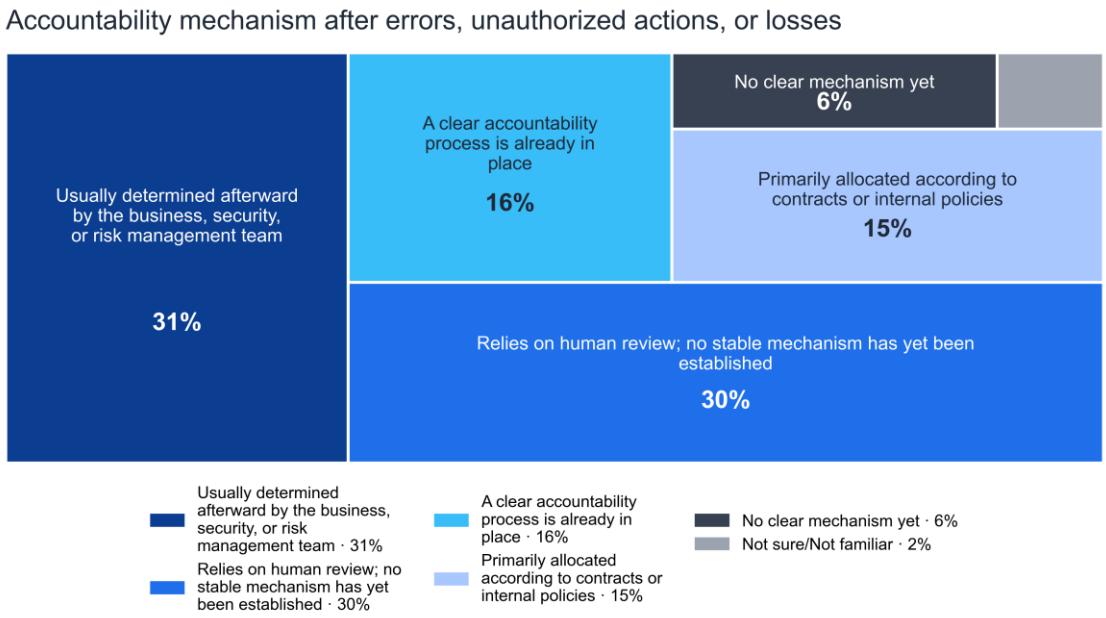


Figure 21. The leading option is "Usually determined afterward by the business, security, or risk management team", at 31.2%.

Interpretation: Accountability mechanisms are weak. Only 17% have a defined process; 31% usually decide after the fact through business, security, or risk teams; 30% rely on human review with no stable mechanism; and 6% have no mechanism at all. In the framework of Section 2.1.2 of the main report, accountability and the operational design domain (ODD) are two direct dimensions of maturity. An institutional vacuum in accountability means most organizations have no stable answer to who authorized, who supervised, and who bears the loss. That inflates post-incident cost—each incident requires an ad hoc adjudication group and a fresh argument over responsibility boundaries—and leaves credit records without a responsibility anchor: fulfillment data that cannot be attributed to a subject cannot enter anyone's credit. Against the design in Section 3.3.7, a complete accountability arrangement includes layered responsibility, an evidence chain, advance compensation, and liability insurance. Today's 17% with a defined process means most organizations have not even finished the first component. Accountability is the link in the three-dimension system that most needs to be institutionalized first, and a reasonable candidate for a veto item in the assessment framework.

Q22. Most significant current weakness among the three KYA dimensions

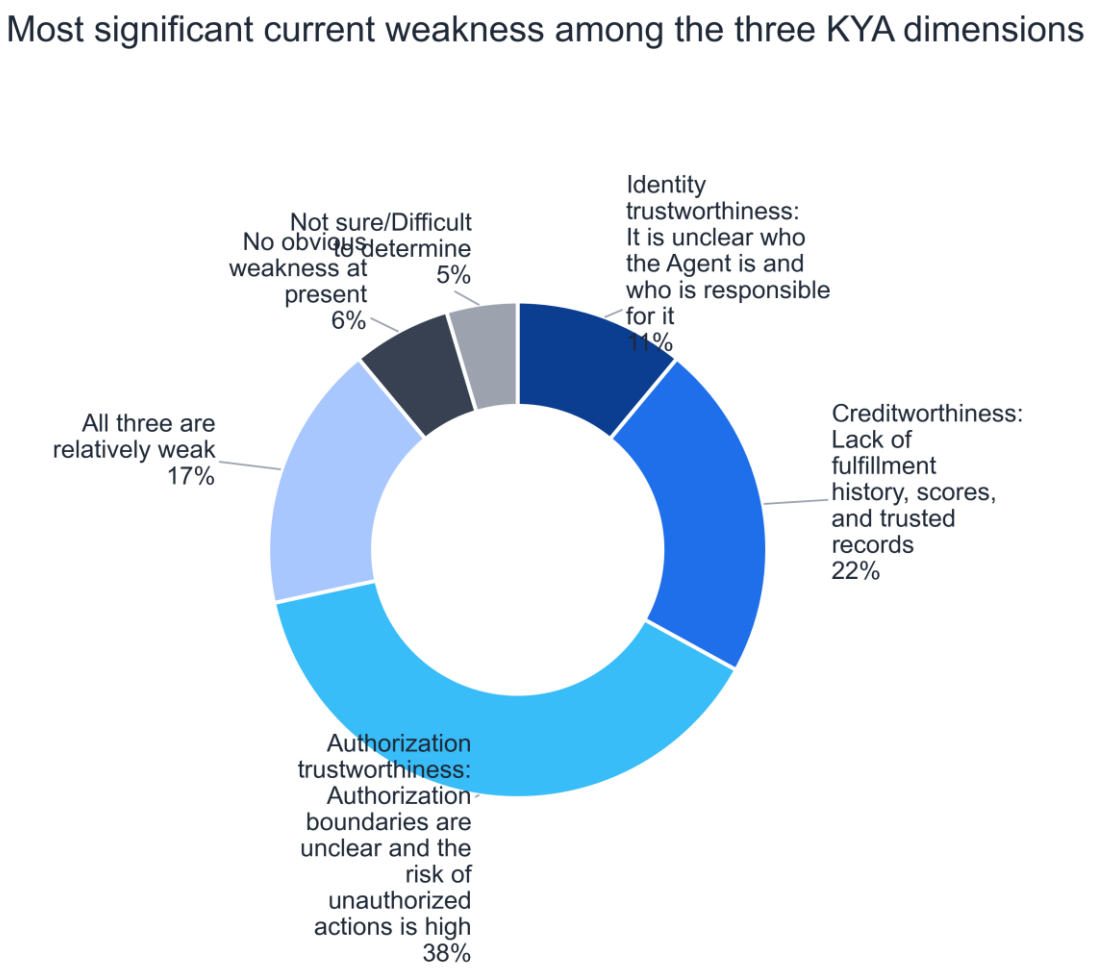


Figure 22. The leading option is "Authorization trustworthiness: Authorization boundaries are unclear and the risk of unauthorized actions is high", at 38.5%.

Interpretation: On the weakest of the three dimensions, 39% chose authorization trustworthiness, 22% creditworthiness, 17% said all three are relatively weak, and 11% chose identity trustworthiness. The subjective ranking matches the capability data: scenario-based trust levels in Q18 have the lowest implementation rate (52%), and difficulty defining authorization boundaries ranks second among Q15 obstacles (51%). Objective gaps and subjective weakness corroborate each other. In the language of Section 3.4.1 of the main report, identity answers attribution, credit answers whether it is worth trusting, and authorization answers whether one dares to grant more power. When authorization boundaries are unclear, even attributable identity and recorded fulfillment still leave authorization decisions dependent on conservative human judgment, and the credit flywheel cannot start: accumulated data cannot be exchanged for larger authorization, so the act of recording itself loses its incentive. The 17% who see all three as weak also matter: a single-point breakthrough will not solve their problem; they need a whole starting checklist of the kind in Section 5.3.4 of the main report. Identity trustworthiness has the lowest selection rate (11%), consistent with Q16 implementation rates above 60% on each element. Identity is currently the strongest of the three dimensions.

3.4 Demand for a Credit System

This chapter focuses on whether the industry needs unified standards, what a framework should prioritize, and how organizations would respond to mature assessment and certification services.

Q23. Need for unified standards for Agent trustworthiness assessment/certification

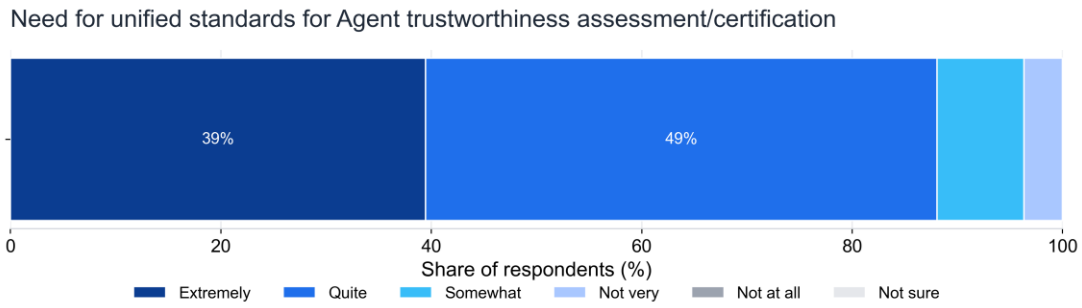


Figure 23. The leading option is "Quite needed; they would benefit industry development", at 48.6%.

Interpretation: 88% of respondents say the industry needs unified Agent trustworthiness assessment or certification standards; 39% say they are extremely needed and should be established urgently; only 4% say they are not very needed. Demand is near consensus. In trust-economics terms, a unified standard lowers cross-organization verification cost. Without an accepted standard, every counterparty must repeat due diligence, credit cannot circulate across scenes, and the market-expansion effect in Section 1.2.4 of the main report is offset by verification cost. This is direct demand-side evidence for the Chapter 5 assessment framework. Combined with Q15, where "lack of external standards" is only a 7% obstacle, an apparent contradiction needs explaining: organizations do not think missing standards block their own upgrade, yet they strongly want a unified industry standard. The reasonable reading is that the main value of a standard is not to guide one's own build, but to lower the cost of mutual trust with external counterparties. A standard is first market infrastructure, and only second a construction guide. That implies the design should first unify verifiable evidence formats, not prescribe internal implementation paths.

Q24. Priority contents and evaluation requirements for the trustworthiness framework

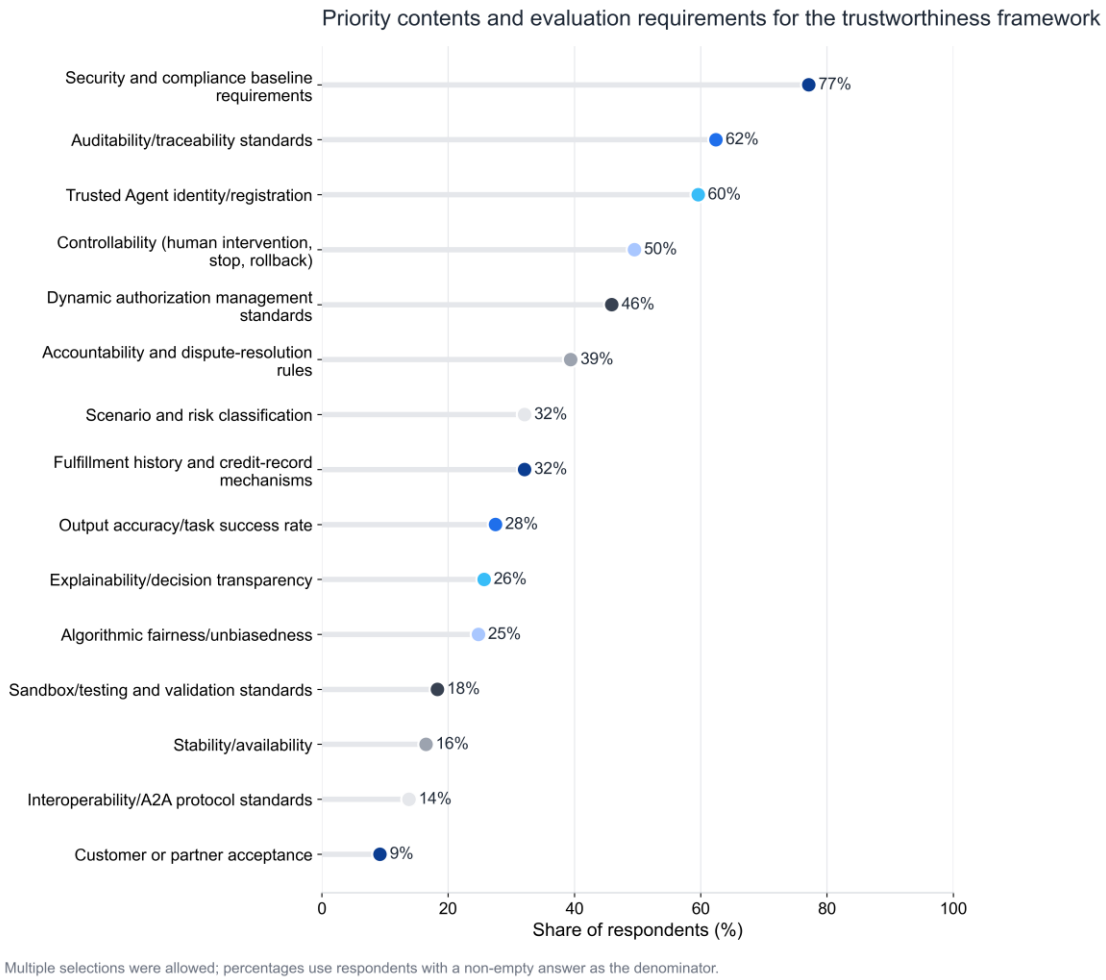


Figure 24. Top multiple-choice items: "Security and compliance baseline requirements" 77.1%; "Auditability/traceability standards" 62.4%; "Trusted Agent identity/registration" 59.6%. Percentages may sum to more than 100% because multiple selections were allowed.

Interpretation: The priority order for framework content is clear. Security and compliance baselines (77%), auditability/traceability standards (62%), and trusted identity registration (60%) occupy the top three; interoperability and A2A protocols (14%) and customer or partner acceptance (9%) sit at the bottom. The market's present core demand is baseline trust first, interconnection later. That order supports a progressive standardization path: first unify the fields and verification methods for identity, authorization, and event records so that credit evidence can be produced and checked; push cross-platform interoperability only after evidence standards stabilize, when the cost is lower. Interoperability without a shared evidence format merely connects incomparable data. The high priority of auditability and traceability (62%) maps directly onto the minimum event set in Section 3.2.3 of the main report: audit presupposes complete event records, and traceability presupposes that records can be joined across steps. Customer acceptance at only 9% does not mean the demand side is unimportant. It reflects a widespread supply-side judgment: before baseline capabilities exist, trust communication toward customers has no facts to stand on.

Q25. Importance of design principles

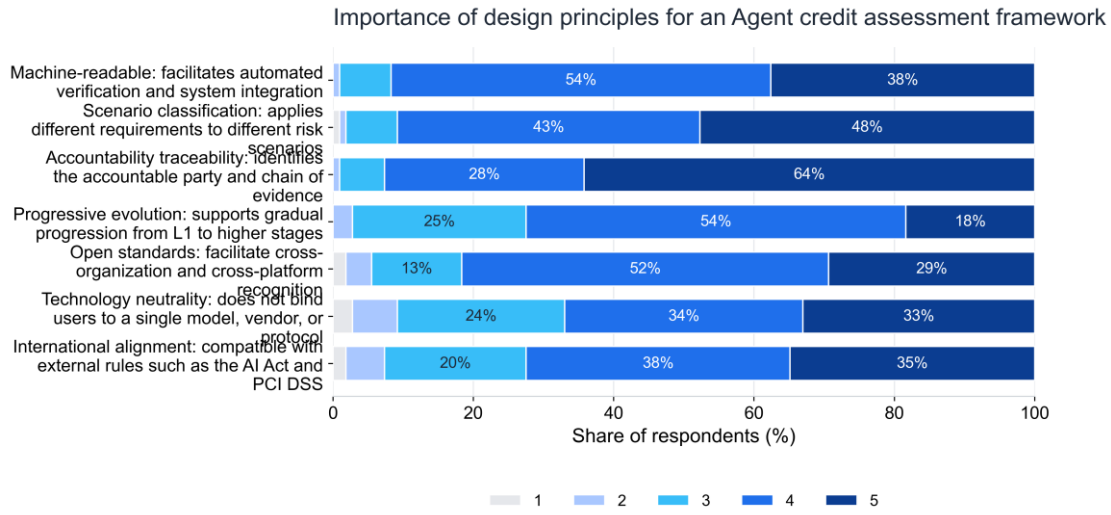


Figure 25. The item with the highest share of a rating of 5 (very important) is "Accountability traceability: identifies the accountable party and chain of evidence" (64.2%).

Interpretation: Among the seven design principles, accountability traceability has a 4–5 score rate of 93% (mean 4.56), machine-readability 92%, and scenario classification 91%. Open standards are at 82%, progressive evolution and international alignment 73% each, and technology neutrality is relatively lower at 67%. Respondents place the most weight on being able to find the authorization source and responsible party for key actions, and on matching metric requirements to action risk, which is consistent with the five principles in Section 5.1.2 of the main report. The first three form a complete evidence chain: scenario classification decides what to record, machine-readability decides how records are used, and accountability traceability decides whom the record ultimately points to. Their joint high scores show that the market's core expectation of an assessment system is usable evidence, not a flattering score. The lower priority of technology neutrality does not mean it is unimportant; it means the present primary contradiction is responsibility and evidence, not the right to choose a technical path. International alignment at 73% sits in the middle and echoes the 50% cross-border compliance pain in Q11, suggesting that standard design should reserve interfaces for cross-border mutual recognition.

Q26. Likely attitude toward mature Agent trustworthiness assessment/certification services

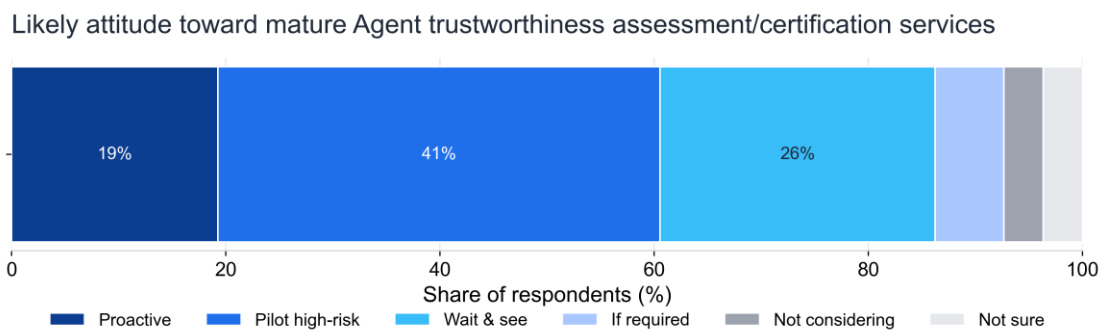


Figure 26. The leading option is "Willing to pilot it for high-risk or critical scenarios", at 41.3%.

Interpretation: Faced with mature assessment and certification services, 41% would first pilot in high-risk or critical scenes, and 19% would adopt proactively and make it a prerequisite for launch or collaboration, for a combined 60% positive stance. 26% would wait and see, pending industry consensus or regulatory requirements. A pilot-first path matches the scene-based design in Section 5.3 of the main report: assessment parameters need to be calibrated on pilot data; high-risk scenes have the strongest need for trust evidence and the clearest willingness to pay; entering through those scenes can produce both data and revenue. The 19% "prerequisite" group is especially important: they have already begun embedding third-party assessment in their own procurement and launch processes. Their existence gives certification services a starting point for network effects—each additional counterparty that requires certification raises the value of being certified. The wait-and-see group is about a quarter of the sample; their conversion depends on verifiable results from the first pilots. Under Section 5.3.3 of the main report, pilots should include explicit threshold calibration and a downgrade/recovery channel, so observers can see a replicable closed loop rather than a one-off demo.

Q27. External support most wanted for building an Agent trustworthiness system

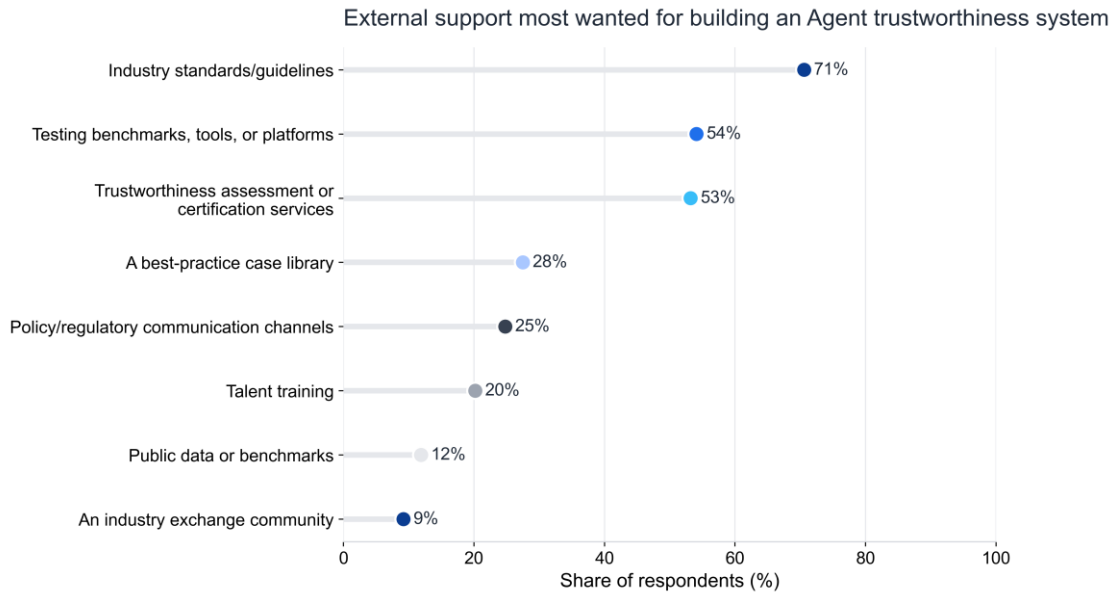


Figure 27. Top multiple-choice items: "Industry standards/guidelines" 70.6%; "Testing benchmarks, tools, or platforms" 54.1%; "Trustworthiness assessment or certification services" 53.2%. Percentages may sum to more than 100% because multiple selections were allowed.

Interpretation: The external support most wanted is industry standards and guidelines (71%), testing benchmarks and tool platforms (54%), and assessment/certification services (53%), followed by a best-practice case library (28%) and regulatory communication channels (25%). Standards, tools, and services all exceed 50%, which means a text standard alone is not enough to land: organizations

also need executable tests to verify compliance, and third-party services to lower the cost of building in-house. That echoes the critical-threshold design in Section 5.2.5 of the main report: a threshold works only if independent testing and verification exist; without a tool platform, a threshold degenerates into a self-declaration. The three high-vote supports also depend on one another. Standards define what to test, tool platforms solve how to test, and certification services answer whose results can be trusted. Together they form a complete market-infrastructure chain; a gap in any link raises overall adoption cost. Regulatory communication channels at 25% are relatively low, comparable to the wait-for-regulation group in Q26, which shows that most organizations currently care more about market-level mutual trust than about regulatory approval.

Q28. Additional suggestions — keyword themes

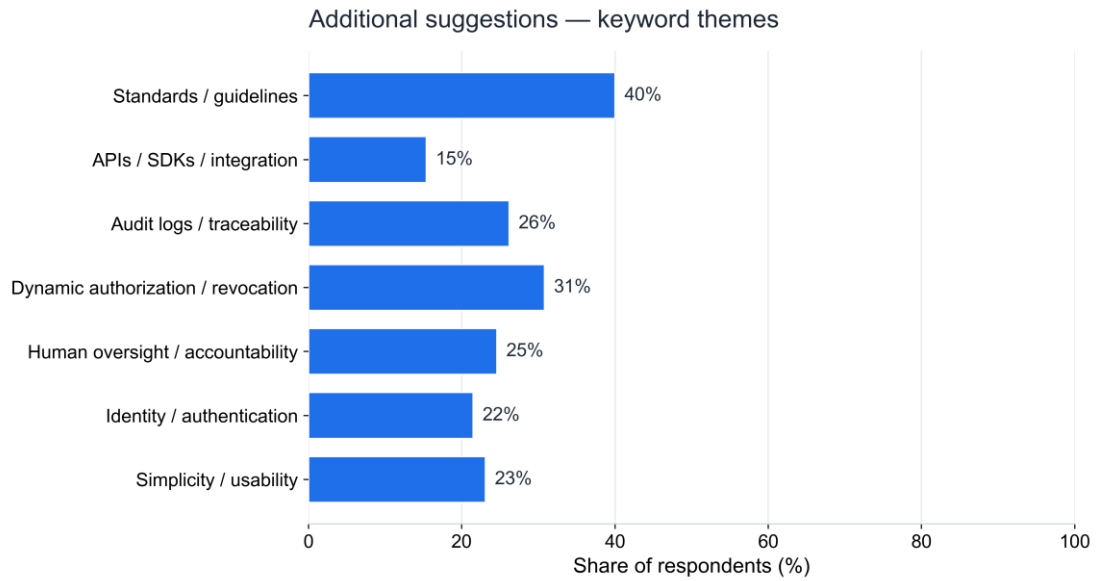


Figure 28. The most frequent theme is "Standards / guidelines" (40.0%).

Interpretation: Keyword themes in the open suggestions concentrate on five directions: standards and guidelines, API/SDK integration, audit logs and traceability, dynamic authorization and revocation, and human oversight and accountability. That is consistent with the earlier quantitative results. The suggestions are mainly engineering asks, with little conceptual debate, which shows that respondents already treat KYA as a system that must land, not a statement of principle. The theme distribution also shows continuing attention to the authorization lifecycle (grant, adjust, revoke), which matches "authorization as data" in Section 3.3.2 of the main report: authorization needs to become an object that can be passed, verified, and revoked, not a one-time configuration. The five themes form a triangle of mutual confirmation with Q24 framework priorities and Q27 external-support demand: standards answer what to build against, APIs/SDKs answer how to connect at low cost, logs and traceability answer where evidence comes from, dynamic authorization answers how permissions move with credit, and human oversight answers who handles exceptions. Open-ended and closed-ended conclusions agree, so the

quantitative results are not an artifact of option wording; the demand structure is robust.

3.5 Optional Module: BNPL / Consumer Finance

This module was completed only by respondents familiar with the relevant business. Interpretations should note that the subsample is smaller than the full sample.

Q29. Extent of Agent use in BNPL/consumer finance

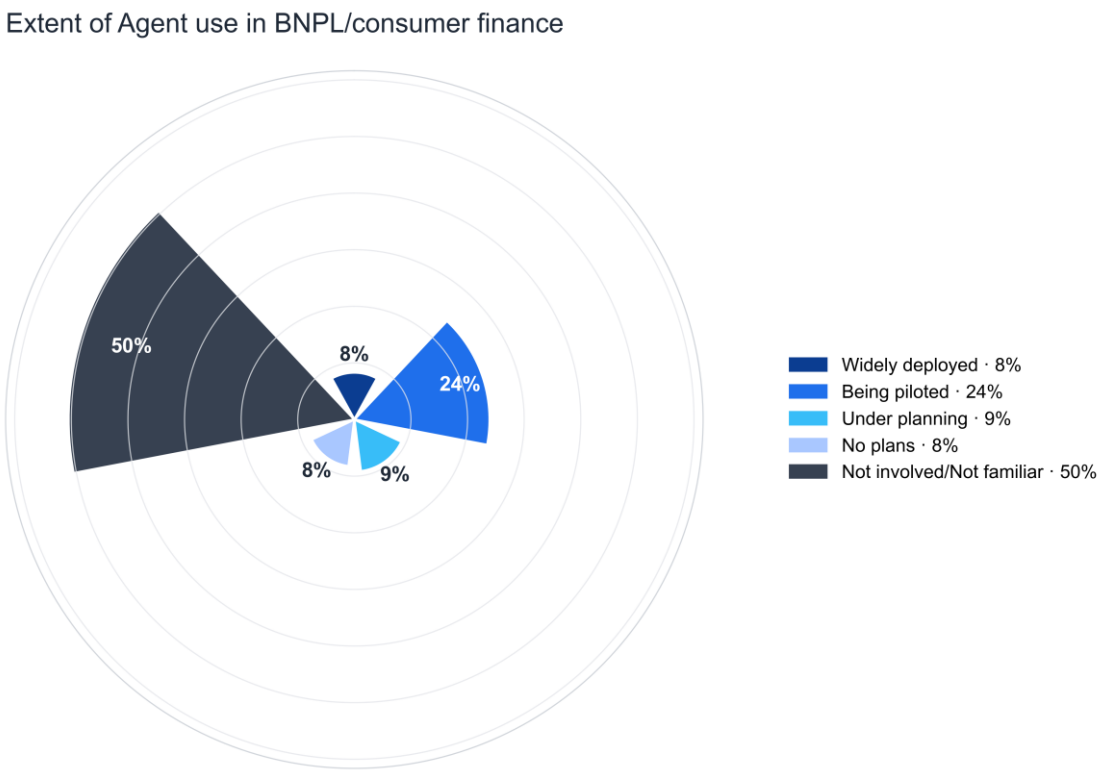


Figure 29. The leading option is "Not involved/Not familiar", at 50.5%.

Interpretation: Among respondents who completed the BNPL/consumer-finance module, 24% are in a pilot stage, 8% have deployed widely, 9% are planning, and 8% have no plans. The overall picture is pilot-led, with little scaled deployment. Consumer finance is directly money-related and demands more trust evidence than ordinary tasks. Current penetration matches the BNPL-scene analysis in Section 5.3.2 of the main report: assessment design for this scene must cover both efficiency and trust metrics; optimizing a single dimension cannot support scale. Pilot dominance also shows that the industry is in the parameter-calibration stage described in Section 5.3.3: reasonable bands for pass rates, limits, and intervention thresholds need to be set from pilot data, not copied from existing credit models. Agent approval has a different error pattern from human approval, so historical risk parameters cannot be reused as-is. The 8% with wide deployment are an important observation sample: their threshold settings, human-review triggers, and bad-debt handling will become de facto industry benchmarks for later entrants.

Q30. Main roles of Agents in BNPL/consumer finance

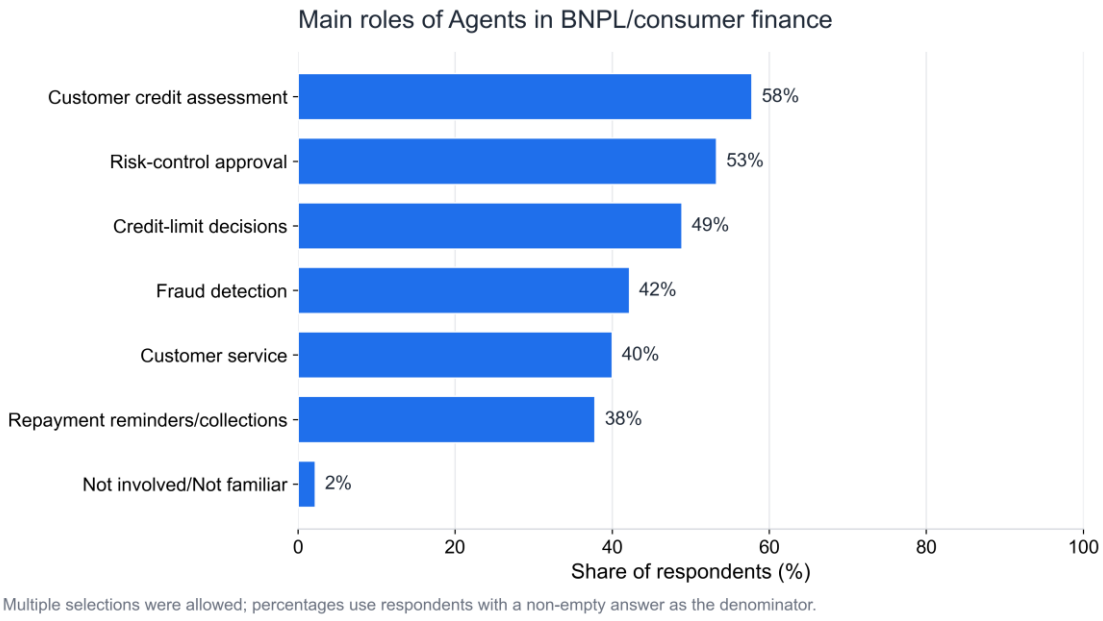


Figure 30. Top multiple-choice items: "Customer credit assessment" 57.8%; "Risk-control approval" 53.3%; "Credit-limit decisions" 48.9%. Percentages may sum to more than 100% because multiple selections were allowed.

Interpretation: Agents in BNPL/consumer finance already sit inside core underwriting: customer credit assessment (58%), risk-control approval (53%), and limit decisions (49%) lead, followed by fraud detection (42%), customer service (40%), and collection reminders (38%). Agent output enters the formation of asset quality directly. Errors travel along the transmission path in Section 1.2.2 of the main report into actual money loss, and they replicate at scale: a bad approval policy can be copied onto thousands of applications in a short time. This scene therefore has no low-risk buffer. Agents that perform credit assessment, limits, and approval should face the highest assessment and authorization requirements. Against Section 3.4.3 of the main report, the three-layer weights in financial lending are identity 15%, capability 40%, and credit 45%, with credit highest. That means historical fulfillment records predict future performance in this scene better than a model-capability demo. Collection reminders at 38% also show that Agents have appeared in post-loan management; the compliance and traceability of that behavior belong in the credit assessment as well.

Q31. Agent vs human decisions

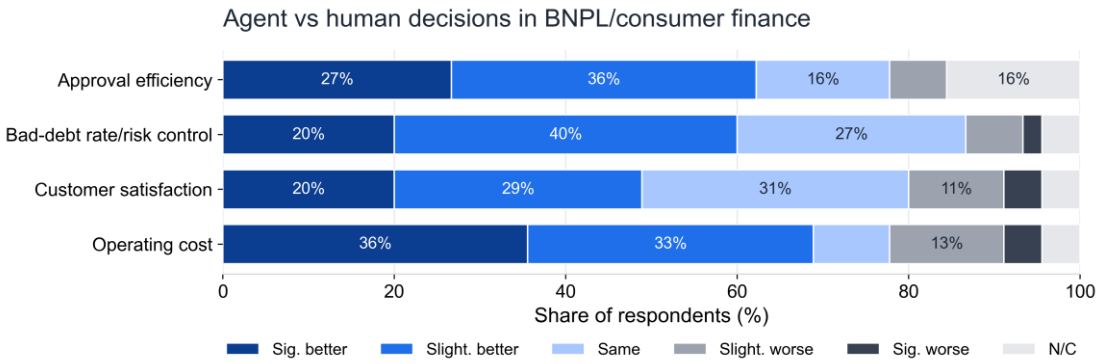


Figure 31. Ranked by the combined share of "significantly better + slightly better", Operating cost received the highest positive rating.

Interpretation: Compared with human decisions, Agents show the clearest advantage on operating cost (69% better) and approval efficiency (62%); bad-debt/risk control (60%) is also ahead. Customer satisfaction is the most divided of the four: only 49% see it as better, and 16% see it as worse. Efficiency and cost gains are already visible; experience and trust remain the short board. Under the scene logic in Section 5.3.2 of the main report, BNPL assessment must hold efficiency and trust together: the faster the approval, the more it needs explainable reasons, a reviewable channel, and attributable records to keep consumer trust. The split on satisfaction is worth a close reading. A 16% "worse" share is the highest of the four items, which means Agent approval in some organizations has already produced perceptible customer friction: unclear decline reasons, no path to appeal, and slow error correction all damage trust directly. Section 5.3.2 also warns that a 98% completion rate is not a repayment probability; efficiency metrics cannot replace asset-quality metrics. Likewise, cost and speed advantages cannot replace the long-term customer-relationship value reflected in satisfaction.

Q32. Observed consumer trust/acceptance of Agent-approved BNPL/consumer finance

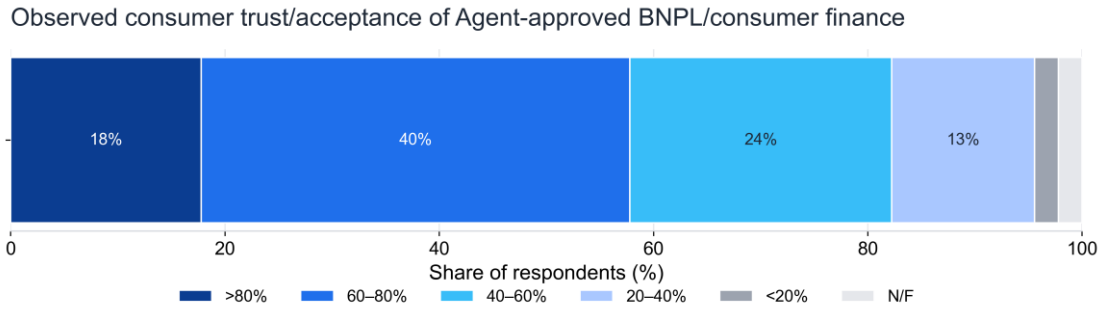


Figure 32. The leading option is "Relatively trusting/acceptance 60–80%", at 40.0%.

Interpretation: Consumer trust in Agent approval is net positive but not strong: 58% have acceptance of 60% or higher (very trusting 18%, relatively trusting 40%), 24% are neutral, and 16% are untrusting. A demand-side base exists. The 24% neutral

group is the main conversion target for trust building; their stance is most sensitive to approval transparency, error-correction speed, and dispute-handling experience. Combined with Q33, explainable approval reasons and connection to credit-reporting systems are the two direct levers for raising that group's acceptance. In the framework of Section 1.2.4 of the main report, consumer trust is the direct determinant of the market-expansion effect in BNPL: each step up in trust expands the serviceable customer boundary. 58% acceptance means the feasible market for Agent approval is still bounded by trust, not by technical capability. The 16% untrusting group should not be ignored. Under social-media amplification, a few negative experiences spread cheaply, and their reputational impact can exceed their headcount share. That is a scene-level footnote to the reputation-manipulation resistance and sentiment monitoring emphasized in Section 5.2.3 of the main report.

Q33. Most critical elements for Agent creditworthiness in BNPL/consumer finance

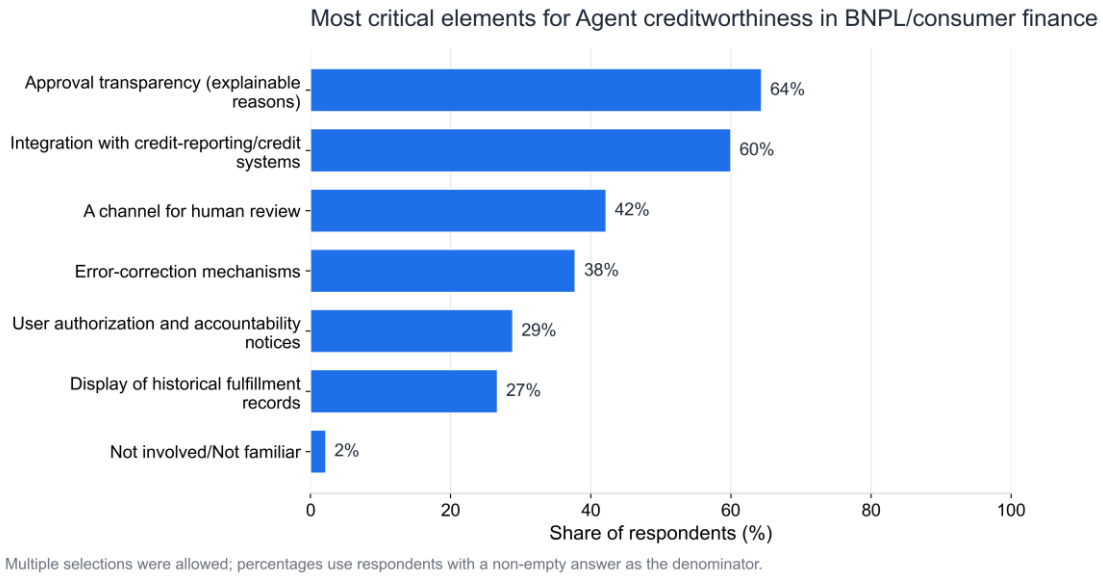


Figure 33. Top multiple-choice items: "Approval transparency (explainable reasons)" 64.4%; "Integration with credit-reporting/credit systems" 60.0%; "A channel for human review" 42.2%. Percentages may sum to more than 100% because multiple selections were allowed.

Interpretation: Among the critical elements of Agent creditworthiness, approval transparency with explainable reasons (64%) and integration with credit-reporting/credit systems (60%) lead, followed by a human-review channel (42%), error-correction mechanisms (38%), authorization and accountability notices (29%), and display of fulfillment records (27%). The first two correspond to "the decision can be understood" and "credit has an external anchor," and they are the two pillars of consumer-finance KYA: explainability solves trust in a single decision, and credit-bureau connection solves the Agent's source of credit as a decision-maker. Human review and error correction rank in the middle, but both are critical-threshold requirements in the sense of Section 5.2.5 of the main report: not bonus points, but entry conditions. An approval system without a review channel cannot make good on a transparency promise. Authorization and accountability notices at 29% rank lower, reflecting a practice that attends more to decision outcomes than to process

disclosure. Under "authorization as data" in Section 3.3.2 of the main report, the notice record itself is key evidence when a dispute arises, and building it costs far less than proving the case after the fact.

3.6 Optional Module: SMEs / Micro-Business Customers

This module observes adoption, concerns, and simplification priorities in SME and micro-business settings.

Q34. Proportion of SME/micro-business customers served

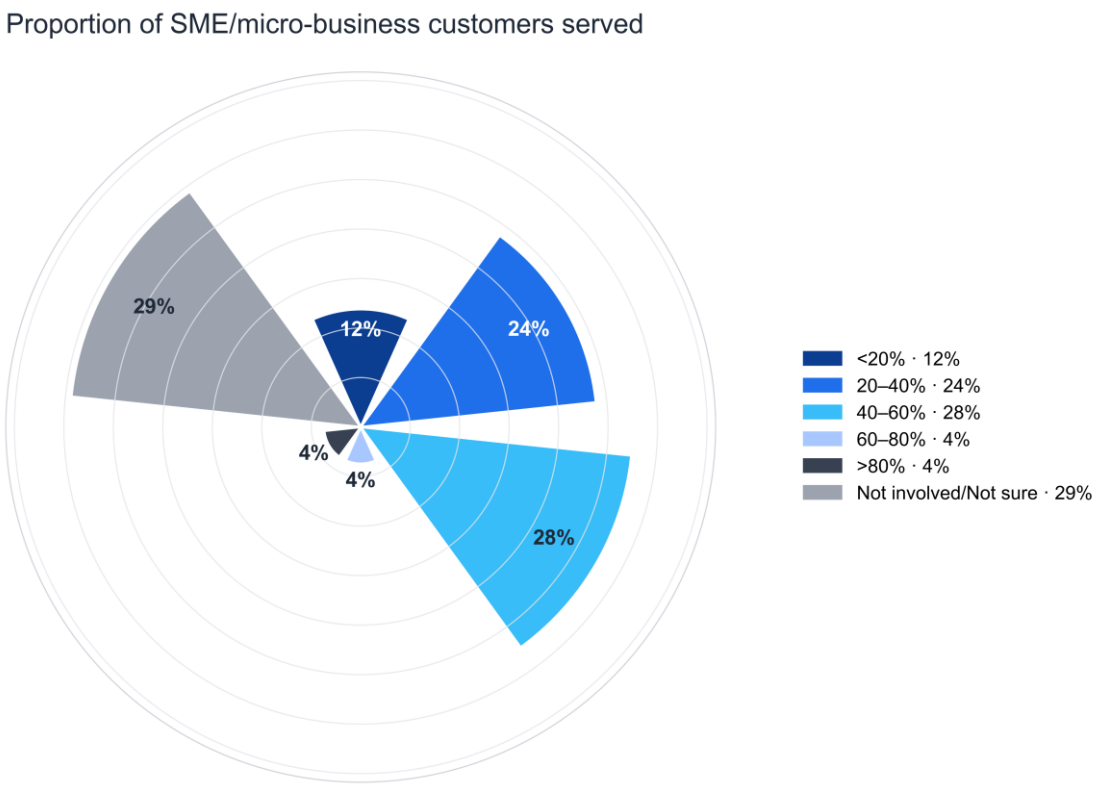


Figure 34. The leading option is "Not involved/Not sure", at 29.4%.

Interpretation: The share of SME/micro-business customers served by respondents' organizations is fairly even: 40–60% accounts for 28%, 20–40% for 24%, under 20% for 12%, and above 60% only 7% combined. SMEs are a common but usually non-dominant customer group. That structure means SME credit demand is easily ignored by standardized products built around large customers. SMEs are also the group with the most severe credit-information asymmetry: informal financials, thin collateral, short credit histories, and limited coverage by traditional credit bureaus. There is a clear gap in credit services for their Agent-use scenes. From the α+β hybrid-credit view in Section 3.2.4 of the main report, SMEs are exactly the segment where hybrid credit has the most value: when subject credit (α) is weak, task-level fulfillment records (β) can accumulate credit over time, and Agent-based services naturally produce structured fulfillment data. The evenness of the share distribution also shows that SME credit solutions are not a niche line for a few institutions; they are a general capability most service organizations will have to face.

Q35. Approximate proportion of observed SME/micro-business customers using Agents

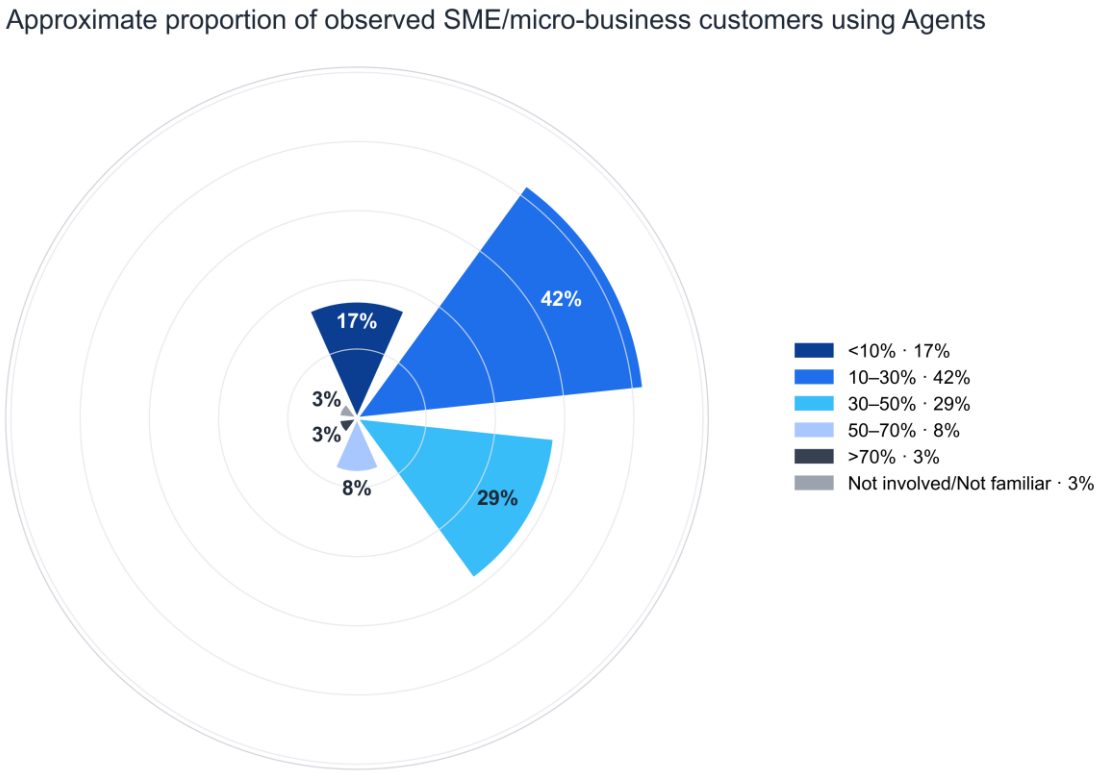


Figure 35. The leading option is "10–30%", at 41.6%.

Interpretation: Among respondents with SME customers, Agent usage by micro-businesses concentrates in the 10–50% band (10–30% at 42%, 30–50% at 29%); only 10% exceed 50%. Penetration is still early. Low penetration is directly tied to the cost and cognitive barriers in Q36: when both understanding cost and implementation cost are high, micro-businesses delay adoption and form a low-use / low-data / low-credit loop. From a cold-start view, this stage most needs reusable identity, authorization, and logging capabilities on the platform side, so each small firm does not have to rebuild them. That is the core of the starting path in Section 5.3.4 of the main report: first meet the four minimums of identity attribution, genuine fulfillment, authorization execution, and exception handling, so that key facts are recorded from day one. 10–30% is the densest band, which means most SME Agent use has only just passed the trial stage. Record quality formed in this period will set the starting point when they later apply for credit services, so the value of platform intervention in record infrastructure is highest now.

Q36. Main credit-related concerns when SME/micro-business customers use Agents

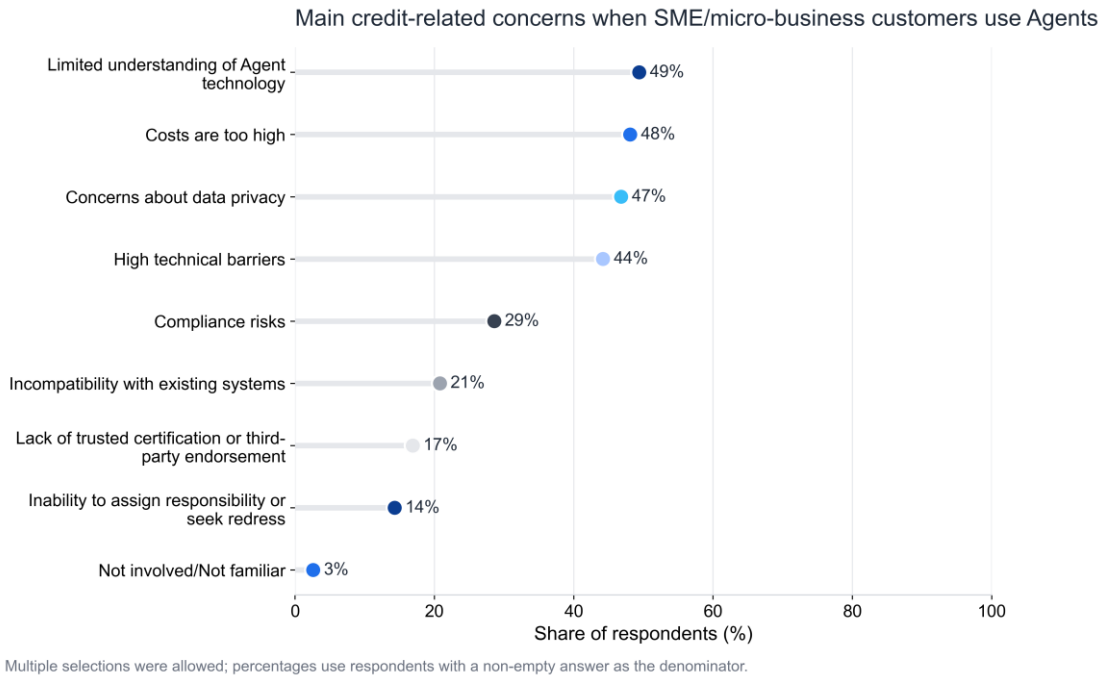


Figure 36. Top multiple-choice items: "Limited understanding of Agent technology" 49.4%; "Costs are too high" 48.1%; "Concerns about data privacy" 46.8%. Percentages may sum to more than 100% because multiple selections were allowed.

Interpretation: The top four credit-related concerns when SMEs use Agents are limited technical understanding (49%), high cost (48%), data-privacy worries (47%), and technical barriers (44%), with compliance risk (29%) and system incompatibility (21%) next. Cognition and cost form a double barrier and rank above institutional concerns. That is clearly different from the large-enterprise sample, where accountability and compliance lead, and it shows that the main contradiction in the SME market is still access, not the fineness of governance. The ranking has direct product implications: an SME version of a credit solution should subtract first, using templates and platforms to lower understanding and implementation cost. What is simplified is process, not the threshold; control requirements on already-opened actions cannot be waived. The starting checklist in Section 5.3.4 of the main report is designed on that logic: identity attribution, genuine fulfillment, authorization execution, and exception handling cannot be cut, but each can be delivered as a platform default. Privacy worries at 47% are close to half, which means a platform solution must also solve trust in data custody; otherwise lower cost will not buy adoption.

Q37. Availability of simplified Agent credit solutions for micro-businesses (< $150,000 annual revenue)

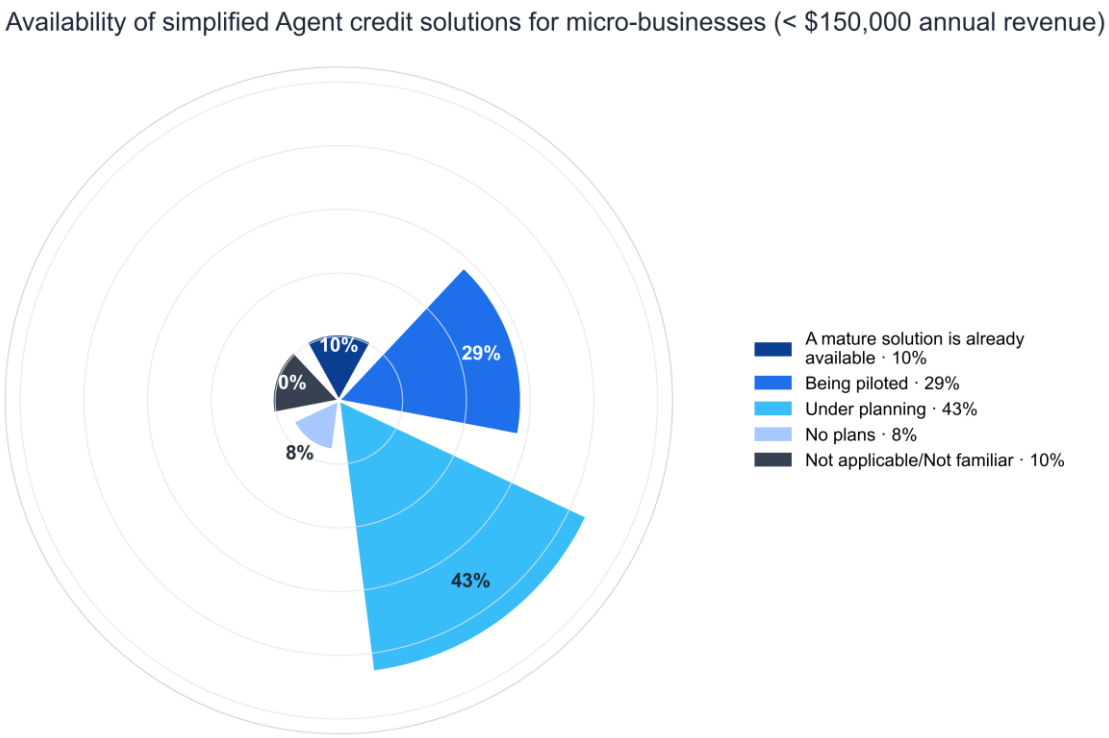


Figure 37. The leading option is "Under planning", at 42.9%.

Interpretation: For simplified Agent credit solutions aimed at micro-businesses with annual revenue below USD 150,000, 43% are under planning, 29% are being piloted, only 10% already have a mature solution, and 10% say not applicable or not familiar. Supply is clearly short. Micro-business credit demand has long existed, but service cost is high: small ticket sizes, rigid due-diligence cost, and uneconomic default handling make traditional models hard to cover. If Agent-based services can reuse records the platform already has, marginal service cost can be brought into a commercializable range. A 10% mature-supply rate means the market is still early; first movers can help define de facto standards. Against the four cold-start paths in Section 3.2.5 of the main report, guaranteed limits and subject anchoring are the realistic starting point for this segment: the platform provides an initial credit anchor from existing transaction data, and the Agent's task-fulfillment records then accumulate into independent credit. A 43% planning share foreshadows a rapid supply increase over the next one to two years. The planning window is the best time to embed record standards and authorization templates; once a solution is live, retrofitting a recording mechanism will cost much more.

Q38. Applicable scenarios for Agent credit solutions for micro-businesses

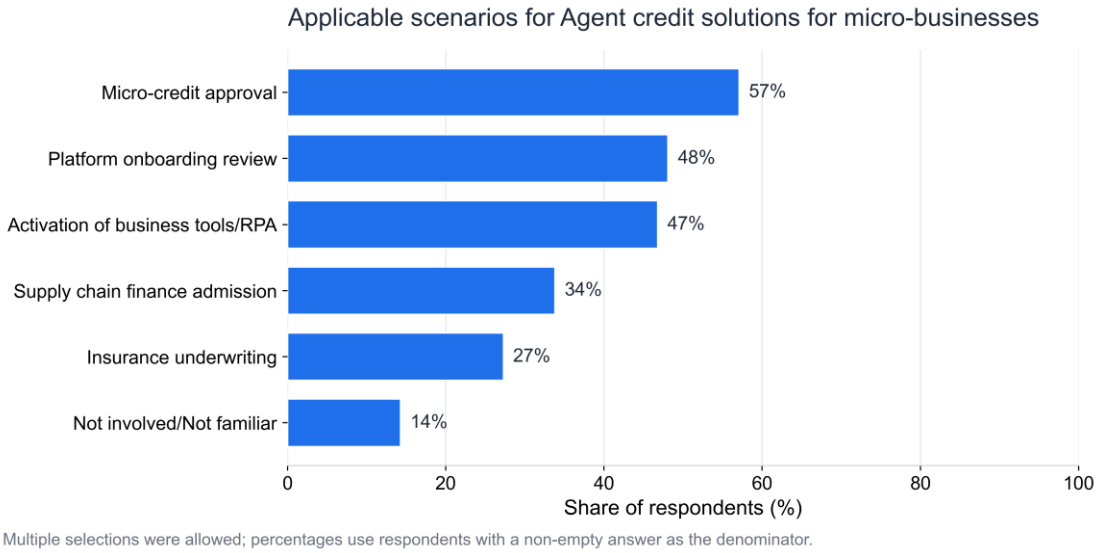


Figure 38. Top multiple-choice items: "Micro-credit approval" 57.1%; "Platform onboarding review" 48.1%; "Activation of business tools/RPA" 46.8%. Percentages may sum to more than 100% because multiple selections were allowed.

Interpretation: Applicable scenes for micro-business Agent credit solutions are led by micro-credit approval (57%), platform onboarding review (48%), and activation of business tools/RPA (47%), with supply-chain-finance admission (34%) and insurance underwriting (27%) next. The first three are admission-type judgments: clear task boundaries, verifiable outcomes, and bounded single-ticket risk. They match the sandbox-task features in Section 3.2.5 of the main report and are suitable as the first scenes for a credit cold start. Accumulating fulfillment samples in admission scenes, then expanding into money scenes, is a risk-controllable sequence. Micro-credit approval involves money, but the "micro" attribute bounds single-ticket loss; with limit controls it can meet the low-risk authorization conditions in Section 3.3.3 of the main report, so its first-place rank does not conflict with risk tiering. Supply-chain finance (34%) and insurance underwriting (27%) rank lower mainly because they involve multi-party data verification and cross-organization delegation chains. Under the delegation-chain inspection requirements in Section 5.2.4 of the main report, their credit assessment is more complex and is better expanded after single-scene validation.

Q39. Parts of an Agent credit solution that should be simplified first for SMEs/micro-businesses

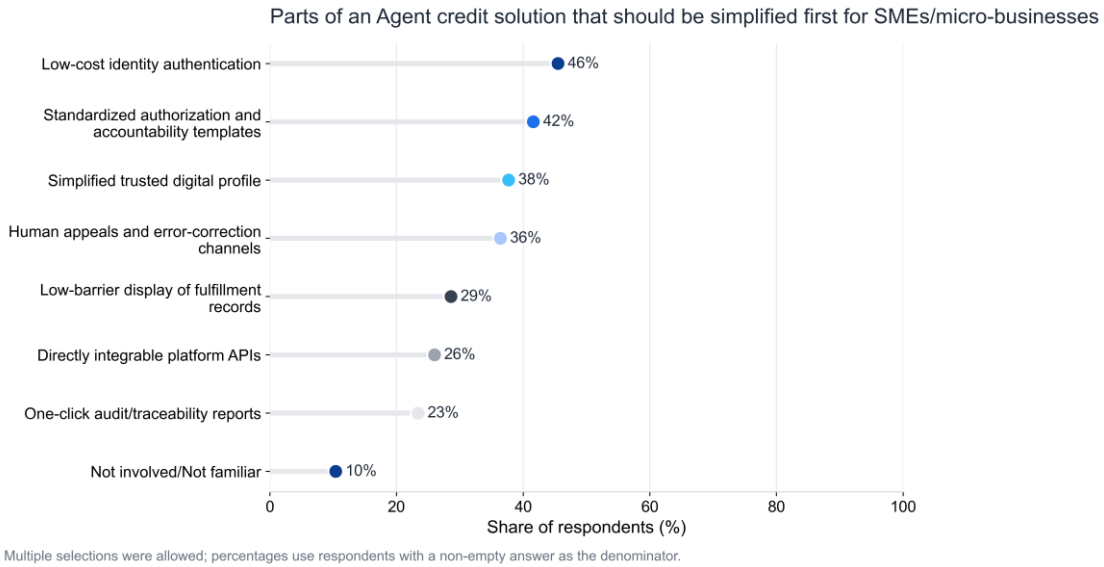


Figure 39. Top multiple-choice items: "Low-cost identity authentication" 45.5%; "Standardized authorization and accountability templates" 41.6%; "Simplified trusted digital profile" 37.7%. Percentages may sum to more than 100% because multiple selections were allowed.

Interpretation: The parts that should be simplified first are low-cost identity authentication (46%), standardized authorization and accountability templates (42%), simplified trusted digital profiles (38%), and human appeal and error-correction channels (36%). The first three are lightweight versions of KYA's three dimensions: verifiable identity attribution, clear authorization boundaries, and accumulable fulfillment records. They map item by item onto the three-dimension framework in Chapter 3 of the main report, which means SME scenes do not need another framework, but a low-cost implementation of the same one. Combined with the cost and cognitive barriers in Q36, the SME breakthrough is not a lower assessment standard, but packaging the three capabilities as out-of-the-box templates so that key facts are recorded from day one. The starting checklist in Section 5.3.4 of the main report—which Agent, what task, what authorization, what outcome, who intervened—is the minimal form of that template. Appeal and error-correction channels at 36% entering the top four is noteworthy: SMEs care as much about what happens after an error as about how not to err. The existence of a correction channel is itself a source of trust, and it should be built in parallel with the other three, not later.

3.7 Scale and Performance Metrics

This chapter provides the interval-based scale and performance profile needed for the development index in Chapter 6 of the main report. The bands follow the original questionnaire design and do not require exact values.

Q40. Scale of Agent deployment

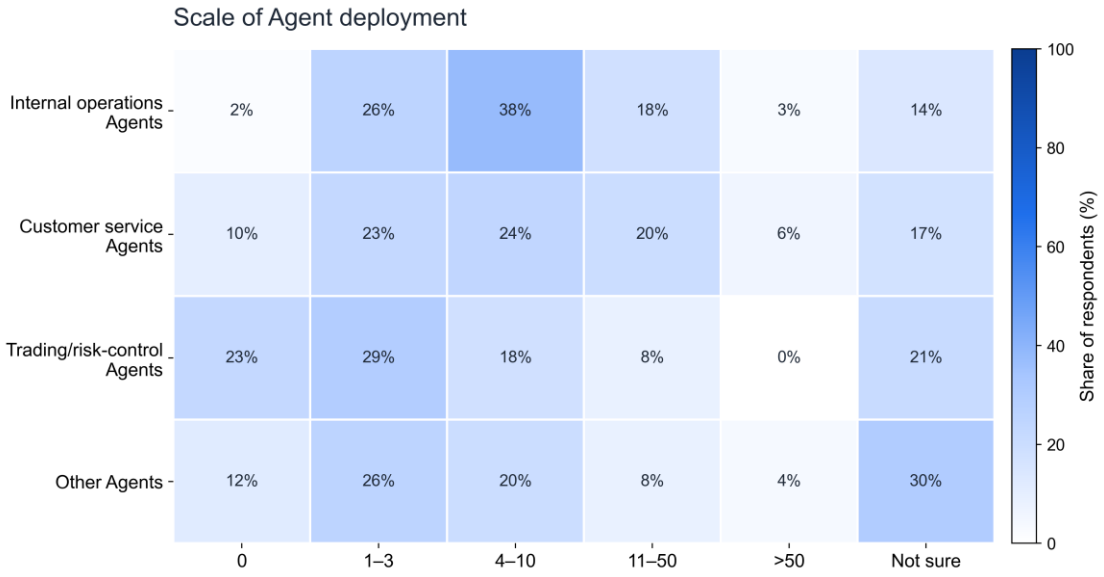


Figure 40. Darker cells indicate a higher share in that deployment-size band; the four Agent types show clearly different scale distributions.

Interpretation: Deployment scale splits by Agent type. Internal-operations Agents are the most widely deployed: the 4–10 band is 38%, and not deployed is only 2%. Customer-service Agents are next, with 26% at 11 or more. Trading/risk-control Agents are the most conservative: 23% not deployed, and only 8% at 11 or more. Scale declines as business risk rises, with a clear gradient that matches the risk logic behind dynamic downgrade in Section 2.8.3 of the main report: organizations are willing to expand authorization in low-risk scenes and stay cautious in money-related ones. The split also confirms the real-world basis of the risk-tiering matrix in Section 3.3.3: organizations are already adjusting Agent density by scene risk; what they lack is making that implicit practice explicit as verifiable authorization records. For credit-system building, dense deployment of internal-operations and customer-service Agents means large volumes of fulfillment data are being produced but not creditized. Low penetration in trading/risk-control foreshadows the largest room for authorization expansion once the trust-evidence problem is solved. The scale data can serve as interval inputs to the “application breadth” dimension of a later development index.

Q41.1 Agent performance metrics — Business automation rate

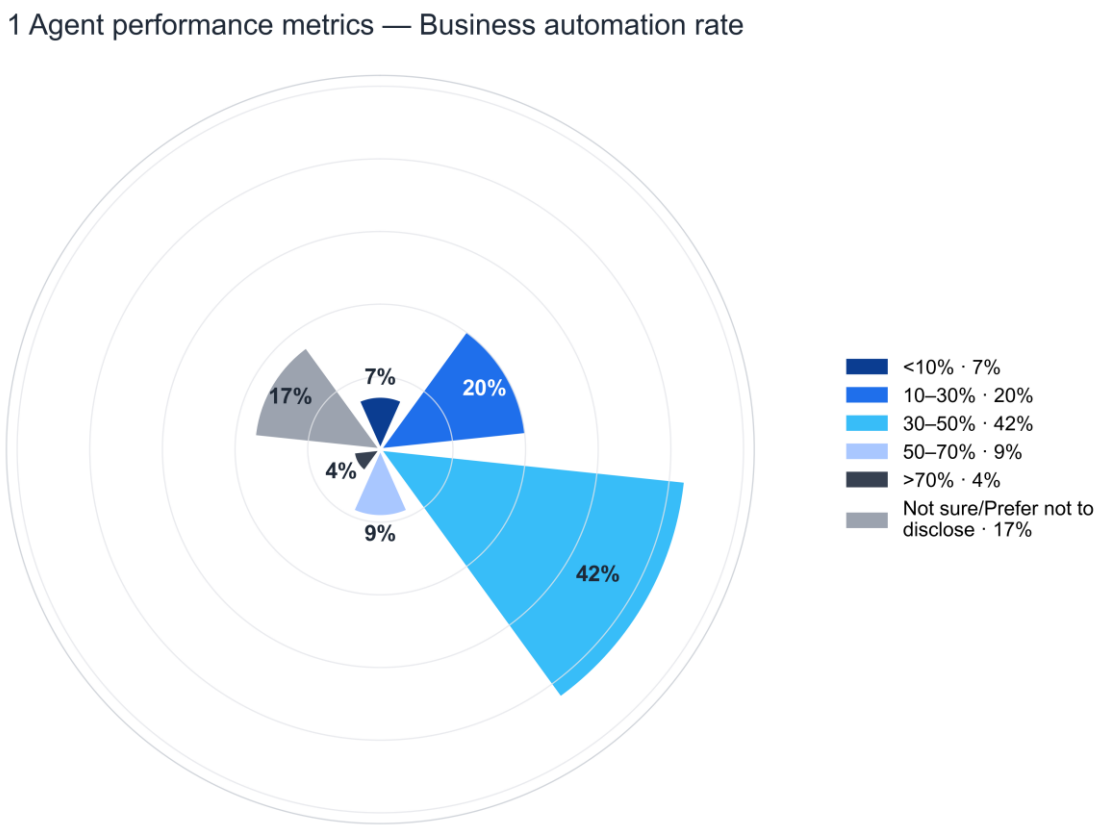


Figure 41. The leading option is "30–50%", at 42.2%.

Interpretation: Business automation rates concentrate in the 30–50% band (42%), with 10–30% at 20% and only 13% combined above 50%. Agents already carry substantive workload, but most organizations have not crossed the "half of the business automated" threshold. Automation depth determines the intensity of credit demand. When Agents handle only edge tasks, trust issues can be absorbed by human process. When automation exceeds half, a human backstop is no longer cost-feasible, and a credit system shifts from an optional investment to an operating necessity. The present distribution shows most organizations sitting just before that conversion. Read against Q14 maturity, a 30–50% automation rate lines up structurally with the L3 hybrid-credit peak (38%): humans still participate deeply as backup takeover, and automation has not yet become autonomy. Under the upgrade loop in Section 2.1.2 of the main report, moving further up from this band does not first require a stronger model; it requires accumulated fulfillment records and tiered authorization. That is where a credit system should intervene.

Q41.2 Agent performance metrics — Agent decision accuracy

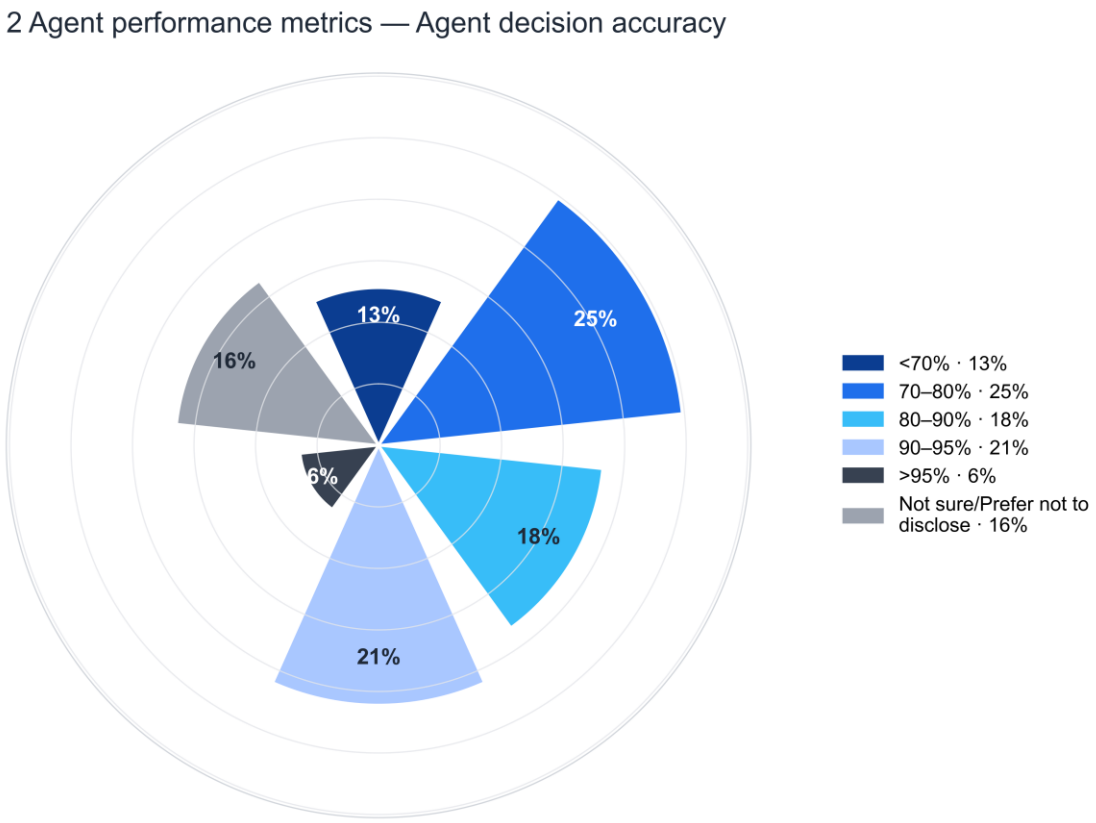


Figure 42. The leading option is "70–80%", at 24.8%.

Interpretation: Agent decision accuracy is dispersed: 70–80% accounts for 25%, 90–95% for 21%, 80–90% for 18%, above 95% only 6%, and 13% below 70%. Against the requirements of money and high-risk scenes, current accuracy cannot support high-grade autonomy, which is directly related to L4-and-above remaining below 15% in Q14. Accuracy is an outcome metric. Under Section 5.1.2 of the main report, outcome metrics must be read together with control metrics; a short run of smooth operation cannot replace active verification of interception and error-correction mechanisms. An Agent that is 95% accurate but cannot intercept unauthorized actions is riskier than one that is 90% accurate with complete controls. Dispersion itself is information: accuracy runs continuously from below 70% to above 95%, which means differences in task type, data quality, and engineering investment are widening gaps across organizations. One value of a unified assessment standard is to make that gap comparable and priceable. The 13% low-accuracy group should be treated as excluded from high-risk-scene admission.

Q41.3 Agent performance metrics — Human intervention rate

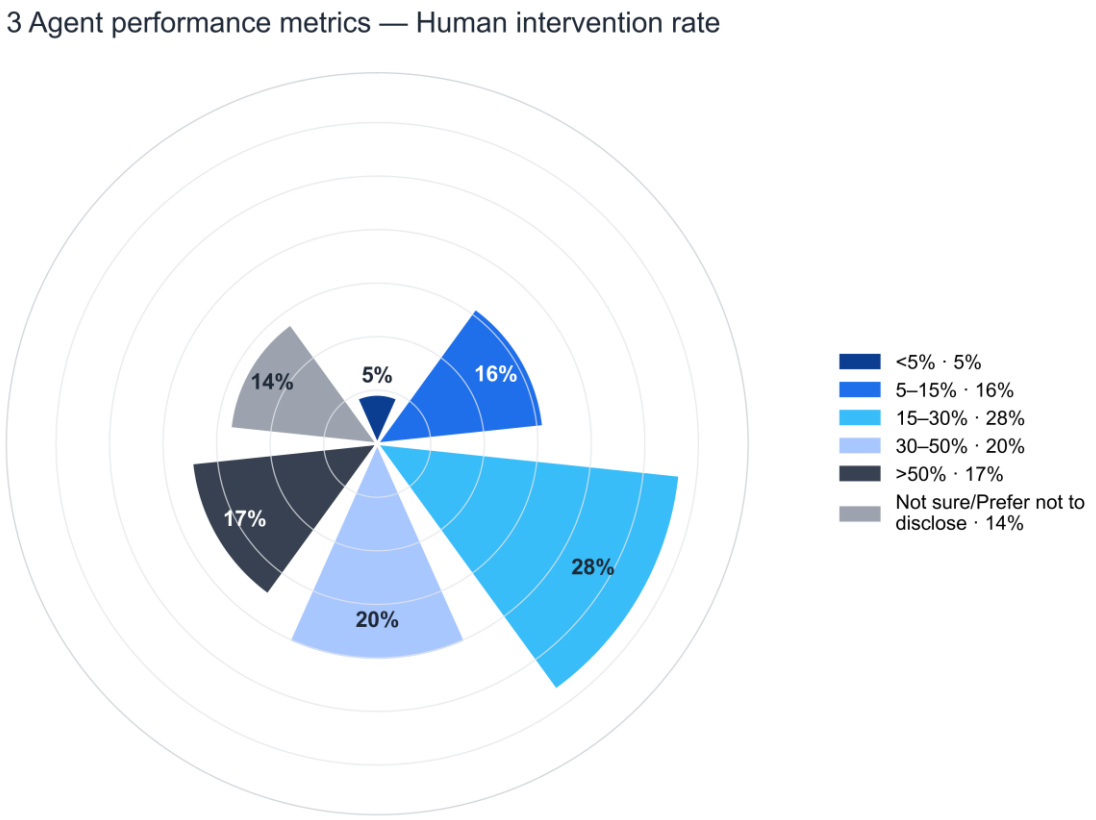


Figure 43. The leading option is "15–30%", at 28.4%.

Interpretation: Human intervention rates are mainly 15–30% (28%) and 30–50% (20%); above 50% is 17%, and below 5% only 5%. Most Agents in fact run in a state that needs frequent human involvement, which matches L1–L2, and "nominal autonomy, actual human watching" is common. Human involvement is not itself a negative signal. Section 3.2.3 of the main report lists human-intervention events as part of the minimum event set; the time, reason, and handling of an intervention are valid data for calibrating authorization. The question is whether intervention events are recorded and used to calibrate authorization. High intervention plus low recording is the real gap in credit building: the organization pays for human supervision but does not accumulate evidence that could lower future supervision intensity. On cost structure, an intervention rate above 30% means one in three items still needs human handling, and labor saved by automation is largely clawed back by supervisory labor. That maps directly onto the Q8 pain of "more human confirmation, lower efficiency" (38%). Turning intervention events into credit data is the key to breaking that cost loop.

Q41.4 Agent performance metrics — Average response latency

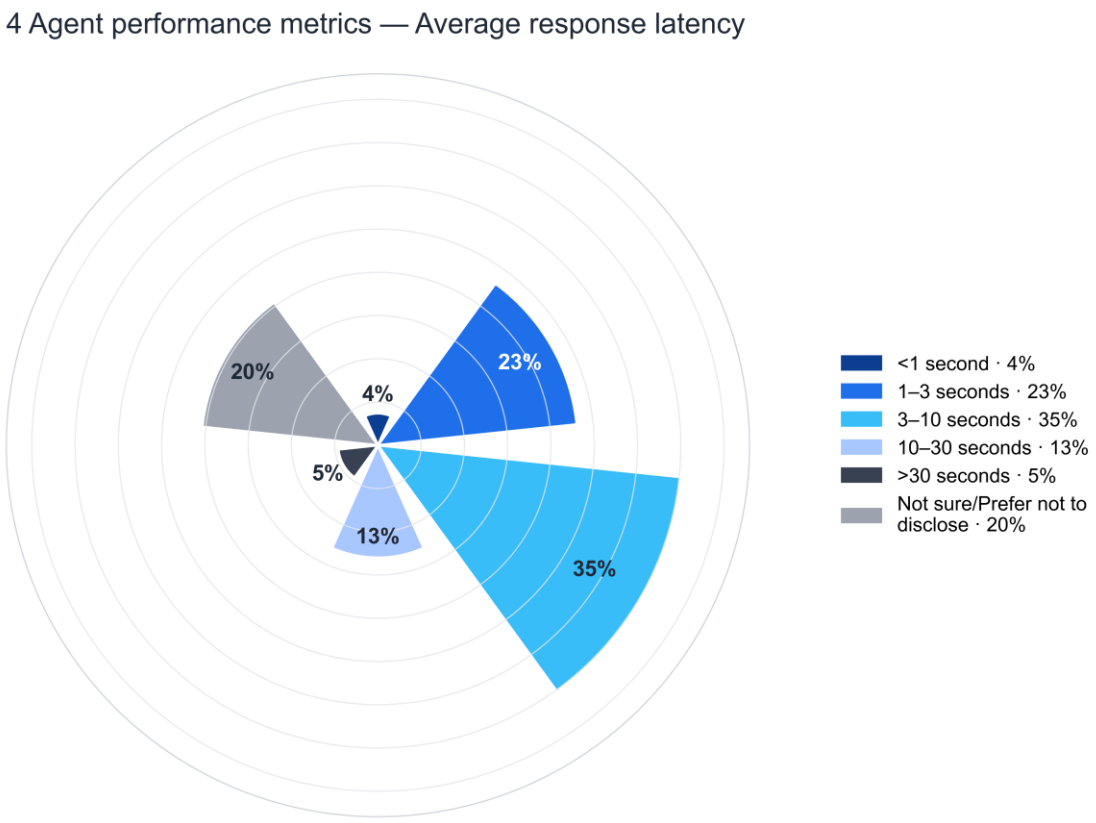


Figure 44. The leading option is "3–10 seconds", at 35.2%.

Interpretation: Response latency is mainly 3–10 seconds (35%) and 1–3 seconds (23%); 18% combined sit above 10 seconds. Latency is overall in an acceptable range, and performance is not the main source of the present trust bottleneck. Read with Q41.2 and Q41.3, resource priority should be accuracy and intervention mechanisms, not response speed. For the assessment system, performance metrics can stay as a basic threshold, while decision quality, unauthorized-action interception, and exception handling should carry more weight. That matches the principle in Section 5.1.2 of the main report of combining outcome and control metrics: latency is an experience-type outcome metric, while interception rates and review-trigger rates are safety-type control metrics, and the latter contribute more at the margin to trust. The 3–10 second mainstream band also has an operational meaning: it leaves a time window for human intervention. Under the emergency mechanisms in Section 3.3.6 of the main report, emergency stop and circuit-breaking need to take effect before the Agent finishes the action. Second-level latency means real-time interception is engineering-feasible, so latency data can serve as an auxiliary input when judging "intervenability" in authorization grading.

3.8 Willingness to Follow Up (Anonymous)

Only the anonymous distribution of willingness to be featured as a case is shown. No names, organizations, emails, or phone numbers are displayed.

Follow-up. Willingness to be featured as a case in the research report

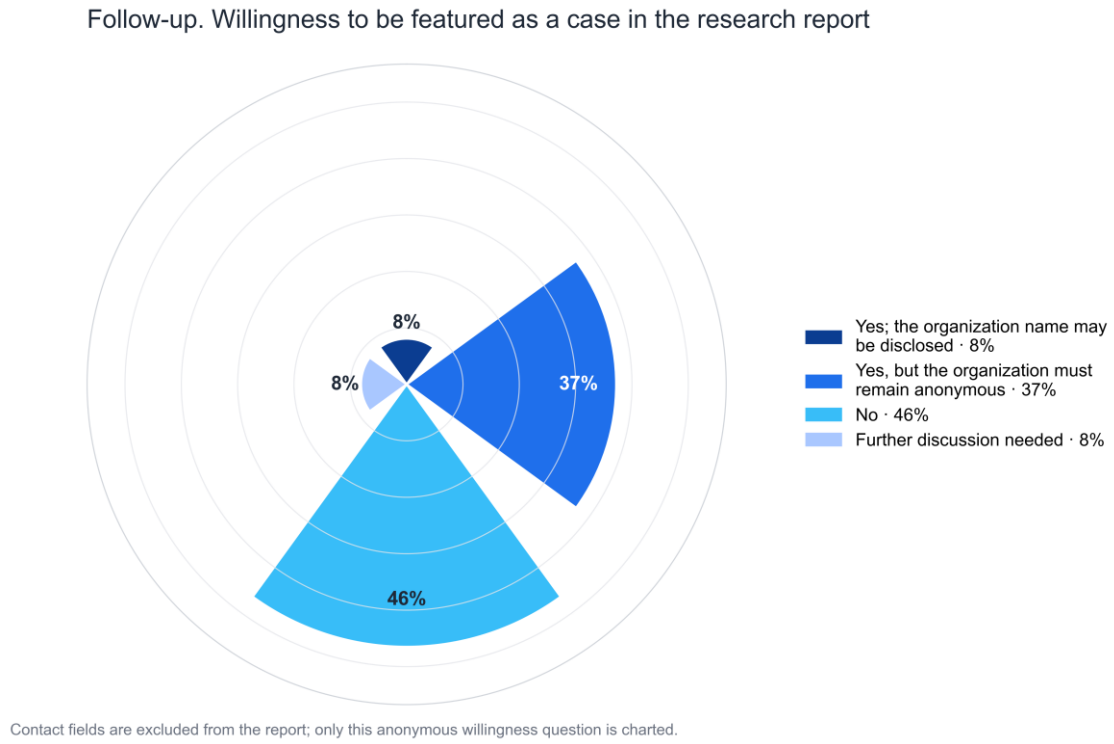


Figure 45. The leading option is "No", at 46.5%.

Interpretation: Willingness to be featured as a case is cautious overall: 46% are unwilling, 37% are willing only if the organization stays anonymous, only 8% would allow the name to be disclosed, and 8% need further discussion. That suggests Agent capability is still treated more as internal operating information than as a credit asset that can be shown outside. One goal of building a credit system is to change that: when trustworthy records are verifiable and portable, displaying fulfillment capability becomes a competitive strategy. As the credit flywheel in Section 2.8.2 of the main report describes, displayable credit records attract larger authorization and more trades, forming a positive loop. A 37% anonymous-sharing willingness is a constructive middle signal: firms do not reject letting practice enter an industry knowledge base; they reject uncontrolled reputational exposure. That supplies a feasible path for an industry case library: collect fulfillment and incident data in anonymized, intervalized form, first solve "is there data," then raise the named share as the trust environment improves. The interview-invitation design in this report's appendix is based on that preference structure.

4. Integrated Conclusions

4.1 Empirical Support for Chapter 1: The New Trust Challenge

- The existence of a trust vacuum is supported quantitatively: 76% of respondents encountered trust, security, accountability, or compliance concerns in the past year, and only 2% never did. The costs first show up as rising compliance and audit pressure (41%), delayed launch or expansion (40%), and efficiency loss from human confirmation (38%), corresponding to monitoring/enforcement cost in the

transaction-cost framework of Section 1.2.4. 23% have already been refused by customers or partners, which shows that missing trust has begun to produce direct market contraction.

- The structure of distrust is a dual gap of capability and institution. Data security and privacy (76%) and hallucinations/unreliable output (62%) sit at the capability layer; unauthorized action (43%) and unclear accountability (40%) sit at the institutional layer. The institutional gap is the increment KYA must cover beyond traditional KYC. KYC answers who the customer is; an Agent scene must also answer who is acting for whom, whether the action exceeded authorization, and who bears the loss if it goes wrong.

- The commercialization constraint has already switched from capability to trust. Low customer and partner trust (53%) outranks ROI calculation (34%) and system integration (35%) as the leading commercial pain point. Value-creating capability is no longer the main constraint; the trust mechanism is.

4.2 Empirical Support for Chapter 2: The Maturity Model

- Industry maturity is spindle-shaped: L3 is 38%, L1 and L2 together are 45%, and L4 and above remain below 15%. Most organizations sit in the transition from human-anchored credit to task autonomy and have not yet moved from a human backstop to authorization based on records.

- Performance data explain the maturity ceiling: decision accuracy above 95% is only 6%, and human intervention above 30% is 37%. Until accuracy and intervention mechanisms improve, the L3-to-L4 jump lacks an engineering base, and independent credit will remain scarce.

- Upgrade obstacles concentrate on authorization and accountability: security risk (60%), authorization boundaries (51%), and unclear accountability (46%) lead, while lack of external standards is only 7%. The credit flywheel stalls at “control establishment”. The constraint is internal engineering capacity, not the supply of external standards. Organizations that first build internal authorization and accountability capability will gain a first-mover advantage.

4.3 Empirical Support for Chapter 3: KYA’s Three Dimensions

- Build-out of the three dimensions is uneven. Identity-trustworthiness elements sit in a narrow 60%–69% implementation band and have started relatively evenly. Creditworthiness is strong on records and weak on scoring, with logs at 81% and internal credit scores at only 52%. Authorization is strong on backstops and weak on tiering, with human confirmation at 86% and scenario-based trust levels at only 52%.

- Authorization trustworthiness is named the weakest dimension by 39% of respondents, first among the three. Capability data and subjective judgment agree: the industry has generally built point-wise human gates, but lacks systematic risk-tiered authorization. When authorization boundaries are unclear, attributable identity and recorded fulfillment still cannot convert into larger authorization.

- Accountability has an institutional vacuum: only 17% have a defined process, and 30% have no stable mechanism. Lifecycle management is heavy on launch and light on operations, with weak coverage of periodic reassessment and retirement. Accountability is the responsibility anchor of credit records and a prerequisite for dynamic assessment and dynamic downgrade. It should be the first link institutionalized in the three-dimension system.

4.4 Demand-Side Evidence for Chapter 5: The Credit Assessment Framework

- Demand for standards is near consensus: 88% see a need for unified assessment and certification standards, and 39% see it as extremely needed. Framework priorities are security and compliance baselines (77%), auditability and traceability (62%), and trusted identity registration (60%), with interoperability protocols last at 14%. A progressive path of baseline trust first and interconnection later is supported by the data.

- The demand ranking of design principles matches Section 5.1.2: accountability traceability (93% scoring 4–5), machine-readability (92%), and scenario classification (91%) occupy the top three. Responsibility and evidence are the primary contradiction at this stage; neutrality of technical path ranks relatively lower.

- Adoption path and support demand are clear: 60% would adopt proactively or first pilot in high-risk scenes. The external support most needed is industry standards (71%), testing benchmarks and tools (54%), and certification services (53%). Standards, tools, and services need to advance together, and threshold-type requirements must be paired with independent testing and verification.

4.5 Scenario-Module Conclusions (BNPL and SME)

- BNPL/consumer finance: Agents have entered core steps such as credit assessment (58%), risk-control approval (53%), and limit decisions (49%), and their output enters the formation of asset quality directly. Versus humans, cost (69%) and efficiency (62%) advantages are clear; customer satisfaction is the most divided (49% better, 16% worse). Consumer acceptance is net positive (58%). Explainable approval and credit-bureau connection are the two trust levers; human review and error correction are threshold-type requirements.

- SME/micro-business: penetration concentrates in the 10–50% band, and concerns are led by cognitive (49%) and cost (48%) barriers. Mature-solution supply for firms with annual revenue below USD 150,000 is only 10%. Priority landing points are admission scenes with clear boundaries (micro-credit 57%, onboarding review 48%). The product direction is to package the three capabilities as lightweight templates: low-cost identity authentication (46%), standardized authorization and accountability templates (42%), and simplified trusted digital profiles (38%).

4.6 Overall Judgment and Action Recommendations

- The first priority is to engineer authorization trustworthiness. Authorization boundaries, dynamic limits, human-confirmation thresholds, emergency stop,

and unauthorized-action blocking should become default capabilities that are verifiable and auditable, moving gradually from point-wise human gates to scenario-based tiered authorization.

- Creditworthiness needs to move from the recording layer to the evaluation layer: first unify the record schema for key decisions, inputs/outputs, tool calls, anomalies, and human interventions, then build internal scoring and cross-system integration.
- Standards and assessment design should hold to accountability traceability, machine-readability, scenario classification, and progressive evolution. Commercialization should enter through high-risk-scene pilots, while building the three external supports of standards, testing benchmarks, and certification services in parallel.
- On scene strategy, BNPL should first fill in explainable approval, human review, error-correction mechanisms, and display of fulfillment records. SME should first provide low-cost identity authentication, simplified trusted digital profiles, and standardized authorization and accountability templates.
- Over the longer run, improvements in accuracy and intervention mechanisms will open room for an L3-to-L4 jump. When trustworthy records are verifiable and portable, Agent credit will shift from an internal management tool to an asset that can be shown outside. Only 8% of organizations are currently willing to share a named case; later movement in that share can serve as an observational indicator of industry trust maturity.

Appendix：Notes on Open-Ended Suggestion Themes

Q28 is an open-ended question. The report does not quote individually identifiable wording. It reports keyword-theme counts only. One response may hit multiple themes. Empty or non-substantive answers (such as N/A or none) have been excluded.

Illustrative suggestions (anonymized; only content that does not point to a personal identity is retained):

- There must be multiple company guidelines and designated people who are responsible.

- Agents need to log their reasoning steps so we can see why an action was taken.

- Keep the standards simple and practical. We need basic, open APIs so different platforms can verify agents easily, without too much overhead.

- There should be clear transparency around an agent's identity, data sources, and permissions. There should also be a clear way to revoke an agent's permissions when its behavior is not as expected.

- Treat agent identity like short-lived tokens; do not give agents static API keys or permanent credentials.

- Talent training and more support from enterprises.

- This would require moving from static API keys to dynamic behavioral governance, cryptographic identity, and continuous auditing.

- Digital certificates, a centralized registry, and token validation.

Copyright and Disclaimer

This report is a research output of the KYA Research Team at The University of Hong Kong. The judgments and views expressed in the report represent only the position of the research team and do not constitute any product commitment, business advice, or investment advice from any institution.

The industry data and case information cited in this report come from public sources, enterprise interviews, and the research team's own compilation, and every effort has been made to attribute the sources; any omissions or discrepancies are welcome to be pointed out.

The content of this report will be revised on an ongoing basis. Without the permission of the research team, no organization or individual may reproduce or excerpt the content of this report in any form; please cite the source when quoting.