From Super-Apps to Agent Economies: Delegated AI Requires Transaction Closure

Chaoyue He *, Xin Zhou, Di Wang, Hong Xu, Wei Liu, Chunyan Miao

Posted Date: 27 April 2026

doi: 10.20944/preprints202604.1860.v1

Keywords: agent economies; delegated AI; transaction closure; action closure; super-app ecosystems; agentic AI; contestable transaction closure; AI governance; ClosureBench; typed mandate; receipt graph

 doi:10.20944/preprints202604.1860.v1

Article

From Super-Apps to Agent Economies: Delegated AI Requires Transaction Closure

Chaoyue He [1,*], Xin Zhou [1], Di Wan [1], Hong Xu [1], Wei Liu [2] and Chunyan Miao [1]

[1] Alibaba–NTU Global e-Sustainability CorpLab (ANGEL), Singapore
[2] Alibaba Group, Hangzhou, China
* Correspondence: cyhe@ntu.edu.sg

Abstract

This position paper argues that delegated AI must be evaluated and governed by **transaction closure**, rather than mere task completion. Agent economies emerge when AI systems execute externally binding commitments—such as purchases, bookings, or procurement orders—on behalf of authorized principals. Safe delegation at this boundary requires environments that support bounded mandates, reversible lifecycle states, verifiable receipts, and contestable failures. While AI-native platforms rapidly advance *action closure*, mature super-app ecosystems and enterprise suites currently offer the clearest testbeds for the payment, refund, and liability semantics required for true transaction closure. This work contributes a formal definition of transaction closure, minimal schema objects for transaction-ready delegation, a comparative ecosystem analysis, a principal-facing mandate-card model, and ClosureBench—a benchmark design for evaluating authorization, split states, recovery, and portability. The long-run goal is contestable transaction closure: portable infrastructure where delegated commitments can be inspected, challenged, and repaired beyond any single host.

Keywords: agent economies; delegated AI; transaction closure; action closure; super-app ecosystems; agentic AI; contestable transaction closure; AI governance; ClosureBench; typed mandate; receipt graph

1. Introduction

Delegated AI requires transaction closure. Agent economies will be defined not by models that browse, reason, or call tools well, but by environments where agents can safely create, reverse, explain, and repair externally binding commitments on behalf of authorized principals.

This paper uses *agent economies* in the plural because delegated commitment will not emerge as one uniform market. Consumer super-apps, enterprise procurement systems, public-service portals, cloud and software operations, AI-native work surfaces, and protocol-native agent markets each expose different authority, settlement, receipt, and recovery problems. The common unit of analysis is the *delegated commitment*: an action that can bind a principal to consequences outside the conversation.

Consider four examples. A traveler may ask an agent to book a refundable flight, arrange ground transport, stay below a budget, cancel if the itinerary changes, and preserve proof for reimbursement. A procurement officer may ask an agent to compare vendor quotes, issue a purchase order within policy, track delivery, and dispute non-performance. A cloud operator may authorize an agent to provision compute under a budget cap, roll back failed deployments, and preserve audit logs. A public-service applicant may authorize an agent to complete a form, pay a fee, schedule an appointment, and escalate if the case is rejected. In each case, the challenge is not finding a webpage or invoking a tool. The agent must interpret a bounded mandate, compare options, reserve scarce capacity, commit or withhold payment, react to drift, preserve receipts, cancel or reverse failed commitments, and escalate when constraints conflict. Figure 1 illustrates this continuum from single-task apps to mature agent economies, while Table 1 provides the definitions for key concepts discussed throughout this work.

 doi:10.20944/preprints202604.1860.v1

Figure 1. Argument map. The structural continuum from traditional apps to agent economies. While integration creates super-app ecosystems and AI delegation enables transaction-entry surfaces, viable agent economies emerge only when environments guarantee completion, transaction, and recovery closure for delegated commitments.

Table 1. Key-terms definitions.

Key term	Definition used in this paper	Source status
AI agent / agentic AI	A goal-directed AI system that can plan, use tools, interact with environments, and act under some level of autonomy or oversight.	Aligns with recent agent reports and governance taxonomies [1–6].
Agent economies	Repeated settings in which AI agents, humans, organizations, services, platforms, or other agents coordinate over authority, commitments, settlement, reversibility, receipts, liability, and redress.	Our substrate-agnostic use of agent-economy, virtual-agent-economy, and governance discussions [7–12].
Super-app ecosystem	A hosted multi-service environment with shared or partially shared identity, discovery, payment, messaging, complementor participation, and host governance, often through mini-apps or adjacent service modules.	Based on the super-app and digital transaction-platform literature [13–18].
Transaction-entry surface	A strategic entry surface that aims to become the default starting point for delegated workflows, regardless of whether it began as a super-app, social network, enterprise suite, operating-system layer, protocol market, or AI-native work surface.	Product and strategy category; public examples include Grab, Paytm, X, and AI-native trajectories [19–22].
Delegated commitment	An externally binding action taken by an agent under a principal's mandate, such as a purchase, booking, official submission, procurement order, cloud-resource action, cancellation, refund request, subscription change, service-level claim, or dispute.	Proposed term, grounded in delegation, identity, payment, and principal-agent work [23–26].
Transaction closure	The environment-side ability to represent bounded authority, move commitments through reversible states, preserve evidence, and route failures into recovery.	Proposed term; motivated by agent-infrastructure, authenticated-delegation, and payment-policy research [23–25,27,28].
Typed mandate	A machine-readable delegation object specifying principal, agent, task class, budget or resource cap, time window, service scope, data scope, revocation, and escalation.	Proposed object; grounded in authenticated delegation and identity work [23,24,29,30].
Receipt graph	A linked trace connecting mandate, action, payment or service state, communication, outcome, and recovery events.	Proposed object; inspired by provenance and audit-trace concepts [27,31,32].
Contestable transaction closure	Transaction closure in which mandates, receipts, and unresolved case state can be inspected, challenged, and exported beyond one host.	Proposed governance requirement; motivated by portability, privacy, and competition concerns [33–36].
Transaction-ready ecosystem	A digital environment that can carry state across services, represent scoped authority, settle and reverse commitments, preserve evidence, allocate responsibility, and route failures into redress.	Proposed synthesis; grounded in agent infrastructure, platform ecosystem, and digital-infrastructure work [27,37,38].

This transition is critical as machine learning moves from content generation toward goal-directed interaction in environments. Benchmarks such as AgentBench, Mind2Web, WebArena, OSWorld, AndroidWorld, and MobileWorld evaluate long-horizon behavior across websites, operating systems, and mobile apps, while evaluation work increasingly warns that deployment requires metrics beyond task success [39–45]. Those benchmarks are necessary, but they still understate the environment problem. Planning does not inherently create budget caps, resource limits, settlement callbacks, refunds, dispute pathways, liability allocation, or proof of what the principal authorized.

2. Transaction Closure: A Formal Object for Agent Evaluation

Here, *transaction* extends beyond payment. A workflow is transactional whenever the agent can create or modify an external commitment: buying something, reserving capacity, scheduling transport, paying a fee, filing a form, issuing a purchase order, provisioning paid compute, entering a service contract, cancelling a booking, initiating a refund, or disputing a failed commitment. Payment serves as the prime example because its authorization and commitment boundary is especially visible, but the broader research problem centers on externally committing action.

Definition 1 (transaction closure). For a delegated workflow w, principal mandate m, agent a, and environment E, E has *transaction closure* for (w, m, a) if it can: (i) represent the principal's authority and constraints in a machine-usable mandate; (ii) transition external commitments through explicit

 doi:10.20944/preprints202604.1860.v1

lifecycle states; (iii) link authorization, action, payment or service state, communication, and outcome into a receipt graph; (iv) support cancellation, refund, dispute, compensation, or human escalation when the workflow fails; and (v) export or hand off the unresolved case without losing authority state or evidence. A minimal lifecycle, defined in Equation 1, is

$$\mathcal{L} = \{\textsc{Proposed}, \textsc{Held}, \textsc{Authorized}, \textsc{Committed}, \textsc{Fulfilled}, \textsc{Cancelled}, \textsc{Refunded}, \textsc{Disputed}, \textsc{Compensated}, \textsc{Escalated}\}. \quad (1)$$

Crucially, this definition centers the environment. The same model can be safe in a high-closure environment and unsafe in a thin one.

Definition 2 (contestable transaction closure). Transaction closure is *contestable* when the mandate, lifecycle state, receipt graph, and unresolved case state can be inspected, challenged, and exported without forcing the principal to remain inside one host's private evidence system. Contestability remains the target: closure should increase recoverability without turning recoverability into lock-in.

Environment maturity is therefore essential independently of model quality (Table 2). Better reasoning reduces navigation and planning errors, but it does not generate settlement callbacks, refund semantics, dispute logs, or portable receipts. Recent research on delegation, identity, payments, and trajectory-level evaluation points to the same requirement: safe deployment requires infrastructure around the model, not only inside it [23–25,27,28,32,46].

Table 2. Three closures required for agent economies. The central term is *transaction closure* because it names the point where delegated action creates external commitments.

Closure	Plain meaning	Typical failure without it
Completion closure (action closure)	Can the agent finish the whole workflow rather than only the first step?	Search succeeds, but identity or state is lost across services.
Transaction closure (commitment closure)	Can the environment safely create, change, and reverse external commitments?	Payment succeeds but reservation fails; cancellation or refund becomes manual.
Recovery closure (accountability closure)	Can people later see what happened, prove authorization, and repair failure?	No linked receipt, no proof of mandate, and no clear escalation path.

Operationally, a transaction-ready environment must represent what the principal authorized, which options the agent saw, when an action is merely proposed versus actually committed, which evidence links the workflow together, who can repair the failure, and whether the case can move outside the host (Figure 2). Some of these objects already exist informally inside mature super-app ecosystems. The research challenge is to make them explicit, measurable, auditable, privacy-preserving, and portable, as summarized by the minimal schemas in Table 3.

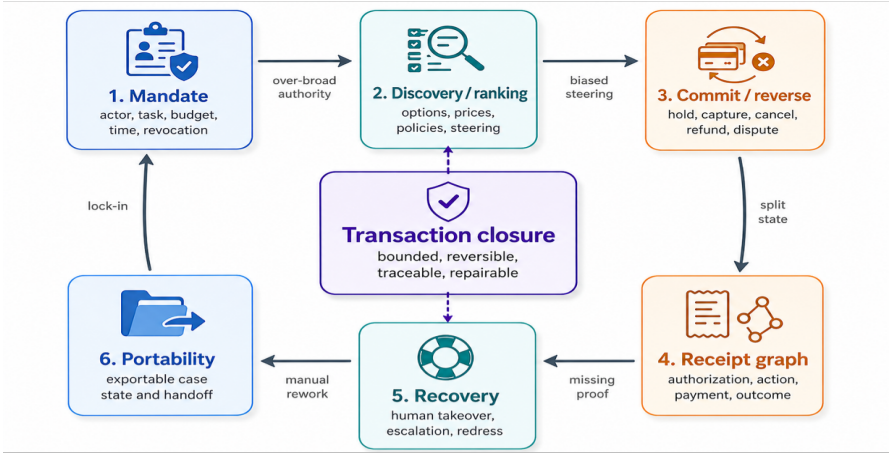


Figure 2. The transaction-closure lifecycle. The same agent can look competent or unsafe depending on whether the environment represents mandates, rankings, commitment states, receipts, recovery, and portability explicitly.

 doi:10.20944/preprints202604.1860.v1

Table 3. Minimal objects for transaction-ready delegation. The schemas are deliberately lightweight: they specify what the community should be able to test, standardize, sign, export, or audit without requiring one implementation.

Object	Minimal schema-like fields	What must be testable
Mandate object	`mandate_id`, `principal_id`, `agent_id`, `task_class`, `budget_or_resource_cap`, `time_window`, `service_scope`, `data_scope`, `confirmation_policy`, `escalation_rule`, `revocation_endpoint`.	Whether the agent acts only within bounded authority and stops or asks again when the mandate envelope is crossed.
Commitment lifecycle	`proposed`, `held`, `authorized`, `committed`, `fulfilled`, `cancelled`, `refunded`, `disputed`, `compensated`, `escalated`, with idempotency and rollback metadata.	Whether split states such as payment-without-booking or cancellation-without-refund are detected and repaired.
Receipt graph	`mandate_id`, `action_ids`, `tool_calls`, `service_ids`, `payment_ids`, `messages`, `timestamps`, `state_transitions`, `evidence_hashes`, `recovery_status`.	Whether humans, auditors, and downstream agents can reconstruct what was authorized, attempted, committed, reversed, or left unresolved.
Handoff packet	`case_id`, `unresolved_branch`, `current_authority_status`, `evidence_bundle`, `responsible_party`, `next_safe_action`, `expiry`.	Whether recovery can move to a support worker, merchant, regulator, or another agent without restarting from screenshots.
Policy surface	`ranking_policy`, `merchant_admission`, `fallback_policy`, `dispute_rules`, `retention_policy`, `export_rules`, `audit_interface`.	Whether closure is contestable rather than hidden inside host-controlled ranking, support, or evidence systems.

The mandate object functions as a signed policy or capability: it is not merely a UI consent event, but a reusable authority object that downstream services can validate, narrow, revoke, and record. The receipt graph is similarly not just a log; it acts as the evidence structure that lets a human, merchant, auditor, or later agent determine whether a commitment stayed within mandate.

Historically, general-purpose technologies create large value only after complementary organizational and infrastructural investments are made [38,47–49]. The same principle applies here. The complements for agentic AI are not solely memory, tool APIs, or stronger planners. They include bounded mandates, settlement rails, machine-readable receipts, exception routing, liability allocation, and ecosystem governance. Consequently, identifying viable agent economies requires analyzing which substrates already provide these complements under realistic pressure from consumers, firms, institutions, and machine counterparties.

2.1. Why Super-Apps and AI-Native Agents Converge on the Same Bottleneck

Mature super-app ecosystems are positioned to set early consumer-facing rules for agent economies by combining three forces rarely present together. First, they possess *reach*: critical edge-case failures like merchant refusals, identity mismatches, and cross-border incompatibilities only become visible across large heterogeneous populations. Second, they feature *transaction depth*: their workflows natively encompass payment, order state, refunds, and support. Third, they exhibit *governance concentration*: hosts can define ranking, admission, and dispute policies at the ecosystem level. This combination normalizes de facto rules before formal standards emerge [11,48,50]. We treat super-apps not as the desired endpoint, but as a high-volume stress test for the broader transaction-closure problem.

Ecosystems are converging on this requirement from diverse origins. Weixin/WeChat illustrates the communication-and-mini-program path enabling high-volume delegation; Grab and Paytm illustrate multi-service and payments-first paths [19,20,51–53]. LINE/PayPay, Kakao, GoTo, and PhonePe extend this pattern across messaging, payments, and everyday services [54–60]. X illustrates a social-native transaction-entry ambition through Grok and X Money, though currently exhibiting less hosted service depth than mature super-app ecosystems [61–63].

AI-native agents serve as an action-side stress test, demonstrating that this convergence can originate from intelligence rather than payments. Codex, for instance, evolved from a cloud-based software-engineering agent into a broad desktop orchestrator featuring computer use, browsing, memory, and parallel task execution [64–67]. This exemplifies strong action closure: the system plans, operates tools, and produces outcomes without constant manual intervention. However, while Codex is highly capable within repositories or sandboxes, it highlights how quickly action closure can mature before payment, contract, liability, and dispute frameworks exist around the agent.

Many workflows in such systems lack transaction closure. Externally binding transactions introduce payment capture, counterparty selection, cancellation windows, identity checks, and rights. If a pure AI-native surface gains a wallet, merchant network, typed mandates, reversible commitment lifecycles, dispute routing, and portable handoff objects, it could evolve into a powerful transaction-

 doi:10.20944/preprints202604.1860.v1

ready agent-economy substrate. This outcome validates the premise that closure, rather than historical origin, determines readiness. Table 4 details these cases and their missing closure elements.

Table 4. Worked examples. AI-native tools highlight the dynamic: strong delegated action remains insufficient unless the environment supports commitment, reversal, receipts, and portability.

Case	What already works	Missing closure	Core dynamic
AI-native action systems (e.g., Codex)	Strong action closure for software work: parallel agents, repository context, desktop orchestration, computer use, browsing, memory, plugins, and ongoing tasks.	Payment or contract authority, counterparty identity, cancellation and refund callbacks, dispute routing, linked receipts, and liability semantics.	AI-native agents become agent-economy substrates only after acquiring transaction closure.
AI-native transaction-entry candidates	Future AI-native surfaces could become the default goal-entry interface across work, commerce, travel, forms, procurement, operations, and daily coordination.	Typed mandates, reversible commitment lifecycles, receipt graphs, audit export, and portability safeguards.	If AI-native systems gain these layers, they validate the premise: closure, not historical origin, determines rule-setting power.
Mature super-app ecosystems	Shared communication, mini-program, payment, merchant, support, and refund layers make high-volume consumer delegation plausible.	More explicit mandates, agent-readable receipts, privacy-preserving audit export, and stronger contestability.	Super-apps are strong early testbeds for the consumer-facing slice of agent economies, not the only target architecture.
Enterprise suites / procurement stacks	Role-based identity, approvals, budgets, procurement records, audit logs, and administrative controls already exist in many organizational systems.	Cross-vendor settlement callbacks, external dispute portability, agent-specific authority objects, and handoff packets recognized outside the firm.	Enterprise environments may define organizational agent economies around procurement, compliance, and operations.
Protocol-native agent markets	MCP-like tool access and A2A-like coordination can expose services and compose agents across hosts.	Mandate semantics, settlement lifecycle, liability allocation, receipt graphs, revocation propagation, and dispute procedures.	Open protocols can overtake integrated hosts if they achieve equal closure with stronger portability.
X / Grok / X Money and social-native challengers	Social distribution, AI assistance, messaging, and payment ambition create an agent-economy path from a different starting point.	Hosted service depth, counterparty ecosystem maturity, refund and dispute semantics, and transaction-volume evidence.	Social-native challengers can compete by accumulating the same closure layers as transaction-heavy ecosystems.

Rule-setting in this domain is operational as well as legal. A high-volume ecosystem sets *technical rules* through API contracts and lifecycle states, *commercial rules* through ranking and support policies, and *legal-facing rules* through the evidence available for disputes or compliance audits. Repeated delegated commitments produce stable mandate and settlement semantics, which then become entrenched through user expectations and counterparty integration, shaping interoperability requirements for future challengers.

2.2. Comparative Landscape, Risks, and Alternative Architectures

Figure 3 summarizes the comparative landscape. The distinction lies between *action closure* and *transaction closure*. AI-native systems traverse the action axis quickly; mature super-app ecosystems often begin farther along the transaction axis for consumer workflows; enterprise suites may begin farther along identity and audit axes; and protocol-native markets may advance portability first. Agent economies require both.

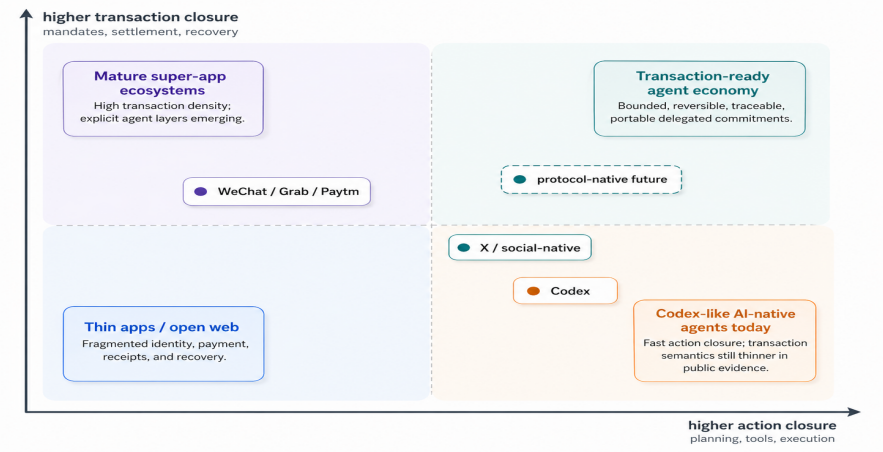


Figure 3. Action closure and transaction closure operate on separate axes. High-volume agent-economy rule-setting will emerge where both axes are integrated.

Public evidence aligns with this directional trajectory. Tencent reports extensive Weixin/WeChat scale alongside high Weixin Pay transaction density; Grab, Paytm, GoTo, LINE/PayPay, Kakao, and PhonePe exhibit related combinations of communication, payments, and emerging AI layers [51–60,68–

 doi:10.20944/preprints202604.1860.v1

72]. Ecosystems like OpenAI's and X demonstrate that AI-native and social-native challengers can rapidly accumulate action and distribution [22,61–63,67,73,74]. Appendix Figure A1 and Table A1 outline the public data and conversion assumptions.

Maturity encompasses more than scale. Raw distribution does not dictate what an agent is permitted to commit, how a partially completed workflow rolls back, or which evidence survives manual intervention. Super-app ecosystems are not the sole plausible substrate, but they remain among the few environments where consumer discovery, settlement, and support pressures operate simultaneously at scale. Enterprise suites and public-service portals expose parallel pressures around role authority and compliance.

Normatively, environments must distinguish between *good closure* (bounded, reversible, auditable, portable) and *bad closure* (opaque, irreversible, locked-in, surveillant). Platform research documents leakage, inference, and self-preferencing risks at large scale, while agentic-commerce security literature highlights that failures propagate across reasoning, authorization, and compliance layers [33–36,75–80]. These concerns reinforce the necessity for portability rather than weakening the closure requirement. Table 5 details these operational implications.

Table 5. Good and bad closure. Contestable transaction closure, rather than platform control, serves as the target framework across consumer, enterprise, public-service, and agent-to-agent domains.

Closure type	Operational meaning	Example in delegated transaction
Good closure	Mandates are scoped; commitments are reversible where possible; receipts are machine-readable; recovery can be exported; privacy exposure is minimized.	Agent books a refundable trip under a visible budget, preserves a receipt graph, cancels after disruption, and exports a dispute packet.
Bad closure	Authority is broad; ranking is hidden; commitments are hard to undo; receipts remain host-private; recovery requires staying inside one platform.	Agent completes a purchase through opaque steering, cannot export evidence, and leaves the principal dependent on host support.

Threat model.

Transaction closure must address more than model errors. Adversaries and systemic risks include malicious merchants manipulating inventory, compromised connectors exfiltrating credentials, prompt-injection attacks, dishonest hosts steering users, buggy non-idempotent retries, and inadvertent over-delegation. Closure mitigates these harms only if mandates adhere to least-privilege principles, revocations propagate effectively, receipt graphs remain tamper-evident, and handoff packets provide adequate evidence without overexposing sensitive traces [75–77,80–83]. Table 6 outlines why common architectural alternatives do not eliminate the transaction-closure bottleneck.

Table 6. Common architectural alternatives and their relationship to the transaction-closure bottleneck.

Architecture / Objection	Significance	Resolution
Protocol-first	MCP and A2A establish interoperability as a core engineering framework [84–86].	Tool and agent protocols still require mandate, settlement, receipt, dispute, revocation, and liability layers. If they achieve comparable closure alongside superior portability, they become the preferred substrate.
Model-first	Enhanced models directly reduce navigation and planning errors.	Superior reasoning does not instantiate budget envelopes, resource caps, reversal semantics, linked receipts, or human takeover interfaces.
Enterprise-first	Business suites frequently feature robust identity, logging, procurement rules, and administrative controls.	This is not a refutation but a parallel route: enterprise systems may define organizational agent economies, while super-apps remain strong consumer-facing testbeds. ClosureBench should compare both rather than assume one universal substrate.
Privacy / competition	Extensive closure can devolve into surveillance, lock-in, and self-preferencing.	Privacy, portability, ranking audits, and contestability represent structural maturity requirements, not extraneous caveats.
Personal-agent future	On-device agents and secure data vaults may eventually secure the control layer.	Localized control still relies on shared commitment and recovery infrastructure to process counterparties, payments, contracts, refunds, reversals, audits, and disputes.

The necessary transition is from thin task-completion benchmarks to portable transaction-closure infrastructure. The community must extract useful primitives from environments where delegated AI first becomes operational and embed them into open layers: typed mandates, commitment lifecycles, receipt graphs, and auditable handoff packets [23,24,27,84–87].

2.3. Principal Control and Benchmarkable Contestability

A transaction-ready environment requires a principal-facing control model. Typed mandates remain inaccessible if people must read policy code before acting, and they remain incomplete if organizational principals cannot express role, budget, or compliance constraints. The interface must

 doi:10.20944/preprints202604.1860.v1

present control surfaces corresponding to actual delegation failures: permitted actions, requisite re-authorization triggers, termination mechanisms, and task routing upon failure. Transaction closure directly intersects with mixed-initiative interaction: principals require sufficient control to prevent over-delegation without introducing interruptions that cripple the workflow [88–92]. Table 7 details these control surfaces.

Table 7. Principal-facing controls for transaction closure. Bounded delegation must be understandable, interruptible, and recoverable.

Control surface	Interface function	Benchmark requirement
Mandate card	Goal, budget, service scope, data scope, deadline, fallback rules, and actions requiring renewed approval.	Evaluation of whether principals can accurately predict agent permissions and whether the agent honors those limits.
Escalation triggers	Conditions demanding re-authorization: price drift, risky substitution, non-refundable commitments, identity mismatches, policy conflicts, or failed partner APIs.	Evaluation of whether the agent interrupts only when necessary while avoiding silent overreach.
Revocation switch	Mechanisms to pause, narrow, or cancel authority during active background execution.	Evaluation of whether downstream tool calls and commitments definitively halt or roll back upon revocation.
Receipt digest	Explanations of authorizations, completed actions, financial commitments, and unresolved branches.	Evaluation of whether humans can reconstruct case events without parsing raw execution logs.
Dispute packet	Exportable evidence bundles for support routing, merchant escalation, regulatory complaints, or inter-agent transfer.	Evaluation of whether recovery proceeds outside the original host environment while preserving mandate and receipt state.

A practical consumer mandate card might state: "Book a refundable flight to Bangkok on April 30, total below $400, cancel automatically if ticketing is not confirmed within 10 minutes, and ask me if the price changes by more than $20." An enterprise mandate might state: "Renew the database-support contract only if the annual price stays below the approved budget, the vendor maintains the required service level, and procurement receives the receipt graph." A receipt digest following a failure might read: "Agent compared 12 fares, selected Flight X, requested a hold, payment succeeded, ticketing failed, and the refund completed after two hours." This principal-facing operationalization of the object model in Table 3 renders authority and recovery legible without exposing raw logs.

This establishes a practical distinction between consent and control. A singular up-front confirmation provides inadequate control if the agent encounters subsequent price increases or refund delays. Conversely, demanding confirmation prior to every tool call reduces autonomy to manual browsing. The appropriate framework is staged autonomy: the principal authorizes a typed mandate, the agent operates within it, and the environment mandates renewed approval when commitments exceed the mandate envelope.

Effective evaluations require workflow families rather than isolated tasks. A comprehensive travel, commerce, or procurement workflow can be executed under identical mandates across a super-app ecosystem, an AI-native surface, an enterprise suite, and a protocol-native market. Comparative analysis should hold the planner constant while varying the underlying substrate. Table 8 outlines the benchmark families facilitating this evaluation.

These dimensions provide falsifiable benchmark criteria. If an AI-native surface employing open protocols matches super-app ecosystems across these families absent centralized host control, the super-app advantage diminishes. Conversely, if a mature super-app fails significantly on revocation, refund processing, or receipt export, it cannot be classified as transaction-ready. Closure, rather than brand identity, forms the metric of evaluation.

 doi:10.20944/preprints202604.1860.v1

Table 8. Workflow families for ClosureBench. The benchmark generalizes beyond consumer tasks to organizational, public-service, operational, and agent-to-agent commitments while preserving realistic recovery failures.

Benchmark family	Delegated workflow	Closure stressors
Travel and mobility	Book a refundable trip under a budget, arrange local transport, react to delay, cancel or rebook if constraints break, preserve receipts for reimbursement.	Price drift, non-refundable fares, booking-payment split, identity mismatch, disruption recovery, refund latency.
Commerce and returns	Find an eligible product, apply coupon or loyalty credit, choose delivery, handle stockout, manage failed delivery, return item, and track refund.	Substitution policy, duplicate execution, delivery dispute, refund degradation, evidence continuity across merchant and logistics.
Public-service filing	Complete a form, verify identity, book appointment, pay fee, submit evidence, receive confirmation, and escalate if the case is rejected.	Identity proofing, policy exception, irreversible submission, proof-of-submission, human takeover burden.
Enterprise procurement	Compare vendors, check policy, issue a purchase order, monitor delivery, reconcile invoice, and dispute non-performance.	Role authority, budget caps, approval chains, vendor accountability, audit evidence, contract reversal.
Cloud and software operations	Provision compute, call paid APIs, change infrastructure, roll back failed actions, and preserve incident or compliance evidence.	Resource-cost overrun, non-idempotent retries, permission scope, rollback, service-level claims, audit trace continuity.
Subscription and contract management	Compare plans, cancel, renew, or downgrade within a deadline, avoid unwanted charges, confirm reversal, and export evidence.	Time-window authority, cancellation proof, recurring-payment state, renewal terms, dispute route after failed cancellation.
Agent-to-agent service markets	A user-side or firm-side agent contracts with another agent for data, compute, logistics, verification, or support under scoped authority.	Counterparty identity, machine-readable terms, settlement, service proof, dispute routing, replay resistance, liability.

3. Alternative Views

While several alternative perspectives challenge an integrated-substrate approach, they become persuasive only when accounting for mandate, commitment, receipt, recovery, liability, and portability semantics.

View 1: Stronger models will make transaction-specific infrastructure unnecessary.

A model-first view argues that improvements in planning, tool use, multimodal perception, and long-horizon reasoning will let AI agents safely complete workflows over existing web and mobile interfaces [39–44,93,94]. **Response.** Better models reduce navigation and interpretation errors, but they do not by themselves create scoped spending authority, resource caps, idempotent commit/cancel states, refund callbacks, contract-reversal semantics, or exportable evidence. Those remain environment properties; agent evaluations that ignore them risk mistaking successful browsing for safe delegated commitment [27,32,45].

View 2: Open protocols will bypass super-app ecosystems.

A protocol-first view holds that MCP-like tool exposure and A2A-like coordination will let agents compose services directly, making integrated hosts less important [84–86]. **Response.** This view is directionally compatible with the paper's normative goal: portable protocols are preferable when they achieve equal or better closure. The current gap is that tool and agent interoperability must still acquire mandate objects, settlement callbacks, receipt graphs, dispute routes, and liability semantics before they can substitute for high-closure transaction environments [23–25,87].

View 3: Personal agents and local data vaults should be the primary substrate.

A user-control view argues that agent economies should start from on-device agents, private user or organizational profiles, and narrowly shared data rather than from host ecosystems, because integrated platforms create surveillance and lock-in risks [77–79,95]. **Response.** Local control is essential for privacy containment, but it does not remove the need for shared commitment infrastructure. A personal or firm-side agent still needs counterparties, identity providers, payment rails, revocation propagation, cancellation and refund semantics, service-level evidence, and dispute evidence that other parties can recognize [23–25,82].

View 4: Enterprise suites, not super-apps, will define the first agent economies.

An enterprise-first view notes that business software often has stronger identity, permissions, logging, procurement rules, and administrative oversight than consumer apps, making it a more natural environment for reliable delegation [27,38,96]. **Response.** Enterprise systems may indeed

 doi:10.20944/preprints202604.1860.v1

define important administrative and compliance patterns, and this paper treats them as candidate substrates rather than afterthoughts. The narrower role of super-app ecosystems is empirical: they expose heterogeneous merchants, payments, logistics, travel, support, refunds, and consumer-facing redress at everyday scale. Enterprise suites and super-apps therefore reveal different portions of the same transaction-closure bottleneck [15,16,18–20].

View 5: Regulation, contract law, and consumer law, rather than architecture, will determine safe delegation.

A law-first view argues that liability rules, consumer-protection duties, procurement rules, contract doctrine, and competition policy will ultimately define how agents may contract, pay, cancel, refund, procure, file, and dispute on behalf of principals [12,33,36,97]. **Response.** Legal rules are necessary, but they need technical evidence and operational hooks. Contestable transaction closure provides those hooks by making mandates, state transitions, receipts, and handoff packets inspectable and exportable. Without such artifacts, regulators, courts, auditors, firms, and support teams may receive screenshots and fragmented support histories rather than reliable proof of authority, action, and recovery [25,31,32].

View 6: Payment-first or blockchain-based agent markets are sufficient.

A payments-first view suggests that once agents can pay each other or settle through purpose-built rails, the agent economy can emerge without super-app-style integration [28,46,98–100]. **Response.** Payment execution is a necessary component, but transactions fail around more than settlement: identity mismatches, stockouts, delivery disputes, booking-confirmation splits, procurement exceptions, cloud-resource overruns, service-level failures, cancellation windows, partial refunds, and recovery evidence all matter. Payment infrastructure therefore becomes transaction-ready only when it is embedded in a lifecycle that includes authority verification, fulfillment state, reversal, dispute, liability, and portable receipts [25,80,82].

4. Research Agenda: ClosureBench and Falsifiable Predictions

Comparative evaluation of candidate substrates for agent economies requires holding model capability constant. Evaluation must shift from asking whether an agent merely completed a task to whether it acted within its mandate, avoided unauthorized commitments, detected split states, reversed failed commitments, preserved receipts, and minimized human recovery burdens. *ClosureBench* proposes a benchmark design wherein identical logical workflows run across mature super-app ecosystems, AI-native surfaces, enterprise suites, open-web stacks, and protocol-native markets, subject to uniform mandates and failure injections. This isolates environment effects from model effects (Table 9).

Table 9. ClosureBench: an actionable benchmark design evaluating both environments and models.

ClosureBench component	Concrete design	Evaluated dimension
Mandate suite	Natural-language and structured mandates detailing task class, budget, time window, merchant scope, data scope, revocation, and escalation rules.	Adherence to bounded authority versus unconstrained task completion.
Substrate swap	Fix planner, principal profile, and goal; vary environment across super-app, AI-native surface, enterprise suite, open-web, or protocol-native stacks.	Isolation of environment-derived utility and risk from model performance.
Failure injections	Expose systems to price drift, stockouts, booking-payment splits, duplicate executions, partner outages, identity mismatches, biased steering, prompt injections, and mid-workflow revocations.	Resilience of autonomy against realistic exception paths.
Trace and recovery audit	Mandate the generation of formal mandate objects, lifecycle states, receipt graphs, and handoff packets.	Reconstructability, challengeability, and reparability by humans and institutions.
Outcome metrics	Track authorization errors, commitment success, split-state rates, reversal success, refund latency, receipt completeness, human recovery burdens, and portability costs.	Validation of transaction and recovery closure beyond binary task success.

This paradigm distinguishes optimal-path agents from recovery-capable agents. Optimal-path agents search, compare, and click under nominal conditions. Recovery-capable agents identify split states, prevent duplicate executions, process revocations, manage cancellations and refunds, escalate appropriately, generate robust receipts, and articulate failures when the environment behaves adversely.

 doi:10.20944/preprints202604.1860.v1

Crucial failures reside not in initial navigation, but in commitment and recovery. ClosureBench tests whether an agent remains functional after the optimal path deteriorates.

This framework yields five testable predictions. First, environment quality heavily determines outcomes as workflows increase in irreversibility. Second, strictly scoped delegation mitigates catastrophic authorization failures more effectively than planner-only enhancements under broad permissions. Third, the completeness of receipt structures strongly predicts principal trust and human-takeover efficiency compared to raw final success metrics. Fourth, mature transaction-heavy ecosystems demonstrate superior performance over thin action surfaces when subjected to failure injections such as price drift, booking-payment splits, refund delays, identity mismatches, partner outages, and revocation. Fifth, governance structures directly impact cost efficiency and principal regret independently of base model capability. These predictions face clear invalidation conditions: if AI-native or protocol-mediated ecosystems quickly add typed mandates, reversible commitment lifecycles, receipt graphs, and auditable handoff semantics, or if mature super-app ecosystems fail to reduce human rework under exceptions, then the projected near-term advantage diminishes significantly [27,32,84–86].

5. Conclusions

Agent economies require AI systems that can safely create, reverse, explain, and repair externally binding commitments. Mature super-app ecosystems provide strong near-term consumer-facing testbeds by merging extensive action surfaces with settlement, receipts, complementor governance, and redress at scale; enterprise suites expose organizational identity, approval, audit, and procurement constraints; protocol-native markets foreground interoperability; and AI-native agents such as Codex demonstrate the rapid convergence of action closure toward this terrain. The imperative is to extract the functional primitives from early operational environments and implement them as portable infrastructure: typed mandates, reversible commitment lifecycles, receipt graphs, auditable handoffs, liability-aware recovery, and contestable governance. Proactively establishing this infrastructure prevents accidental defaults and transforms the super-app-to-agent-economy transition into an open, accountable foundation for delegated AI.

Acknowledgments: Funding in direct support of this work: RIE2025 Industry Alignment Fund (Award I2301E0026) and the Alibaba–NTU Global e-Sustainability CorpLab.

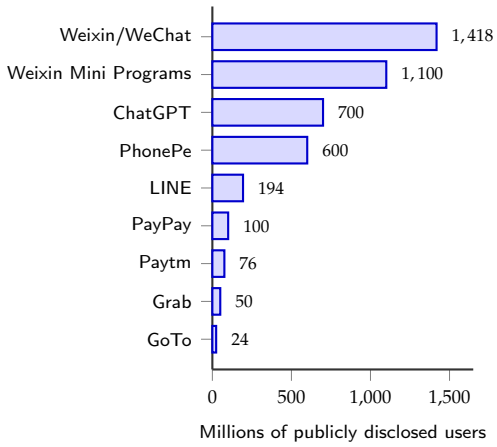
Appendix A. Additional Data and Convergence Signals

Company disclosures provide motivating evidence for the convergence of high-volume action, settlement, and agentic layers, especially in consumer-facing ecosystems. Appendix Figure A1 offers a comparative visualization, acknowledging differences in user denominators, reporting cadences, and GMV scopes. Rounded exchange rates are used strictly to establish a common visual scale for value flows.

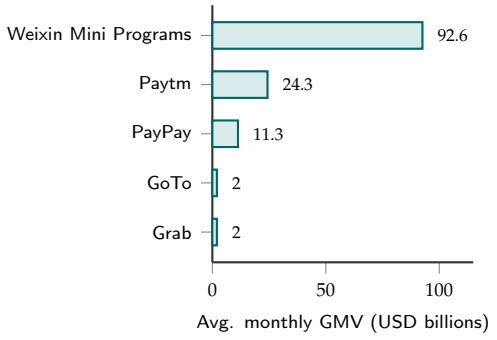
Appendix B. Extended Literature Map and Scope Clarifications

The broader bibliography spans multiple disciplines. Classical AI, browser-assisted agents, and safe-autonomy literature establishes the agent as a goal-directed system under constrained authority [102–109]. Regulatory, market-design, fintech, and payment frameworks detail liability, execution semantics, and economically meaningful delegation [97–100,110–113]. Super-app and service-bundling studies map integrated consumer platforms across geographies, highlighting transit utilization, service synergy, and localization [114–122]. Platform, ecosystem, and service-innovation theory characterizes complementors, boundary resources, and network effects, demonstrating how closure operates as either portable infrastructure or host-specific control [123–130]. Super-app privacy, compliance, and public-market data contextualize the empirical landscape [95,131–138].

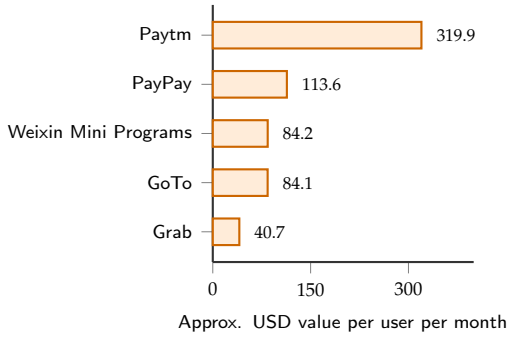
 doi:10.20944/preprints202604.1860.v1



(a) Reach: public user-scale disclosures



(b) Approx. transaction depth: monthly USD value



(c) Approx. per-user transaction intensity

Ecosystem	Everyday entry	Native tx	Extensible ecosystem	Explicit AI/agent
Mature super-app ecosystems				
Weixin / WeChat	●	●	●	◐
Grab	●	●	●	●
LY / LINE / PayPay	●	●	●	●
Kakao	●	◐	◐	●
Payments-first ecosystems				
Paytm	●	●	◐	●
PhonePe	●	●	◐	◐
AI-native / social-native challengers				
X	◐	◐	○	●
AI-native platforms (e.g., OpenAI/Codex)	◐	○	◐	●

(d) Agent-economy convergence map by ecosystem type

Figure A1. Public convergence signals. Panels (a)–(c) indicate scale; panel (d) assesses functional layers of agent-economy readiness [22,51–64,66–74,101]. Filled circles denote substantial public evidence; half-filled circles denote emerging or partial evidence. Context outside panels (b)–(c): Apps in ChatGPT can reach 800M+ users, Codex has 3M+ weekly developers, and PhonePe reported 600M active users.

Table A1. Derived quantities informing Figure A1.

Ecosystem	Transformation used	Approx. monthly USD-equivalent GMV / GTV	Approx. value per disclosed user per month	Interpretation caveat
Weixin Mini Programs	RMB 8T / 12 / 7.2	92.6B	84.2	Numerator is 2024 Mini Program GMV; denominator is 1.1B+ Mini Program MAU.
Paytm	INR 6.2 lakh crore / 3 / 85	24.3B	319.9	Quarterly GMV divided by monthly transacting users; payments-heavy mix makes intensity high.
PayPay	USD 100B / 12	8.3B	115.7	Uses Reuters-reported annual transaction volume and roughly 72M registered users; using monthly transacting users would produce a higher intensity.
GoTo	Rp95.3T / 3 / 15,600	2.0B	84.1	Uses quarterly core GTV and MTU disclosure from the same quarterly report.
Grab	USD 6.1B / 3	2.0B	40.7	Uses on-demand GMV rather than full ecosystem value capture.

 doi:10.20944/preprints202604.1860.v1

Adjacent work on agentic infrastructure, provenance, and evaluation.

These works are included only where they inform evaluation traces, provenance, workflow governance, agent-control infrastructure, or portable audit objects for transaction-ready delegation. Harness engineering and OpenClaw emphasize agent-control frameworks [139] and public agent ecosystems [140]. Acceptance-test methodologies [141,142] and automated-research structures [143–145] inform evaluation artifacts such as discovery traces and review pipelines. Furthermore, recommender systems research directly maps sequential recommendation [146] and preference learning [147] to agent trajectories [148]. Applied domains including sustainability [149–151], synthetic media [152–154], and climate-adaptation [155,156] motivate verifiable provenance graphs, lineage markets [157], and structured workforce governance [158].

Appendix B.1. What Counts as the Agentic AI Wave?

The *agentic AI wave* denotes the paradigm shift from content-centric foundation models toward systems pursuing multi-step goals in digital environments via tool integration, memory architectures, environmental interaction, and bounded autonomy [93,94,159–161]. This progression is tracked across web, desktop, operating-system, and mobile evaluation benchmarks [39–44,162–172].

Appendix B.2. What Counts as a Transaction-Ready Ecosystem?

A transaction-ready ecosystem is any digital environment that enables a delegated system to persist state across service boundaries, receive scoped authorization, execute and reverse value transfers, generate machine-readable traces, and route failures into structured redress pathways [23–25,27]. Super-app ecosystems represent one primary instantiation, but not an exclusive one.

Appendix B.3. Why Platform and Ecosystem Theory Matters Here

The construct of the *super-app ecosystem* is grounded in platform and ecosystem theory addressing digital transaction costs, multisided governance, boundary resources, and complementor dynamics [37,96,173–189]. *Ecosystem* defines the structural substrate, while *agent economy* describes the repeated delegated market interactions utilizing that substrate.

Appendix B.4. Why Interoperability Remains Central

Open protocols provide the architectural mechanism by which functionalities demonstrated in integrated environments transition into widespread utility. MCP addresses model-tool connectivity and resource access, whereas A2A targets cross-agent discovery, communication, and task coordination [84–86]. Should these protocols develop robust identity, payment, receipt, and liability layers, architectural preference will legitimately pivot away from tightly integrated host environments.

Appendix C. Operational Maturity Rubric for Transaction-Ready Ecosystems

Table A2 formalizes substrate-agnostic criteria for assessing integrated ecosystems, enterprise suites, federated stacks, and protocol-native markets.

Appendix D. Failure Taxonomy for Delegated Transaction Workflows

Transaction-ready environments must be evaluated under failure modes (Table A3), providing an explicit blueprint for benchmark injections and post-deployment monitoring.

Appendix E. Protocol Decomposition and Portability Targets

Translating integrated testbeds into robust architectures requires decomposing protocols into portable layers (Table A4).

 doi:10.20944/preprints202604.1860.v1

Table A2. Operational maturity rubric for transaction-ready ecosystems.

Dimension	Low signal	Intermediate signal	High signal	Closure risk if absent
Identity continuity	Frequent reauthentication or manual account stitching across services.	Some account-linked handoffs, but state resets under exceptions.	Shared identity, case state, and entitlements persist through normal and recovery paths.	Agents lose context at service boundaries; humans must repeatedly restate intent or credentials.
Mandate precision	Binary allow/deny permissions or one-time confirmation with little semantic scope.	Budget or merchant constraints exist, but time windows, categories, or fallback rules remain coarse.	Delegation can bind actor, task, budget, category, time window, revocation, and escalation semantics.	Over-broad authority increases catastrophic spend and makes human oversight unintelligible.
Spend control	Charges happen only at final checkout with weak pre-commitment controls.	Pre-authorizations or caps exist for some services, but cross-service budgeting is limited.	Holds, staged approvals, multi-step captures, and category caps are supported across workflows.	Agents either cannot act at all or can commit value without meaningful least-privilege safeguards.
Settlement reversibility	Cancellation or refund is manual, slow, or external to the workflow surface.	Some services support refunds or reversals, but partial failures remain opaque.	Commit, cancel, partial capture, refund, and dispute are first-class workflow states.	Errors that should be recoverable become costly or irreversible.
Receipt completeness	Users receive fragmented notifications or screenshots with little machine readability.	Some order histories or case logs exist, but not all actions are linked.	Authorization, execution, reversal, communication, and escalation emit linked machine-readable traces.	Trust erodes because neither users nor auditors can reconstruct what happened.
Handoff fidelity	Search, booking, payment, messaging, and support each reset task state.	Some cross-service context persists on the happy path.	Intent, constraints, and evidence survive both normal handoffs and exception-heavy recovery flows.	Agents appear capable in demos but collapse during realistic multi-service workflows.
Escalation routes	Human takeover requires abandoning the workflow or switching to unrelated channels.	Some service-specific support exists, but authority and evidence transfer poorly.	Human escalation preserves the mandate, prior actions, receipts, and unresolved branches of the task.	Partial automation creates new burdens because recovery starts from scratch.
Complementor transparency	Ranking, admission, and fallback rules are opaque.	Policies exist but are hard to audit across services.	Hosts expose legible rules for ranking, admission, fallback, and dispute treatment.	Agents may optimize against hidden steering or self-preferencing incentives.
Portability	Data and receipts are difficult to export or reinterpret elsewhere.	Some export tools exist but omit action traces or policy semantics.	Receipts, mandates, and task state can be exported or mapped into portable representations.	High closure becomes lock-in rather than transferable maturity.
Privacy containment	Cross-service actions are logged broadly with few user-visible controls.	Some data minimization and disclosure controls exist.	Granular disclosures, revocation, scoped data sharing, and retention limits constrain observation.	A powerful testbed becomes socially unacceptable because surveillance risk scales with closure.
Ranking neutrality	Discovery layers routinely steer users toward opaque host preferences.	Some neutrality policies exist but without robust audit support.	External audits can test for steering, self-preferencing, and unexplained fallback differences.	Completion gains may conceal ecosystem-level distortions in user welfare or complementor opportunity.
External standard support	The ecosystem speaks only host-specific APIs and identifiers.	Limited support exists for external identity or tool standards.	Core objects can map into open standards for identity, tooling, coordination, and audit export.	Lessons learned inside the ecosystem remain trapped there rather than becoming infrastructure.

 doi:10.20944/preprints202604.1860.v1

Table A3. Failure taxonomy for delegated transaction workflows.

Failure mode	Example injection	Immediate harm	Closure primarily stressed	Representative metric or test
Price drift before commit	Fare or checkout total changes after the agent selects an option.	Budget overrun or silent deviation from user constraints.	Transaction closure + mandate precision	Regret relative to user cap; rate of unauthorized upward price acceptance.
Stockout with substitution	Ordered item becomes unavailable and the merchant proposes a substitute.	Wrong-good delivery or coercive acceptance of lower-value items.	Completion closure + mandate precision	Substitution acceptance accuracy; post-hoc user regret.
Booking-confirmation split	Payment succeeds but reservation, ticket, or appointment issuance fails.	User is charged without receiving the intended service.	Transaction closure + recovery closure	Mean time to detect split state; refund latency; receipt linkage quality.
Duplicate execution	Network retries or tool ambiguity lead to repeated orders or bookings.	Double charge, duplicated reservations, or inventory exhaustion.	Recovery closure + reversibility	Duplicate rate under retry stress; idempotency coverage.
Travel disruption	Delay or cancellation forces downstream rebooking, ride changes, or reimbursement.	Cascading service failures and time-sensitive losses.	Completion closure + recovery closure	Recovered utility after disruption; minutes saved; residual out-of-pocket cost.
Delivery failure	Courier cannot complete delivery or marks the order delivered incorrectly.	Lost goods, spoiled goods, or fraudulent completion.	Recovery closure + escalation	Successful recovery rate; time to redress; evidence completeness.
Identity mismatch	Merchant, platform, or public-service endpoint rejects the user's identity state.	Service denial despite prior workflow progress.	Completion closure + recovery closure	Recovery rate after KYC/identity friction; state persistence through retry.
Partner outage	Mini app, merchant API, or support channel becomes unavailable mid-task.	Workflow stalls after partial commitment.	Completion closure + escalation	Safe rollback rate; fallback success; user rework burden.
Prompt injection or malicious content	Third-party content instructs the agent to deviate from user goals or leak secrets.	Unauthorized action or credential misuse.	Recovery closure + privacy containment	Success rate of defense; harmful action rate under adversarial prompts.
Biased steering	Ranking or tool suggestions push the agent toward host-preferred merchants.	Higher prices, lower quality, or reduced complementor fairness.	Governance transparency + ranking neutrality	Excess cost versus neutral ranking baseline; steering detectability.
Procurement-policy violation	Agent selects a vendor, term, or purchase amount outside organizational policy.	Unauthorized spend, compliance failure, or invalid approval chain.	Mandate precision + recovery closure	Policy-violation rate; re-approval success; audit evidence completeness.
Resource-cost overrun	Agent provisions paid compute, API calls, or infrastructure beyond the approved resource cap.	Unexpected operational cost or degraded service reliability.	Transaction closure + mandate precision	Cost overrun relative to cap; rollback success; resource-revocation latency.
Service-level dispute	Counterparty claims fulfillment while the principal or monitoring system records failed or degraded service.	Payment without promised performance or unresolved liability.	Recovery closure + receipt graph	Dispute success rate; service-proof quality; handoff completeness.
Refund degradation	Refund is partial, delayed, or routed outside the original workflow.	Liquidity loss and user distrust even when the agent recognized the failure.	Transaction closure + recovery closure	Refund completion rate; time-to-refund; documentation burden.
Mid-workflow revocation	User retracts authority while a background task is still executing.	Confusion over whether downstream actions should continue or roll back.	Mandate precision + escalation	Revocation latency; fraction of actions correctly halted or compensated.

 doi:10.20944/preprints202604.1860.v1

Table A4. Protocol decomposition of the portability problem.

Layer	Typical realization in integrated ecosystems	Portable target primitive	Remaining gap
Identity and authentication	Shared account systems, session state, or host-provided sign-in	Interoperable agent identity plus user-linked authentication context	Current standards identify principals well, but agent-specific delegation and liability semantics remain immature.
Mandate and consent	Host-specific checkout prompts, app permissions, and background-task approvals	Transferable mandate objects with budgets, categories, revocation, and escalation rules	Most ecosystems still treat consent as UI events rather than reusable machine-readable objects.
Tool and data access	Host APIs, mini-app SDKs, and connector frameworks	MCP-like tool exposure with standardized policy envelopes	Tool access is advancing faster than spend control, redress, or audit export.
Cross-agent coordination	Internal orchestration or proprietary task routing	A2A-like task-handoff objects with state, scope, and responsibility metadata	Coordination standards rarely encode liability-sensitive workflow state.
Settlement lifecycle	Platform-native checkout, refunds, and partner-specific exception handling	Standardized commit, cancel, refund, dispute, and compensation callbacks	Payment rails remain fragmented across merchants, categories, and jurisdictions.
Receipts and provenance	In-app order histories, notifications, and case records	Receipt graphs linking authorization, execution, reversal, and escalation	There is no widely adopted portable schema for agent-action receipts with audit semantics.
Governance and ranking	Centralized admission, ranking, and fallback policy	Auditable policy surfaces, neutrality probes, and exportable logs	Ecosystem-level steering remains difficult to test independently of host cooperation.
Privacy-preserving export	Platform-specific exports or user data downloads	Minimal, scoped export of traces needed for audit, switching, and dispute	The field lacks consensus on how to maximize contestability without overexposing sensitive behavioral traces.

Appendix F. Benchmark Sketches for Future Work

Travel and mobility closure.

Tasks should chain search, comparison, booking, disruption response, ride-hailing, reimbursement, and proof preservation. Metrics capture recovered utility under disruption, delay minutes saved, refund latency, and user correction burden.

Commerce and returns closure.

Tasks should chain discovery, merchant verification, coupon usage, checkout, delivery selection, failed delivery handling, returns, and refund tracking. Metrics encompass total cost efficiency, return success rates, dispute resolution burdens, and receipt completeness.

Public-service closure.

Tasks should integrate identity proofing, form completion, appointment booking, fee remittance, submission verification, and exception escalation. Metrics measure policy compliance, bureaucratic burden reduction, proof-of-submission stability, and time-to-redress.

Enterprise procurement closure.

Tasks should chain vendor discovery, policy checking, quote comparison, purchase-order issuance, invoice reconciliation, delivery verification, and dispute escalation. Metrics capture policy adherence, approval latency, evidence continuity, and cost recovery.

Cloud and software-operations closure.

Tasks should chain infrastructure provisioning, paid API use, rollback, incident response, and audit-log preservation. Metrics capture resource-cap violations, rollback success, idempotency failures, and service-level evidence quality.

 doi:10.20944/preprints202604.1860.v1

Agent-to-agent market closure.

Tasks should involve one agent contracting with another for data, compute, logistics, verification, or support. Metrics capture counterparty authentication, settlement success, service-proof completeness, replay resistance, and portability of dispute packets.

Cross-substrate panels.

For each benchmark family, the core logical workflow must be reproduced across integrated ecosystems, enterprise suites, federated service stacks, and protocol-native architectures, quantifying specific environmental closure capabilities separate from base model variations.

References

1. Maslej, N.; Fattorini, L.; Perrault, R.; Gil, Y.; Parli, V.; Kariuki, N.; Capstick, E.; Reuel, A.; Brynjolfsson, E.; Etchemendy, J.; et al. Artificial intelligence index report 2025. *arXiv preprint arXiv:2504.07139* **2025**.
2. Staufer, L.; Feng, K.; Wei, K.; Bailey, L.; Duan, Y.; Yang, M.; Ozisik, A.P.; Casper, S.; Kolt, N. The 2025 AI agent index: Documenting technical and safety features of deployed agentic AI systems. *arXiv preprint arXiv:2602.17753* **2026**.
3. OECD. The Agentic AI Landscape and Its Conceptual Foundations, 2026. OECD report.
4. Infocomm Media Development Authority of Singapore. Model AI Governance Framework for Agentic AI, 2026. Policy framework.
5. Kasirzadeh, A.; Gabriel, I. Characterizing AI agents for alignment and governance. *arXiv preprint arXiv:2504.21848* **2025**.
6. Feng, K.J.; McDonald, D.W.; Zhang, A.X. Levels of autonomy for ai agents. *arXiv preprint arXiv:2506.12469* **2025**.
7. Rothschild, D.M.; Mobius, M.; Hofman, J.M.; Dillon, E.W.; Goldstein, D.G.; Immorlica, N.; Jaffe, S.; Lucier, B.; Slivkins, A.; Vogel, M. The agentic economy. *arXiv preprint arXiv:2505.15799* **2025**.
8. Hadfield, G.K.; Koh, A. An economy of AI agents. *arXiv preprint arXiv:2509.01063* **2025**.
9. Tomasev, N.; Franklin, M.; Leibo, J.Z.; Jacobs, J.; Cunningham, W.A.; Gabriel, I.; Osindero, S. Virtual agent economies. *arXiv preprint arXiv:2509.10147* **2025**.
10. Chaffer, T.J. Can We Govern the Agent-to-Agent Economy? *arXiv preprint arXiv:2501.16606* **2025**.
11. Liu, X.; Shang, H.; Jin, H. When Agent Markets Arrive. *arXiv preprint arXiv:2604.06688* **2026**.
12. Busch, C. Consumer Law for AI Agents. *arXiv preprint arXiv:2507.11567* **2025**.
13. Diaz Baquero, A.P. Super Apps: opportunities and challenges. PhD thesis, Massachusetts Institute of Technology, 2021.
14. Hasselwander, M. Digital platforms' growth strategies and the rise of super apps. *Heliyon* **2024**, *10*.
15. Lucas Jr, D.; Lopes, E.L. Defining a super app and analyzing it from an ecosystemic perspective **2024**.
16. Hasselwander, M.; Kriswardhana, W.; Esztergár-Kiss, D.; Lah, O.; Steinberg, M. One App for Everything: A Multidisciplinary Review of Super Apps. *Available at SSRN 5772697*.
17. van der Vlist, F.N.; Helmond, A.; Dieter, M.; Weltevrede, E. Super-appification: Conglomeration in the global digital economy. *New Media & Society* **2025**, *27*, 3314–3337.
18. Goggin, G.; Athique, A.M. Super apps as digital transaction platforms: What Southeast Asia's Grab tells us. *Platforms & Society* **2026**, *3*, 29768624251412170.
19. Grab. Grab – Your Everyday Everything App, 2026. Accessed April 12, 2026.
20. Paytm. Consumer Engagement on Paytm Super App Is at Its Highest, 2023. January 23, 2023.
21. Reuters. X Seals Payments Deal with Visa in Push Toward Musk's "Everything App" Goal, Source Says, 2025. January 28, 2025.
22. Reuters. OpenAI Plans Desktop "Superapp" to Streamline User Experience, 2026. March 19, 2026.
23. South, T.; Marro, S.; Hardjono, T.; Mahari, R.; Whitney, C.D.; Greenwood, D.; Chan, A.; Pentland, A. Authenticated delegation and authorized ai agents. *arXiv preprint arXiv:2501.09674* **2025**.
24. South, T.; Nagabhushanaradhya, S.; Dissanayaka, A.; Cecchetti, S.; Fletcher, G.; Lu, V.; Pietropaolo, A.; Saxe, D.H.; Lombardo, J.; Shivalingaiah, A.M.; et al. Identity Management for Agentic AI: The new frontier of authorization, authentication, and security for an AI agent world. *arXiv preprint arXiv:2510.25819* **2025**.
25. Consumer Bankers Association. Agentic AI Payments: Navigating Consumer Protection, Innovation, and Regulatory Frameworks, 2026. White paper.

 doi:10.20944/preprints202604.1860.v1

26. Rauba, P.; Cepenas, S.; van der Schaar, M. Multi-Agent Systems Should be Treated as Principal-Agent Problems. *arXiv preprint arXiv:2601.23211* **2026**.
27. Chan, A.; Wei, K.; Huang, S.; Rajkumar, N.; Perrier, E.; Lazar, S.; Hadfield, G.K.; Anderljung, M. Infrastructure for AI agents. *arXiv preprint arXiv:2501.10114* **2025**.
28. Safwan Uddin, M.; Mouzam, M.; Imran, M.; Faizan, S.B.U. APEX: Agent Payment Execution with Policy for Autonomous Agent API Access. *arXiv e-prints* **2026**, pp. arXiv–2604.
29. Hardt, D. The OAuth 2.0 authorization framework. Technical report, 2012.
30. Sakimura, N.; Bradley, J.; Jones, M.; De Medeiros, B.; Mortimore, C. OpenID Connect Core 1.0 incorporating errata set 1. *The OpenID Foundation, specification* **2014**, *335*.
31. Belhajjame, K.; B'Far, R.; Cheney, J.; Coppens, S.; Cresswell, S.; Gil, Y.; Groth, P.; Klyne, G.; Lebo, T.; McCusker, J.; et al. Prov-dm: The prov data model. *W3C Recommendation* **2013**, *14*, 15–16.
32. Ye, B.; Li, R.; Yang, Q.; Liu, Y.; Yao, L.; Lv, H.; Xie, Z.; An, C.; Li, L.; Kong, L.; et al. Claw-Eval: Toward Trustworthy Evaluation of Autonomous Agents. *arXiv preprint arXiv:2604.06132* **2026**.
33. Vezzoso, S. 'Super-apps' and the Digital Markets Act. *Journal of Antitrust Enforcement* **2024**, *12*, 331–337.
34. Treasury, H. Unlocking digital competition, report of the digital competition expert panel (2019) **2019**.
35. Crémer, J.; De Montjoye, Y.A.; Schweitzer, H. *Competition policy for the digital era*; Publications Office of the European Union, 2019.
36. UK Competition and Markets Authority. Mobile Ecosystems Market Study, 2022. Final report.
37. Jacobides, M.G.; Cennamo, C.; Gawer, A. Towards a theory of ecosystems. *Strategic management journal* **2018**, *39*, 2255–2276.
38. Tilson, D.; Lyytinen, K.; Sørensen, C. Research commentary—Digital infrastructures: The missing IS research agenda. *Information systems research* **2010**, *21*, 748–759.
39. Liu, X.; Yu, H.; Zhang, H.; Xu, Y.; Lei, X.; Lai, H.; Gu, Y.; Ding, H.; Men, K.; Yang, K.; et al. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688* **2023**.
40. Deng, X.; Gu, Y.; Zheng, B.; Chen, S.; Stevens, S.; Wang, B.; Sun, H.; Su, Y. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems* **2023**, *36*, 28091–28114.
41. Zhou, S.; Xu, F.F.; Zhu, H.; Zhou, X.; Lo, R.; Sridhar, A.; Cheng, X.; Ou, T.; Bisk, Y.; Fried, D.; et al. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854* **2023**.
42. Xie, T.; Zhang, D.; Chen, J.; Li, X.; Zhao, S.; Cao, R.; Hua, T.J.; Cheng, Z.; Shin, D.; Lei, F.; et al. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems* **2024**, *37*, 52040–52094.
43. Rawles, C.; Clinckemaillie, S.; Chang, Y.; Waltz, J.; Lau, G.; Fair, M.; Li, A.; Bishop, W.; Li, W.; Campbell-Ajala, F.; et al. Androidworld: A dynamic benchmarking environment for autonomous agents. *arXiv preprint arXiv:2405.14573* **2024**.
44. Kong, Q.; Zhang, X.; Yang, Z.; Gao, N.; Liu, C.; Tong, P.; Cai, C.; Zhou, H.; Zhang, J.; Chen, L.; et al. MobileWorld: Benchmarking Autonomous Mobile Agents in Agent-User Interactive and MCP-Augmented Environments. *arXiv preprint arXiv:2512.19432* **2025**.
45. Kapoor, S.; Stroebl, B.; Siegel, Z.S.; Nadgir, N.; Narayanan, A. Ai agents that matter. *arXiv preprint arXiv:2407.01502* **2024**.
46. Chua, J.K.; Huang, D.; Wang, Z. A Novel Hierarchical Multi-Agent System for Payments Using LLMs. *arXiv preprint arXiv:2602.24068* **2026**.
47. Brynjolfsson, E. The productivity paradox of information technology. *Communications of the ACM* **1993**, *36*, 66–77.
48. David, P.A. The dynamo and the computer: an historical perspective on the modern productivity paradox. *The American economic review* **1990**, *80*, 355–361.
49. Brynjolfsson, E.; Hitt, L.M. Beyond computation: Information technology, organizational transformation and business performance. *Journal of Economic perspectives* **2000**, *14*, 23–48.
50. North, D.C. Institutions, institutional change and economic performance. *Cambridge Univ Pr* **1990**.
51. Grab. Grab Unveils 13 AI-Powered Experiences at GrabX 2026 as Southeast Asia's Intelligent Everyday Guide, 2026. April 8, 2026.
52. Paytm Editorial Team. Paytm Launches the All-New App, Combining Clean Interface and AI-First Features to Redefine Everyday Payments, 2025. November 10, 2025. Paytm blog.
53. Paytm Editorial Team. Paytm Bets on AI in Travel, Launches Dedicated AI Travel Booking App "Paytm Checkin", 2025. November 11, 2025. Paytm blog.

 doi:10.20944/preprints202604.1860.v1

54. GoTo Group. GoTo Group Records First Quarterly Adjusted Pre-Tax Profit and Raises Full Year Guidance as It Reports 2025 Third Quarter Earnings, 2025. November 12, 2025. Official earnings release.
55. LY Corporation. Business Development Outside of Japan, 2025. Accessed April 19, 2026. Official company page reporting global LINE monthly active users as of March 31, 2025.
56. LY Corporation. FY2025 Q2 Earnings Highlights, 2025. October 30, 2025. Official earnings presentation.
57. LY Corporation. FY2025 Q3 Earnings Highlights, 2026. February 6, 2026. Official earnings presentation.
58. LY Corporation. FY2024 Full Year & Q4 Earnings Highlights, 2025. May 8, 2025. Official earnings presentation.
59. Kakao. Kakao Hits Record-High Q3 2025 Results — KRW 2,087 Billion in Revenue, KRW 208 Billion in Operating Profit, 2025. November 7, 2025. Official results release.
60. PhonePe Team. PhonePe 2025: The Year at a Glance, 2025. December 2025. Official year-end blog post.
61. X Help Center. About Grok, Your Humorous AI Assistant on X, 2026. Accessed April 17, 2026.
62. Reuters. Elon Musk Says X Money to Enter Early Public Access Next Month, 2026. March 10, 2026.
63. X Corp.. XChat App Store Listing, 2026. Accessed April 17, 2026. App Store listing.
64. OpenAI. Introducing Codex, 2025. May 16, 2025. OpenAI product announcement.
65. OpenAI. Codex Is Now Generally Available, 2025. October 6, 2025. OpenAI product announcement.
66. OpenAI. Introducing the Codex App, 2026. February 2, 2026. OpenAI product announcement.
67. OpenAI. Codex for (Almost) Everything, 2026. April 16, 2026. OpenAI product announcement.
68. Tencent. Tencent Announces 2025 Annual and Fourth Quarter Results, 2026. March 18, 2026. Annual results release.
69. Tencent. Corporate Overview, 2026. March 27, 2026. Corporate overview deck. https://static.www.tencent.com/uploads/2026/03/27/c8736de5d9eb9f55aaa90c99212ca9a8.pdf.
70. Tencent. Weixin Empowers Local Businesses with Mini Program and Payment Solutions Ahead of Visit Malaysia 2026, 2026. April 17, 2026. Tencent article.
71. Grab. Grab Reports Fourth Quarter and 2025 Results with First Full Year Net Profit, 2026. February 12, 2026. Earnings release.
72. Paytm. Earnings Release: Q3 FY 2026, 2026. January 29, 2026. Earnings release.
73. OpenAI. GPT-5 and the New Era of Work, 2025. August 7, 2025. OpenAI product announcement.
74. OpenAI. Introducing Apps in ChatGPT and the New Apps SDK, 2025. October 6, 2025. OpenAI product announcement.
75. Baskaran, S.; Zhao, L.; Mannan, M.; Youssef, A. Measuring the leakage and exploitability of authentication secrets in super-apps: The wechat case. In Proceedings of the Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses, 2023, pp. 727–743.
76. Zhang, Y.; Yang, Y.; Lin, Z. Don't leak your keys: Understanding, measuring, and exploiting the appsecret leaks in mini-programs. In Proceedings of the Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security, 2023, pp. 2411–2425.
77. Li, W.; Yang, B.; Ye, H.; Xiang, L.; Tao, Q.; Wang, X.; Zhou, C. Minitracker: Large-scale sensitive information tracking in mini apps. *IEEE Transactions on Dependable and Secure Computing* **2023**, *21*, 2099–2114.
78. Wang, M.; Lin, P.; Knockel, J.; Greenberg, W.; Mayer, J.; Mittal, P. What WeChat Knows: Pervasive First-Party Tracking in a Billion-User Super-App Ecosystem. *Proceedings on Privacy Enhancing Technologies* **2025**.
79. Cai, Y.; Zhang, Z.; Yao, M.; Liu, J.; Zhao, X.; Fu, X.; Li, R.; Liu, Z.; Chen, X.; Guo, Y.; et al. I can tell your secrets: Inferring privacy attributes from mini-app interaction history in super-apps. In Proceedings of the 34th USENIX Security Symposium (USENIX Security 25), 2025, pp. 6541–6560.
80. Mao, Q.; Wang, J.; Liu, Y.; Zhu, L.; Ma, C.; Yan, J. SoK: Security of Autonomous LLM Agents in Agentic Commerce. *arXiv preprint arXiv:2604.15367* **2026**.
81. Debi, T.; Zhu, W. Whispers of Wealth: Red-Teaming Google's Agent Payments Protocol via Prompt Injection. *arXiv preprint arXiv:2601.22569* **2026**.
82. Acharya, V. Secure Autonomous Agent Payments: Verifying Authenticity and Intent in a Trustless Environment. *arXiv preprint arXiv:2511.15712* **2025**.
83. Zhang, J.; Yang, L.; Han, Y.; Xiang, Z.; Hei, X. A small leak will sink many ships: Vulnerabilities related to mini-programs permissions. In Proceedings of the 2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC). IEEE, 2023, pp. 595–606.
84. Model Context Protocol. Specification, 2025. Version 2025-11-25.
85. Google Cloud. Announcing the Agent2Agent Protocol (A2A), 2025. April 9, 2025.

 doi:10.20944/preprints202604.1860.v1

86. Linux Foundation. A2A Protocol Surpasses 150 Organizations, Lands in Major Cloud Platforms, and Sees Enterprise Production Use in First Year, 2026. April 9, 2026.
87. Tomašev, N.; Franklin, M.; Osindero, S. Intelligent AI delegation. *arXiv preprint arXiv:2602.11865* **2026**.
88. Horvitz, E. Principles of mixed-initiative user interfaces. In Proceedings of the Proceedings of the SIGCHI conference on Human Factors in Computing Systems, 1999, pp. 159–166.
89. Parasuraman, R.; Sheridan, T.B.; Wickens, C.D. A model for types and levels of human interaction with automation. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans* **2000**, *30*, 286–297.
90. Lee, J.D.; See, K.A. Trust in automation: Designing for appropriate reliance. *Human factors* **2004**, *46*, 50–80.
91. Amershi, S.; Weld, D.; Vorvoreanu, M.; Fourney, A.; Nushi, B.; Collisson, P.; Suh, J.; Iqbal, S.; Bennett, P.N.; Inkpen, K.; et al. Guidelines for human-AI interaction. In Proceedings of the Proceedings of the 2019 chi conference on human factors in computing systems, 2019, pp. 1–13.
92. Shneiderman, B. *Human-centered AI*; Oxford University Press, 2022.
93. Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; Cao, Y. ReAct: Synergizing Reasoning and Acting in Language Models. In Proceedings of the International Conference on Learning Representations (ICLR), 2023.
94. Schick, T.; Dwivedi-Yu, J.; Dessì, R.; Raileanu, R.; Lomeli, M.; Hambro, E.; Zettlemoyer, L.; Cancedda, N.; Scialom, T. Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems* **2023**, *36*, 68539–68551.
95. Adavi, K.A.K.; Acker, A. A platform perspective for personal information practices on superapps. *Platforms & Society* **2026**, *3*, 29768624261418430.
96. De Reuver, M.; Sørensen, C.; Basole, R.C. The digital platform: a research agenda. *Journal of information technology* **2018**, *33*, 124–135.
97. Gardhouse, K.; Oueslati, A.; Kolt, N. Regulating AI Agents. *arXiv preprint arXiv:2603.23471* **2026**.
98. Xu, M. The Agent Economy: A Blockchain-Based Foundation for Autonomous AI Agents. *arXiv preprint arXiv:2602.14219* **2026**.
99. Noel, T. Purpose-Built Payment Infrastructure for Autonomous AI Agents. *Available at SSRN 6265040* **2026**.
100. Zhang, Y.; Xiang, Y.; Lei, Y.; Wang, Q.; Qiu, T.; Sun, Y.; Zarkov, S.; Yuen, T.H.; Deppeler, A.; Yu, J.; et al. SoK: Blockchain Agent-to-Agent Payments. *arXiv preprint arXiv:2604.03733* **2026**.
101. Reuters. SoftBank-backed PayPay valued at $12.7 billion in Nasdaq debut as shares jump. https://www.reuters.com/business/media-telecom/softbanks-paypay-set-hotly-anticipated-nasdaq-debut-after-raising-880-million-2026-03-12/, 2026. Reuters.
102. Russell, S.; Norvig, P.; Intelligence, A. A modern approach. *Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs* **1995**, *25*, 79–80.
103. Nakano, R.; Hilton, J.; Balaji, S.; Wu, J.; Ouyang, L.; Kim, C.; Hesse, C.; Jain, S.; Kosaraju, V.; Saunders, W.; et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332* **2021**.
104. Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; Mané, D. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565* **2016**.
105. Altman, E. *Constrained Markov decision processes*; Routledge, 2021.
106. Achiam, J.; Held, D.; Tamar, A.; Abbeel, P. Constrained policy optimization. In Proceedings of the International conference on machine learning. Pmlr, 2017, pp. 22–31.
107. Garcıa, J.; Fernández, F. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research* **2015**, *16*, 1437–1480.
108. Dulac-Arnold, G.; Mankowitz, D.; Hester, T. Challenges of real-world reinforcement learning. *arXiv preprint arXiv:1904.12901* **2019**.
109. Hadfield-Menell, D.; Dragan, A.D.; Abbeel, P.; Russell, S. The off-switch game. In Proceedings of the AAAI Workshops, 2017.
110. Sanabria, J.M.; Vecino, P.A. Beyond the sum: Unlocking ai agents potential through market forces. *arXiv preprint arXiv:2501.10388* **2024**.
111. Arner, D.W.; Barberis, J.; Buckley, R.P. The evolution of Fintech: A new post-crisis paradigm. *Geo. J. Int'l L.* **2015**, *47*, 1271.
112. Philippon, T. The fintech opportunity. Technical report, National Bureau of Economic Research, 2016.
113. Brunnermeier, M.K.; James, H.; Landau, J.P. The digitalization of money. Technical report, National Bureau of Economic Research, 2019.

 doi:10.20944/preprints202604.1860.v1

114. Ye, S. The rise of superapps in emerging countries. In Proceedings of the 2022 4th international Conference on economic Management and cultural industry (ICEMCI 2022). Atlantis Press, 2022, pp. 1847–1856.
115. Balogh, A.; Varga, J. Superapps: At the crossroads of enhanced customer experience and innovation management theories. *Edelweiss Applied Science and Technology* **2024**, *8*, 848–860.
116. Hasselwander, M.; Weiss, D. Consumer preferences for super app services: E-commerce, social media, and banking dominate. *European Research on Management and Business Economics* **2025**, *31*, 100284.
117. Hasselwander, M. Who wants to use transport super apps? Insights from combined PLS-SEM and NCA methods. *Transportation Research Part F: Traffic Psychology and Behaviour* **2025**, *113*, 1–13.
118. Fang, Y.H.; Liao, C.H.; Li, C.Y. Super app on demand: Exploring the impact of service synergy on willingness to use a new service. *Electronic Commerce Research and Applications* **2024**, *67*, 101430.
119. Gupta, S.; Gupta-Rawal, S.; Shrivastava, P. Dynamic AI-Embedded Super App: A Design-Based Process Innovation for Customer Engagement and Value Creation. *Journal of Product Innovation Management* **2026**, *43*, 99–124.
120. Loh, X.M.; Shyu, W.H.; Lee, V.H.; Huang, H.L.; Tan, G.W.H.; Ooi, K.B. Super App: A Multi-Analytical Cross-Country Approach on the "Everything" App. *Journal of Computer Information Systems* **2025**, pp. 1–14.
121. Şimşekler, S. The effects of service design on super app brand perception and user experience. Master's thesis, Middle East Technical University (Turkey), 2024.
122. Minghai, Y.; Wenqing, L.; Khan, W.A.; Nurhalim, W. The SuperApp Implementation in Business: Revolutionizing Business Operations for a Seamless Future. *Bincang Sains dan Teknologi* **2023**, *2*, 118–123.
123. Parker, G.G.; Van Alstyne, M.W. Two-sided network effects: A theory of information product design. *Management science* **2005**, *51*, 1494–1504.
124. Evans, D.S.; Schmalensee, R. *Matchmakers: The new economics of multisided platforms*; Harvard Business Review Press, 2016.
125. Parker, G.G.; Van Alstyne, M.W.; Choudary, S.P. *Platform revolution: How networked markets are transforming the economy and how to make them work for you*; WW Norton & Company, 2016.
126. Cennamo, C.; Santalo, J. Platform competition: Strategic trade-offs in platform markets. *Strategic management journal* **2013**, *34*, 1331–1350.
127. Adner, R. Ecosystem as structure: An actionable construct for strategy. *Journal of management* **2017**, *43*, 39–58.
128. Lusch, R.F.; Nambisan, S. Service innovation: A service-dominant logic perspective1. *MIS quarterly* **2015**, *39*, 155–175.
129. Kapoor, R. Ecosystems: broadening the locus of value creation. *Journal of Organization Design* **2018**, *7*, 1–16.
130. Engert, M.; Hein, A.; Maruping, L.M.; Thatcher, J.B.; Krcmar, H. Self-organization and governance in digital platform ecosystems: an information ecology approach. *Mis Quarterly* **2025**, *49*, 91–122.
131. Fasnacht, D. Super apps as catalysts: designing value constellations in open and digital ecosystems. *Journal of Business Strategy* **2026**, pp. 1–20.
132. Lin, J.; Zhu, J.; Zhou, Z.; Xi, Y.; Liu, W.; Yu, Y.; Zhang, W. Superplatforms Have to Attack AI Agents. *arXiv preprint arXiv:2505.17861* **2025**.
133. Xiang, D.; Jin, L.; Wu, S.; Lin, W.; Ding, Z. MiniEval: Automated detection of compliance violations and quantitative privacy risk assessment in MiniApps. *Information and Software Technology* **2026**, *195*, 108103. https://doi.org/https://doi.org/10.1016/j.infsof.2026.108103.
134. Tencent. Business Overview, 2025. December 22, 2025. Corporate overview deck.
135. Grab. Product Innovation, 2026. Accessed April 17, 2026. Grab Inside Grab product innovation page.
136. xAI. Grok 4, 2025. July 9, 2025. xAI product announcement.
137. Reuters. Uber and iFood Announce Strategic Partnership in Brazil, 2025. May 14, 2025.
138. Reuters. Russia Sees China's WeChat, Douyin as Models for Its Homegrown Max Messenger, 2026. April 8, 2026.
139. He, C.; Zhou, X.; Wang, D.; Xu, H.; Liu, W.; Miao, C. Harness Engineering for Language Agents: The Harness Layer as Control, Agency, and Runtime. *Preprints* **2026**. https://doi.org/10.20944/preprints202603.1756.v1.
140. He, C.; Zhou, X.; Wang, D.; Xu, H.; Liu, W.; Miao, C. OpenClaw as Language Infrastructure: A Case-Centered Survey of a Public Agent Ecosystem in the Wild. *Preprints* **2026**. https://doi.org/10.20944/preprints202603.1060.v1.
141. He, C.; Zhou, X.; Wang, D.; Xu, H.; Liu, W.; Miao, C. Human-AI Productivity Claims Should Be Reported as Time-to-Acceptance under Explicit Acceptance Tests. *TechRxiv* **2026**. TechRxiv preprint, https://doi.org/10.36227/techrxiv.177040595.50580086/v1.

 doi:10.20944/preprints202604.1860.v1

142. He, C.; Zhou, X.; Wang, D.; Xu, H.; Liu, W.; Miao, C. Beyond Human Measure: ASI Should Be Guided by Open-World Alignment. *Preprints* **2026**. https://doi.org/10.20944/preprints202604.1749.v1.
143. He, C.; Zhou, X.; Wang, D.; Xu, H.; Liu, W.; Miao, C. The AutoResearch Moment: From Experimenter to Research Director. *Preprints* **2026**. https://doi.org/10.20944/preprints202603.1329.v1.
144. He, C.; Zhou, X.; Wang, D.; Xu, H.; Liu, W.; Miao, C. Let Papers Flow: AI Conferences Should Embrace Submission Explosion via Autonomous Review Pipelines. *Preprints* **2026**.
145. He, C.; Zhou, X.; Wang, D.; Xu, H.; Liu, W.; Miao, C. The One-Person Laboratory Should Be a First-Class Unit of Evaluation in Dry-Lab AI Research. *Preprints* **2026**.
146. He, C.; Liu, Y.; Guo, Q.; Miao, C. Multi-scale quasi-RNN for next item recommendation. *arXiv preprint arXiv:1902.09849* **2019**.
147. He, C.; Zhou, X.; Wang, D.; Xu, H.; Liu, W.; Miao, C. Recommender Systems Should Now Be Designed Towards Agents. *Preprints* **2026**.
148. He, C.; Zhou, X.; Wang, D.; Xu, H.; Liu, W.; Miao, C. SP$\hat{2}$DPO: An LLM-assisted Semantic Per-Pair DPO Generalization. *arXiv preprint arXiv:2601.22385* **2026**.
149. He, C.; Zhou, X.; Wang, D.; Xu, H.; Liu, W.; Miao, C. PCA-OS: A Planetary Climate Adaptation Operating System. *Preprints* **2026**. Preprint.
150. He, C.; Zhou, X.; Yu, X.; Zhang, L.; Zhang, Y.; Wu, Y.; Xiao, L.; Li, L.; Wang, D.; Xu, H.; et al. SSKG Hub: An Expert-Guided Platform for LLM-Empowered Sustainability Standards Knowledge Graphs. *arXiv preprint arXiv:2603.00669* **2026**. https://doi.org/10.48550/arXiv.2603.00669.
151. He, C.; Zhou, X.; Wang, D.; Yu, X.; Xiao, L.; Li, L.; Xu, H.; Liu, W.; Miao, C. KG4ESG: The ESG Knowledge Graph Atlas. *Preprints* **2026**. https://doi.org/10.20944/preprints202602.1970.v2.
152. He, C.; Zhou, X.; Wang, D.; Xu, H.; Liu, W.; Miao, C. ESGlass: Glass-Box ESG and Sustainability Reports. *Preprints* **2026**. https://doi.org/10.20944/preprints202603.2187.v1.
153. He, C.; Zhou, X.; Wu, Y.; Yu, X.; Zhang, Y.; Zhang, L.; Wang, D.; Lyu, S.; Xu, H.; Xiaoqiao, W.; et al. Esgenius: Benchmarking llms on environmental, social, and governance (esg) and sustainability knowledge. In Proceedings of the Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 2025, pp. 14612–14653. https://doi.org/10.18653/v1/2025.emnlp-main.739.
154. Zhang, L.; Zhou, X.; He, C.; Wang, D.; Wu, Y.; Xu, H.; Liu, W.; Miao, C. Mmesgbench: Pioneering multimodal understanding and complex reasoning benchmark for esg tasks. In Proceedings of the Proceedings of the 33rd ACM International Conference on Multimedia, 2025, pp. 12829–12836.
155. He, C.; Zhou, X.; Wang, D.; Xu, H.; Liu, W.; Miao, C. The Synthetic Media Exchange: When Lineage Becomes Currency. *Preprints* **2026**.
156. He, C.; Zhou, X.; Wang, D.; Xu, H.; Liu, W.; Miao, C. From Prompts to Portfolios: AI Agents as Agentic Multimedia Firms. *Preprints* **2026**.
157. He, C.; Zhou, X.; Wang, D.; Xu, H.; Liu, W.; Miao, C. Elevate Before You Eliminate: Firms Should Redesign High-Risk Roles Before Any AI-Attributed Layoffs **2026**.
158. He, C.; Zhou, X.; Wang, D.; Xu, H.; Liu, W.; Miao, C. Remote-Capable Knowledge Work Should Default to AI-Enabled Flexibility **2026**.
159. Wang, G.; Xie, Y.; Jiang, Y.; Mandlekar, A.; Xiao, C.; Zhu, Y.; Fan, L.; Anandkumar, A. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291* **2023**.
160. Shinn, N.; Cassano, F.; Gopinath, A.; Narasimhan, K.; Yao, S. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems* **2023**, *36*, 8634–8652.
161. Paranjape, B.; Lundberg, S.; Singh, S.; Hajishirzi, H.; Zettlemoyer, L.; Ribeiro, M.T. Art: Automatic multi-step reasoning and tool-use for large language models. *arXiv preprint arXiv:2303.09014* **2023**.
162. Mialon, G.; Fourrier, C.; Wolf, T.; LeCun, Y.; Scialom, T. Gaia: a benchmark for general ai assistants. In Proceedings of the The Twelfth International Conference on Learning Representations, 2023.
163. Koh, J.Y.; Lo, R.; Jang, L.; Duvvur, V.; Lim, M.; Huang, P.Y.; Neubig, G.; Zhou, S.; Salakhutdinov, R.; Fried, D. Visualwebarena: Evaluating multimodal agents on realistic visual web tasks. In Proceedings of the Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 881–905.
164. Kapoor, R.; Butala, Y.P.; Russak, M.; Koh, J.Y.; Kamble, K.; AlShikh, W.; Salakhutdinov, R. Omniact: A dataset and benchmark for enabling multimodal generalist autonomous agents for desktop and web. In Proceedings of the European Conference on Computer Vision. Springer, 2024, pp. 161–178.

 doi:10.20944/preprints202604.1860.v1

165. Zhang, C.; Yang, Z.; Liu, J.; Li, Y.; Han, Y.; Chen, X.; Huang, Z.; Fu, B.; Yu, G. Appagent: Multimodal agents as smartphone users. In Proceedings of the Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, 2025, pp. 1–20.
166. Wang, J.; Xu, H.; Ye, J.; Yan, M.; Shen, W.; Zhang, J.; Huang, F.; Sang, J. Mobile-agent: Autonomous multi-modal mobile device agent with visual perception. *arXiv preprint arXiv:2401.16158* **2024**.
167. Chu, X.; Qiao, L.; Lin, X.; Xu, S.; Yang, Y.; Hu, Y.; Wei, F.; Zhang, X.; Zhang, B.; Wei, X.; et al. Mobilevlm: A fast, strong and open vision language assistant for mobile devices. *arXiv preprint arXiv:2312.16886* **2023**.
168. Ran, D.; Wu, M.; Yu, H.; Li, Y.; Ren, J.; Cao, Y.; Zeng, X.; Lu, H.; Xu, Z.; Xu, M.; et al. Beyond pass or fail: Multi-dimensional benchmarking of foundation models for goal-based mobile UI navigation. *arXiv preprint arXiv:2501.02863* **2025**.
169. Wu, Q.; Yang, Z.; Li, H.; Gao, P.; Liu, W.; Luan, J. MobileBench-OL: A Comprehensive Chinese Benchmark for Evaluating Mobile GUI Agents in Real-World Environment. *arXiv preprint arXiv:2601.20335* **2026**.
170. Li, Y.; Liu, Y.; Lu, H.; Xia, Z.; Wang, H.; Han, K.; Yang, C.; Wu, J.; Xu, J.; Shi, R.; et al. GUI-CEval: A Hierarchical and Comprehensive Chinese Benchmark for Mobile GUI Agents. *arXiv preprint arXiv:2603.15039* **2026**.
171. Nie, H.; Liu, X.; Bai, Y.; Wang, Y.; Liu, Y.; Yao, Q.; Wang, Z. PSPA-Bench: A Personalized Benchmark for Smartphone GUI Agent. *arXiv preprint arXiv:2603.29318* **2026**.
172. Sun, J.; Li, M.; Zhang, Y.; Niu, J.; Wu, Y.; Jin, R.; Lei, S.; Tan, P.; Zhang, Z.; Wang, R.; et al. AmbiBench: Benchmarking Mobile GUI Agents Beyond One-Shot Instructions in the Wild. *arXiv preprint arXiv:2602.11750* **2026**.
173. Coase, R.H. The nature of the firm (1937). *The nature of the firm: origins, evolution, and development* **1993**, pp. 18–33.
174. Williamson, O.E. The economics of organization: The transaction cost approach. *American journal of sociology* **1981**, *87*, 548–577.
175. Malone, T.W.; Yates, J.; Benjamin, R.I. Electronic markets and electronic hierarchies. *Communications of the ACM* **1987**, *30*, 484–497.
176. Bakos, J.Y. A strategic analysis of electronic marketplaces. *MIS quarterly* **1991**, *15*, 295–310.
177. Rochet, J.C.; Tirole, J. Platform competition in two-sided markets. *Journal of the european economic association* **2003**, *1*, 990–1029.
178. Rochet, J.C.; Tirole, J. Two-sided markets: a progress report. *The RAND journal of economics* **2006**, *37*, 645–667.
179. Boudreau, K. Open platform strategies and innovation: Granting access vs. devolving control. *Management science* **2010**, *56*, 1849–1872.
180. Eisenmann, T.; Parker, G.; Van Alstyne, M. Platform envelopment. *Strategic management journal* **2011**, *32*, 1270–1285.
181. Tiwana, A. *Platform ecosystems: Aligning architecture, governance, and strategy*; Newnes, 2013.
182. Gawer, A. Bridging differing perspectives on technological platforms: Toward an integrative framework. *Research policy* **2014**, *43*, 1239–1249.
183. Constantinides, P.; Henfridsson, O.; Parker, G.G. Introduction—platforms and infrastructures in the digital age, 2018.
184. Hein, A.; Schreieck, M.; Riasanow, T.; Setzke, D.S.; Wiesche, M.; Böhm, M.; Krcmar, H. Digital platform ecosystems. *Electronic markets* **2020**, *30*, 87–98.
185. McIntyre, D.; Srinivasan, A.; Afuah, A.; Gawer, A.; Kretschmer, T. Multisided platforms as new organizational forms. *Academy of management perspectives* **2021**, *35*, 566–583.
186. Ghazawneh, A.; Henfridsson, O. Balancing platform control and external contribution in third-party development: the boundary resources model. *Information systems journal* **2013**, *23*, 173–192.
187. Eaton, B.; Elaluf-Calderwood, S.; Sørensen, C.; Yoo, Y. Distributed Tuning of Boundary Resources: The Case of Apple's iOS Service System1. *MIS quarterly* **2015**, *39*, 217–243.
188. Chen, L.; Yi, J.; Li, S.; Tong, T.W. Platform governance design in platform ecosystems: Implications for complementors' multihoming decision. *Journal of management* **2022**, *48*, 630–656.
189. Costabile, C. Digital platform ecosystem governance of private companies: Building blocks and a research agenda based on a multidisciplinary, systematic literature review. *Data and Information Management* **2024**, *8*, 100053.