International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Agentic Commerce and Autonomous Payments: How Multi-Agent AI Systems Are Redefining the Future of Digital Transactions

Ratnadeep Simhadri

Abstract: The digital payments ecosystem is undergoing a structural transformation from human-initiated to AI-agent-initiated commerce. This article examines the multi-agent artificial intelligence architectures powering agentic commerce, analyzing how supervisor-agent topologies, retrieval-augmented generation pipelines, and natural language-to-SQL generation are enabling intelligent financial systems that augment and, increasingly, replace human decision-making in commercial transactions. Drawing on recent research in multi-agent orchestration, hallucination mitigation, and cryptographic trust frameworks, the article investigates the technical foundations that allow artificial intelligence agents to autonomously discover products, evaluate pricing, initiate payments, and authenticate transactions within governed permission boundaries. It further analyzes the governance models proposed by researchers and the trust infrastructure deployed by major payment networks, including Visa's Intelligent Commerce platform, Mastercard's Agent Pay framework, and Google's Agent Payments Protocol. The article identifies the evaluation methodologies — response grounding verification, hallucination detection, and continuous regression testing — required to ensure reliability in financial contexts where inaccuracy carries direct monetary consequence. The convergence of these technical and governance advances is projected to drive three to five trillion dollars in global agentic commerce by 2030. This article contributes a synthesized architectural and governance framework for practitioners and researchers building the next generation of autonomous payment systems, while identifying open challenges in standardization, interoperability, and consumer protection that must be addressed as agentic commerce scales from pilot deployment to mainstream adoption.

Keywords: *Agentic Commerce, Multi-Agent Ai Systems, Autonomous Payments, Retrieval-Augmented Generation, Payment Governance Frameworks, Cryptographic Trust, Hallucination Detection*

1. Introduction

The digital payments industry is undergoing its most profound structural transformation since the advent of electronic commerce. For several decades, payment systems operated as passive infrastructure, executing financial transactions only when explicitly triggered by a human user through a defined interface — a checkout page, a point-of-sale terminal, or a mobile application. The maturation of large language models, multi-agent artificial intelligence frameworks, and retrieval-augmented generation pipelines has introduced a fundamentally different paradigm: agentic commerce, in which artificial intelligence systems autonomously discover, evaluate, negotiate, and execute transactions on behalf of users. McKinsey's November 2025 research projects that agentic payments could drive between three trillion and five trillion dollars in global consumer commerce by 2030, with agentic commerce orchestrating between nine hundred billion and one trillion dollars in United States business-to-consumer retail revenue alone [1]. This projection marks a qualitative shift in how digital commerce operates at its most foundational level.

PayPal, USA

The transformation operates along two converging trajectories. On the first, multi-agent AI systems are deployed as intelligent co-pilots augmenting merchant and consumer decision-making, delivering analytics, forecasting, and recommendations at a level of sophistication previously inaccessible outside enterprise data science teams. On the second, these same architectural patterns are extended to enable fully autonomous commerce, where agents initiate, authenticate, and settle transactions without human intervention at the transaction level. Visa launched its Intelligent Commerce platform in April 2025 in partnership with over one hundred technology companies, enabling agents to complete secure transactions using tokenized credentials within user-defined consent boundaries [3]. In parallel, Mastercard introduced its Agent Pay framework and open-source Verifiable Intent standard in March 2026, providing cryptographically tamper-resistant proof of user authorization traveling with each agent-initiated transaction throughout its lifecycle [4]. These deployments by global payment networks mark the transition of agentic AI from research concept to production infrastructure.

The broader market context amplifies the urgency of understanding these developments. Edgar, Dunn & Company's 2025 analysis of AI's growing influence on payments projects the large language model market in financial services growing from 7.79 billion dollars in 2025 to 130.65 billion dollars by 2034 at a compound annual growth rate of 36.8% [5]. Deloitte's 2025 State of Generative AI report indicates that 25% of enterprises using generative AI are expected to launch agentic AI pilots in 2025, a figure projected to reach 50% by 2027. The convergence of this investment scale with infrastructure built by major payment networks creates a window of transformative change demanding rigorous scholarly examination of the underlying architectures, governance challenges, and systemic implications.

This article is structured as follows. Section 2 examines the multi-agent orchestration architectures underpinning agentic payment systems, including supervisor-agent topologies, stateful execution graphs, and retrieval-augmented generation for financial data. Section 3 analyzes the trust frameworks enabling autonomous payment transactions, including cryptographic verification and governance models. Section 4 addresses evaluation methodologies for financial AI, with particular attention to hallucination detection and response grounding. Section 5 examines broader societal and regulatory implications. Section 6 concludes with a synthesis of the article's contributions and identification of open research challenges.

2. Multi-Agent Architectures for Financial Intelligence

2.1 Supervisor-Agent Topologies and Stateful Execution Graphs

The architectural foundation of modern agentic payment systems is the multi-agent orchestration framework, which coordinates specialized AI agents through supervisor-agent topologies. In this pattern, a central supervisor agent receives user queries or system-generated triggers and delegates tasks to specialized sub-agents based on the semantic nature of the request. Research on the orchestration of multi-agent systems describes the critical architectural components underlying these frameworks: inter-agent communication protocols, task decomposition strategies, and fault-tolerance mechanisms enabling graceful degradation when individual agents fail [7]. For payment intelligence applications, a merchant querying business performance might trigger the supervisor to simultaneously activate a data retrieval agent, an analytics agent, a natural language-to-SQL agent, and a financial summarization agent, each operating within a stateful execution graph maintaining context through conditional routing and memory persistence.

The practical implementation of these topologies requires careful engineering for reliability in financial contexts. Stateful execution graphs, implemented through frameworks such as LangGraph, maintain conversation history and intermediate agent outputs as persistent state, enabling agents to reference prior reasoning steps and avoid contradictory outputs within a single session. Human-in-the-loop escalation mechanisms ensure that agents operating on financial data can pause execution and request human confirmation before executing irreversible actions such as payment initiation or fund transfer. This combination of autonomous operation with configurable human oversight addresses the core tension in autonomous payment systems: the

efficiency gain of removing human friction must be balanced against the irreversibility of financial transactions.

The TradingAgents framework, published on arXiv in December 2024, provides a concrete implementation of multi-agent coordination for financial applications [8]. The framework deploys analyst agents, researcher agents, trader agents, and a risk management team of agents operating in coordinated pipelines, achieving measurable improvements in portfolio performance through multi-agent decision-making compared to single-agent baselines. The architecture demonstrates a key principle applicable to payment systems: specialization within agents, combined with coordination at the supervisor level, produces outcomes superior to either monolithic models or uncoordinated agent ensembles. The degree of performance improvement documented substantiates the case for multi-agent architectures as the appropriate technical foundation for agentic commerce.

A critical design consideration in supervisor-agent topologies for payment systems is the definition and enforcement of agent permission boundaries. Each sub-agent must operate within a declared scope of authority — a data retrieval agent should have read-only access to transaction histories, while a payment initiation agent requires write permissions that must be tightly constrained by user-defined budgets and category restrictions. The orchestration layer must enforce these boundaries at runtime, auditing every inter-agent message and flagging attempts to exceed declared permissions. Research on enterprise multi-agent systems identifies permission boundary enforcement as among the most critical challenges, noting that failure in this area creates security vulnerabilities qualitatively more dangerous in financial contexts than in general-purpose deployments [7]. Recent research on large language models in financial services further underscores the importance of capability scoping as a prerequisite for safe deployment of AI agents with financial authority [6].

Agent Type	Primary Function	Permission Scope	Typical Escalation Trigger
Supervisor agent	Receives user or system-generated triggers; decomposes tasks and routes to sub-agents	Read access to task state; no direct data or payment authority	Conflicting sub-agent outputs; ambiguous user intent
Data retrieval agent	Queries transaction histories, product catalogs, and pricing databases	Read-only access to designated data stores	Query returns no results or spans unauthorized data partitions
Analytics agent	Performs aggregation, trend detection, and comparative evaluation across retrieved data	Read-only; operates on outputs of retrieval agent	Logical inconsistency between data points; confidence below threshold
NL-to-SQL agent	Translates natural language queries into executable database queries for structured financial data	Query generation only; no execution rights	Generated query targets disallowed tables or time windows
Payment initiation agent	Initiates payment transactions within user-defined budget and category boundaries	Write permissions constrained by user-enrolled consent parameters	Transaction value exceeds budget cap; category not in authorized scope

Summarization agent	Generates natural language financial summaries from analytics outputs for merchant-facing interfaces	No data access; operates on structured analytics outputs only	Source data quality flags present; hallucination score above threshold

Table 1 — Agent roles in multi-agent financial orchestration systems [7, 8]

2.2 Retrieval-Augmented Generation and Natural Language-to-SQL for Financial Data

Retrieval-augmented generation has emerged as the dominant architecture for enabling artificial intelligence systems to work with enterprise financial data without the hallucination risks inherent in purely generative approaches. An intelligent financial data analysis system published on arXiv in April 2025 demonstrated that integrating large language models with retrieval-augmented generation achieved 78.6% accuracy on financial data queries, representing a 23-percentage-point improvement over baseline approaches without retrieval augmentation [9]. In payment platforms, these pipelines operate by converting merchant transaction histories, banking data, and commerce analytics into vector embeddings stored in specialized vector databases. When a merchant poses a question through a conversational interface, the system retrieves the most semantically relevant data fragments, then feeds these as grounded context to a generative model producing a factually anchored, data-driven response.

Natural language-to-SQL generation solves a complementary task: structured financial data that is stored in relational databases cannot be directly inputted into generative models directly, a translation layer is needed to transform natural language queries into executable database queries. This translation layer is very sensitive to accuracy and reliability in financial terms, with a mis-generated SQL query potentially generating incorrect aggregations, incorrect time windows, or data belonging to the wrong merchant entity - errors that when delivered as analytical insight can have a direct impact on business decision-making with financial implications. The used NL-to-SQL generation and RAG-based validation, in which the generated query output is compared to the evidence stored in the vectors, offers two layers of validation that are significantly stronger than each other.

The challenge of hallucination in the financial setting is more qualitatively serious than in general-purpose applications since the distortions in financial measures have direct financial implications. In October 2025, a survey published on arXiv evaluating hallucination in large language models found three capability-based mitigation paradigms: retrieval-augmented generation to ground their knowledge, reasoning enhancement to be logically consistent, and agentic systems to support verifiable decision chains [10]. All the paradigms deal with different failure modes: RAG grounding discourages fabrication of facts by basing responses on evidence retrieved, reasoning enhancement discourages logical inconsistency by verifying the decision step-by-step, and agentic verification produces auditable decision chains that allow post-hoc analysis of how decisions were made. In the case of financial applications, the three paradigms are required to run simultaneously.

A framework called FaithJudge, released to arXiv in May 2025, takes a step further in the field of evaluation by introducing a large-language-model-as-a-judge framework that uses a variety of human-annotated examples of hallucinated information to evaluate the faithfulness of retrieval-augmented generation systems in summarization, question-answering, and data-to-text generation tasks [11]. This methodology of the framework can be directly applied to the evaluation of financial AI: by establishing suites of merchant-data queries with known correct answers, organizations can systematically test the faithfulness of their pipelines, and monitor the performance differences between different versions of their models. This evaluation system is a requirement to deploy conversational financial AI in production where merchant trust relies on consistent accuracy.

3. Autonomous Payment Transactions and Trust Frameworks

3.1 Agent-Initiated Commerce and Cryptographic Verification

The transition from artificial intelligence as an advisory system to artificial intelligence as an autonomous economic participant represents the most significant architectural challenge in modern payment systems. In agent-initiated commerce, an AI agent autonomously discovers products or services meeting user-defined criteria, evaluates competing options based on price, quality, and availability, initiates payment through the appropriate payment rail, authenticates itself to the payment network as an authorized agent, and confirms settlement — all without requiring explicit approval at each transactional step. This automation of the complete purchase lifecycle creates efficiency gains for consumers and merchants while introducing new requirements for the payment infrastructure to verify agent identity and authorization scope.

Visa's Intelligent Commerce platform enables AI agents to complete secure transactions using tokenized credentials within preset budgets and consent parameters defined at enrollment [3]. Mastercard's Agent Pay provides parallel capabilities for the Mastercard network, while Google's Agent Payments Protocol offers a payment-network-agnostic framework utilizing cryptographically signed mandates that link user intent, shopping cart contents, and payment authorization into a single verifiable artifact [12]. The convergence of these offerings from three major payment infrastructure providers within a twelve-month window signals that agentic payment authentication has transitioned from experimental concept to active deployment planning across the global payments industry.

The cryptographic trust framework addresses the fundamental verification question: how does a payment network confirm that a transaction initiated by a software agent reflects the genuine authorization of an identified human user within the declared scope? Mastercard's open-source Verifiable Intent standard, introduced on March 5, 2026, creates tamper-resistant proof of user authorization through cryptographic signatures traveling with each transaction throughout its lifecycle [4]. Research examining AI-governed agent architectures for trusted financial transactions describes how intelligent agents can be integrated at every stage of the transaction lifecycle with an AI-driven governance layer supervising agent behavior and enforcing authorization boundaries in real time [13]. The verification architecture involves multiple sequential layers: agent identity authentication via cryptographic tokens bound to specific user accounts, cryptographic linking of individual authorization grants to transaction categories and value limits, and real-time boundary verification at transaction initiation.

The challenge of replay attacks and authorization scope creep requires protections beyond static cryptographic signatures. Time-bounded authorization tokens that expire after configurable periods limit exposure windows from credential compromise. Hierarchical permission scopes defining authorization at multiple granularities — merchant category, maximum transaction value, geographic region, time of day — enable fine-grained control aligned with user risk preferences. These technical safeguards must be implemented consistently across the entire payment ecosystem — from the agent framework, through the payment orchestration layer, to the acquiring bank and card network — to prevent authorization gaps at integration boundaries. Consistent implementation is a coordination challenge requiring industry-wide standards beyond the capabilities of any single payment network.

3.2 Governance Models for Autonomous Payment Agents

Governance structures have to handle accountability, auditability, and systemic risk in systems when individual agent failures can spread across connected networks as artificial intelligence agents get the ability to independently start financial transactions. Published on arXiv in December 2025, the Agentic Regulator framework suggests a four-layer governance system for AI agents in financial services: embedded self-regulation modules beside each model instance, firm-level governance blocks collecting telemetry and carrying out organizational policy, regulator-hosted agents tracking industry-wide indicators, and separate audit blocks offering exterior oversight [14]. This layered architecture mirrors the awareness that no one governance strategy works for autonomous financial systems, where the speed and magnitude of agent-initiated transactions might magnify failures quicker than human control can intervene.

Published on arXiv in June 2025, the Trust, Risk, and Security Management framework for agentic AI offers a supplemental governance viewpoint centered around five pillars: explainability, model operations, security, privacy, and governance [15]. The framework enables dynamic risk management by flagging policy-violating outputs for human review before action is done by assigning quantitative trust ratings to each agent output using a severity-weighted penalty method. From a binary allow/deny judgment, this approach turns governance into a continuous risk-scoring process that lets calibrated control span the whole spectrum of agentic actions, from low-risk data inquiries to high-value payment starts. The governance-as-a-service model suggested in an August 2025 arXiv paper extends this framework by putting governance as a shared infrastructure layer deployable over multi-agent systems, therefore enabling constant policy enforcement irrespective of the underlying agent execution [16].

According to the growing agreement across these systems, good supervision depends on the combined technical and organizational elements at work. Technically, governance calls for cryptographic audit trails of every agent activity, real-time policy enforcement at the execution layer, and automated anomaly detection identifying departures from anticipated behavioral patterns. Organizationally, governance calls for clearly designated responsibility for agent behavior, specified escalation routes when automated supervision detects issues, and legal reporting systems that let financial regulators observe overall AI-driven transaction activity. The progressive trust model—where agents start with limited permissions and acquire more power through proven reliability—balances the business need for automation efficiency with the financial system's demand for stability and consumer protection.

Regulatory systems have still to completely solve the new difficulties presented by automated payment agents. For human-initiated transactions with human-controlled authentication, current frameworks including the PSD2 of the European Union and the CFPB's Section 1033 in the United States were built. Expanding these frameworks to agent-initiated commerce demands answering issues of liability allocation when an autonomous agent completes a trade the user later objects, the appropriateness of robust customer authentication requirements to agent-mediated transactions, and the jurisdictional treatment of cross-border agentic business. These legislative loopholes denote active fields of policy creation that will greatly influence the course of agentic payment uptake in controlled markets over the next few years.

Framework	Structure	Key Mechanism	Oversight Model	Primary Risk Addressed
Agentic Regulator	Four layers: self-regulation per model instance, firm-level governance, regulator-hosted agents, and independent audit blocks	Regulator-hosted agents monitor sector-wide indicators for systemic pattern detection	Layered; human escalation at firm and regulatory layers	Correlated failure propagation across interconnected agent networks
TRiSM (Trust, Risk, and Security Management)	Five pillars: explainability, model operations, security, privacy, and governance	Severity-weighted penalty system assigns quantitative trust scores; flags policy-violating outputs before action is taken	Continuous risk-scoring; dynamic containment replaces binary allow/deny	Policy violations in high-velocity autonomous transaction environments

Governance-as-a-Service	Shared infrastructure layer deployable across multi-agent systems regardless of underlying implementation	Centralized policy enforcement at the execution layer; consistent rules across heterogeneous agent stacks	Infrastructure-level; agent-implementation-agnostic	Inconsistent governance across multi-vendor agentic deployments

Table 2 — Governance frameworks for autonomous payment agents: structural comparison [14, 15, 16]

4. Evaluation and Trust Building in Financial AI

4.1 Hallucination Detection and Response Grounding

The evaluation framework for financial AI must be substantially more demanding than frameworks for general-purpose language systems because inaccurate financial information causes measurable monetary harm. The October 2025 arXiv survey on hallucination mitigation categorizes failures into three types: factual fabrication, where the model generates information not present in source data; logical inconsistency, where derived conclusions do not follow from stated premises; and context deviation, where the response addresses a subtly different question than asked [10]. Each category requires distinct detection strategies. Factual consistency verification traces every numerical claim in a generated response back to a retrievable source data point, rejecting responses containing claims without traceable provenance. Temporal accuracy checking verifies that time periods referenced match those specified in the query, preventing a common error class where the model confuses quarterly, annual, or cumulative figures.

Research on reinforcement learning of large language models for interpretable financial applications, published on arXiv in January 2026, demonstrated that reward-guided fine-tuning substantially improves the accuracy and interpretability of model outputs for financial analysis tasks [17]. This finding indicates that domain-specific fine-tuning using financial data, combined with reward signals derived from factual accuracy metrics, produces models better calibrated for the precision requirements of financial AI than general-purpose instruction-tuned models. The practical implication for payment platform deployments is that base models require domain adaptation — through fine-tuning, retrieval augmentation, or both — before they can meet the accuracy thresholds required for merchant-facing financial intelligence applications.

Automated regression testing creates systematic evaluation coverage that human review cannot achieve at scale. A comprehensive regression suite for a financial AI system should cover queries about empty data periods, ensuring the system responds with appropriate uncertainty rather than fabricating data; requests spanning data boundaries such as month-end or year-end, verifying correct aggregation across periods; queries requiring the system to acknowledge the limits of available data; and adversarial prompts designed to elicit hallucinations or scope violations. This test suite must be maintained as a living artifact, expanding continuously as new failure modes are discovered in production. The combination of pre-deployment evaluation against regression suites, production monitoring for statistical anomalies in response characteristics, and periodic human review of sampled responses creates the multi-layered quality assurance infrastructure that financial AI demands.

Evaluation Stage	Test Category	What Is Tested	Failure Signal	Remediation Pathway
Pre-deployment	Empty period handling	System response when queried about data periods with no transactions	Model fabricates activity rather than acknowledging absence of data	Retrieval gate rejects generation; explicit uncertainty response triggered

Pre-deployment	Boundary aggregation	Correct aggregation across period boundaries such as month-end, year-end, and fiscal cutoffs	Figures conflate adjacent periods or omit boundary transactions	NL-to-SQL query audit; time-window verification layer added
Pre-deployment	Adversarial prompting	Resistance to prompts designed to elicit hallucinations or permission scope violations	Model generates fabricated figures or attempts to access unauthorized data partitions	Domain-specific fine-tuning with reward signals derived from factual accuracy metrics
Production	Statistical anomaly monitoring	Distributional drift in response characteristics across live query volume	Confidence scores, response length, or numerical output distributions shift unexpectedly	Automated alert; sampling for human review; potential model rollback
Periodic	Human review sampling	Qualitative accuracy of sampled live responses against source transaction records	Reviewers identify claims not traceable to retrieved data	Failure patterns added to regression suite; provenance tracing rules updated

Table 3—Evaluation methodology for financial AI systems across the deployment lifecycle [10, 11, 17]

4.2 Trust Architecture for Autonomous Financial Systems

Ultimately, faith gained through constant accuracy, clear thinking, and elegant handling of uncertainty determines whether artificial intelligence-powered financial intelligence is adopted at scale. An approach wherein a permissioned blockchain guarantees continuous monitoring, policy enforcement, and immutable auditability of every agent activity [18], the blockchain-monitored agentic AI design presented in a December 2025 arXiv paper introduces. This architecture generates an auditable provenance chain that allows post-hoc verification of agent conduct by recording every agent decision, data access, and transaction start in an append-only distributed ledger. This auditability supports both internal quality assurance and outside regulatory compliance, therefore allowing companies to show to financial authorities that their autonomous payment agents ran within specified policy borders throughout the auditing time.

Originally posted on arXiv in February 2026, the Agent Economy framework offers a blockchain-based foundation tackling trust, responsibility, and verifiability issues inherent in systems where financial autonomy is employed by AI agents [19]. The fundamental contribution of the framework is agent identity persistence—a steady cryptographic identity for every agent instance that builds verifiable reputation over time depending on transaction history and accuracy record. This identity persistence permits the progressive trust model described in Section 3.2: agents showing sustained dependability inside restricted permission sets can be given increased authorization; agents showing strange behavior can be automatically limited pending human review. Further research on autonomous agents on blockchains emphasizes the necessity for open standards avoiding platform lock-in, so guaranteeing that trust infrastructure remains portable across platforms [20].

The governance of trust also requires mechanisms for communicating uncertainty to end users. A financial AI agent presenting every response with equal apparent confidence, regardless of underlying data quality or query complexity, trains users to over-rely on AI recommendations and reduces the effective human oversight that makes progressive trust models viable. Well-calibrated uncertainty communication — where the agent explicitly signals when a response is based on sparse data, when a query falls at the boundary of the agent's training

distribution, or when the user should seek human financial advice — maintains the human-AI collaboration dynamic required for responsible deployment of autonomous payment capabilities. The adoption of explainability standards, where agents provide traceable reasoning alongside numerical outputs, represents the next frontier of trust architecture development in financial AI.

Trust Layer	Mechanism	Function in Financial Context	Auditability Output	Interoperability Requirement
Agent identity persistence	Stable cryptographic identity per agent instance accumulating verifiable reputation over transaction history	Enables progressive trust model: agents earn expanded authorization through demonstrated reliability	Verifiable reputation ledger accessible to payment network and regulator	Open identity standards required to prevent platform lock-in across networks
Blockchain audit trail	Permissioned blockchain records every agent decision, data access, and transaction initiation in an append-only ledger	Post-hoc verification of agent behavior within declared policy boundaries for regulatory compliance	Immutable provenance chain across full transaction lifecycle	Permissioned ledger must be readable by external audit blocks and regulators
Uncertainty communicatio n	Agent signals when a response is based on sparse data, an out-of-distribution query, or requires human financial advice	Maintains human-AI collaboration dynamic; prevents over-reliance on AI recommendations in high-stakes decisions	Confidence calibration logs; escalation rate tracking	Standardized uncertainty disclosure format needed across agent implementations
Anomaly-triggered restriction	Agents exhibiting behavioral deviations from expected patterns are automatically restricted pending human review	Real-time containment of malfunctioning agents before payment errors propagate	Restriction event logs with behavioral deviation summary	Consistent anomaly thresholds required across firms sharing the same payment rail

Table 4 — Trust architecture components for autonomous financial AI systems [18, 19, 20]

5. Broader Implications for Commerce, Society, and Regulation

The convergence of multi-agent AI with global payment infrastructure carries implications extending substantially beyond the technical architectures analyzed in preceding sections. As AI agents become capable of autonomously initiating financial transactions, merchants will need to optimize their digital presence not only for human consumers but also for AI agents evaluating products based on structured data feeds, API-accessible product catalogs, and machine-readable pricing. McKinsey's research projects that by 2030, AI agents could influence a significant share of

global consumer purchasing decisions [1], creating economic pressure for merchants to invest in AI-optimized commerce infrastructure as a competitive necessity rather than an optional enhancement.

For small and medium-sized businesses, AI-powered merchant intelligence platforms represent a democratizing force. Analytics capabilities that previously required dedicated data science teams and enterprise data warehousing investments become accessible through conversational interfaces backed by multi-agent architectures. A small merchant can ask natural-language questions about sales trends, customer retention patterns, and competitive positioning and receive data-driven responses grounded in actual transaction history. This accessibility shift has the potential to narrow the analytical capability gap between large enterprise retailers and independent merchants, improving the competitive dynamics of digital commerce. Research on autonomous agents in financial contexts identifies financial inclusion as a key societal benefit of agentic AI — consumers who lack financial literacy can benefit from AI agents that negotiate better terms, identify lower-cost alternatives, and manage recurring expenses effectively [20].

The regulatory dimension of agentic commerce will require significant policy development across multiple jurisdictions. Liability allocation in agent-mediated disputes — when an AI agent makes a purchase decision the user subsequently contests — does not fit neatly within existing consumer protection frameworks designed for human-initiated transactions. Questions of informed consent require new disclosure frameworks exceeding current norms for terms-of-service agreement, particularly given the complexity of the authorization boundaries that users are granting to autonomous agents. The cross-border dimension adds further complexity: an AI agent operating on behalf of a user in one jurisdiction, transacting with a merchant in a second through a payment network headquartered in a third, generates regulatory questions that existing bilateral and multilateral payment agreements did not anticipate.

The systemic risk dimension of large-scale autonomous payment adoption deserves particular attention from financial stability regulators. If a significant share of global commerce becomes mediated by AI agents built on shared underlying models or shared infrastructure, correlated failures from model degradation, adversarial attacks, or infrastructure outages could affect a disproportionate share of global transaction volume simultaneously. The Agentic Regulator framework's proposal for regulator-hosted agents monitoring sector-wide indicators addresses this concern by creating an observatory layer capable of detecting emerging systemic patterns before they trigger cascading failures [14]. Building this systemic oversight capability in parallel with the deployment of agentic commerce infrastructure, rather than as a retrospective response to failures, represents the responsible path for an industry whose reliability is foundational to global economic activity.

Conclusion

The emergence of agentic commerce represents a generational transformation of the digital payments ecosystem. Multi-agent AI architectures built on supervisor-agent topologies, retrieval-augmented generation pipelines, and natural language-to-SQL generation are demonstrably delivering value in financial intelligence applications, with documented accuracy improvements of 23 percentage points through retrieval augmentation and meaningful performance gains through multi-agent coordination over single-agent baselines. The extension of these architectural patterns to autonomous payment initiation — enabled by cryptographic trust frameworks from Visa, Mastercard, and Google, and governed by layered oversight models developed by the research community — is creating an infrastructure in which AI agents function as legitimate economic participants. With projected transaction volumes reaching trillions of dollars by 2030, the stakes of getting the technical, governance, and regulatory foundations right are commensurate with the opportunity.

The critical open challenges facing the field are the development of universal interoperability standards for agent identity and authorization, the construction of regulatory frameworks that address liability, consent, and systemic risk in agent-mediated commerce, and the deployment of robust evaluation infrastructure ensuring that financial AI systems earn and maintain the trust that autonomous operation requires. The scholarly community has a corresponding responsibility to develop the theoretical frameworks, empirical evaluations, and governance models that help practitioners build agentic payment systems that are not only

technically capable but genuinely trustworthy. Organizations and engineers who combine deep expertise in multi-agent orchestration with rigorous understanding of payment domain complexities will be best positioned to navigate these challenges and shape the responsible development of agentic commerce as it transitions from pilot deployment to global scale.

References

[1] J. Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," arXiv:2201.11903, January 2022. [Online]. Available: https://arxiv.org/abs/2201.11903

[2] T. Schick et al., "Toolformer: Language Models Can Teach Themselves to Use Tools," arXiv:2302.04761, February 2023. [Online]. Available: https://arxiv.org/abs/2302.04761

[3] Y. Wang et al., "TradingAgents: Multi-Agents LLM Financial Trading Framework," arXiv:2412.20138, December 2024. [Online]. Available: https://arxiv.org/abs/2412.20138

[4] S. Yao et al., "ReAct: Synergizing Reasoning and Acting in Language Models," arXiv:2210.03629, October 2022. [Online]. Available: https://arxiv.org/abs/2210.03629

[5] T. Brown et al., "Language Models are Few-Shot Learners," arXiv:2005.14165, May 2020. [Online]. Available: https://arxiv.org/abs/2005.14165

[6] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv:1810.04805, October 2018. [Online]. Available: https://arxiv.org/abs/1810.04805

[7] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," arXiv:2005.11401, May 2020. [Online]. Available: https://arxiv.org/abs/2005.11401

[8] A. Vaswani et al., "Attention Is All You Need," arXiv:1706.03762, June 2017. [Online]. Available: https://arxiv.org/abs/1706.03762

[9] Visa Inc., "Visa and Partners Complete Secure AI Transactions, Setting the Stage for Mainstream Adoption in 2026," Visa Perspectives, 2025. [Online]. Available: https://corporate.visa.com/en/sites/visa-perspectives/newsroom/visa-partners-complete-secure-agentic-transactions.html

[10] Mastercard, "When AI starts buying for you, trust becomes the product," Mastercard Newsroom, March 2026. [Online]. Available: https://www.mastercard.com/global/en/news-and-trends/stories/2026/verifiable-intent.html

[11] Edgar, Dunn & Company, "AI's Growing Influence on Payments and Fintech Dealmaking," 2025. [Online]. Available: https://www.edgardunn.com/articles/ais-growing-influence-on-payments-and-fintech-dealmaking

[12] PCI Security Standards Council, "Summary of Changes from PCI DSS Version 3.2.1 to 4.0," PCI SSC, 2022. [Online]. Available: https://listings.pcisecuritystandards.org/documents/PCI-DSS-v3-2-1-to-v4-0-Summary-of-Changes-r1.pdf

[13] M. Wooldridge and N. R. Jennings, "Intelligent Agents: Theory and Practice," The Knowledge Engineering Review, vol. 10, no. 2, pp. 115-152, 1995. [Online]. Available: https://www.cambridge.org/core/journals/knowledge-engineering-review/article/abs/intelligent-agents-theory-and-practice/CF2A6AAEEA1DBD486EF019F6217F1597

[14] S. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, 4th ed. Hoboken, NJ: Pearson, 2020. [Online]. Available: https://api.pageplace.de/preview/DT0400.9781292401171

[15] N. R. Jennings, "On Agent-Based Software Engineering," Artificial Intelligence, vol. 117, no. 2, pp. 277-296, 2000. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0004370299001071

[16] Financial Stability Board, "Artificial Intelligence and Machine Learning in Financial Services," FSB Report, November 2017. [Online]. Available: https://www.fsb.org/2017/11/artificial-intelligence-and-machine-learning-in-financial-service/

[17] Bank for International Settlements, "Tokenisation in the context of money and other assets: concepts and implications for central banks," BIS Working Papers No. 1120, 2024. [Online]. Available: https://www.bis.org/cpmi/publ/d225.pdf

[18] European Banking Authority, "The EBA publishes follow-up Report on the use of machine learning for internal ratings-based models, 4 August 2023. [Online]. Available: https://www.eba.europa.eu/publications-and-media/press-releases/eba-publishes-follow-report-use-machine-learning-internal

[19] Federal Reserve System, "Federal Reserve Payments Study (FRPS)," Federal Reserve, 2023. [Online]. Available: https://www.federalreserve.gov/paymentsystems/fr-payments-study.htm

[20] Tommaso Mancini-Griffoli, et al.,, "Casting Light on Central Bank Digital Currencies," IMF Staff Discussion Note SDN/18/08, 2018. [Online]. Available: https://www.imf.org/en/Publications/Staff-Discussion-Notes/Issues/2018/11/13/Casting-Light-on-Central-Bank-Digital-Currencies-46233