Tabb FORUM
WHERE CAPITAL MARKETS SPEAK

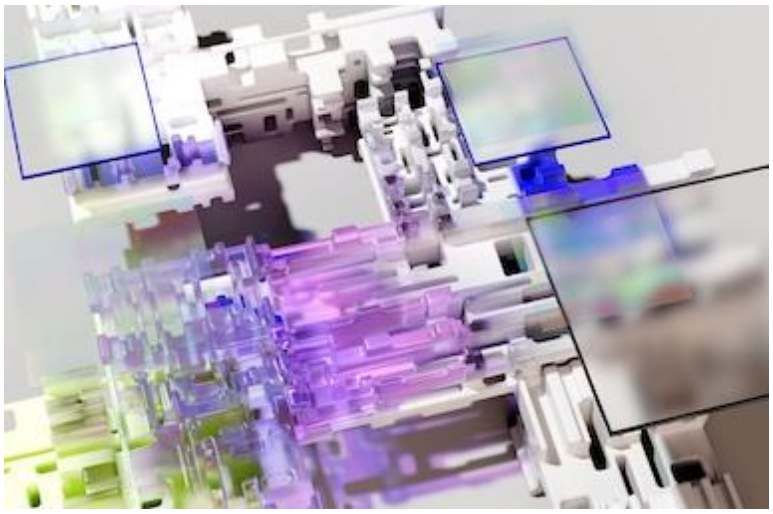
October 13, 2026

State of AI in the Capital Markets 2026: The Era of Execution at Scale



Sean Creamer
Contributing Writer
Follow | Profile | More

Share



AI in capital markets is moving beyond proofs of concept and into production-grade agentic systems that can execute trades, adjust risk models and automate compliance workflows. That shift is exposing a tougher challenge: individual productivity gains mean little if firms lack the architecture, controls and economics to scale them across the enterprise. As autonomous systems spread, traders, analysts and developers are increasingly becoming supervisors of AI-driven workflows, while firms wrestle with governance, testing and rapidly rising compute costs, writes financial technology journalist Sean Creamer.

Executive Summary

Artificial intelligence is reshaping capital markets, but the terms of engagement have changed. If 2024 and 2025 were defined by generative proofs of concept and retrieval-

augmented generation (RAG), 2026 marks the arrival of production-grade agentic architectures operating directly within live market workflows. As a March 2025 report from Moody's Analytics confirmed, the industry has shifted toward multi-agent frameworks capable of autonomously triggering pre-approved trades, dynamically adjusting risk models, and automating compliance reporting.

Yet as institutions push autonomous tools into high-volume environments, they collide with a sober operational reality. Gartner predicted in June 2025 that over 40% of enterprise agentic AI projects will be canceled by 2027 due to unclear business value and inadequate risk controls, underscoring the industry's struggle to create a bridge between pilot efficiency and sustainable production scale.

Handing portfolio managers, quants, and developers frontier AI models creates isolated pockets of speed, but without an enterprise architecture engineered to absorb that velocity, the return on investment quickly evaporates.

"AI is an unbelievable tool for individual productivity, but individual productivity does not directly translate to business benefits," points out **Medan Gabbay, Co-CEO of Quod Financial**. He says that an individual developer or trader can become highly efficient, but "unless the process, the architecture, and the whole organization is able to absorb and make use of that efficiency, you are essentially running in place."

In 2026, execution at scale is defined not by how fast a user can prompt a model, but by how effectively the firm's infrastructure can govern and execute the result.
This shift from passive desktop assistance to networked, multi-step agentic systems is forcing an enterprise-wide re-architecture of human capital. Across trading desks and quantitative operations, professionals are moving away from manual execution and pivoting toward supervisory oversight—a transition aligned with November 2025 guidance from the Authority for the Financial Markets (AFM) stressing that visible human involvement is essential to maintain trust in autonomous models.

For **David A. Chapa, Chief AI Strategist at Hitachi Vantara**, this evolution elevates domain expertise over replacing it. Because specialists are now planning parameters rather than manually executing them, Chapa notes that human workers are becoming "choreographers more than actual dancers participating in the theater." The mandate is not replacing human talent, as agentic systems still depend on professionals who understand market nuances well enough to steer probabilistic models into deterministic execution pipelines.

On the buy side, this dynamic is redefining the daily responsibilities of investment teams. As autonomous agents take on the heavy lifting of parsing unstructured research, extracting earnings nuances, and drafting allocation models, the traditional junior analyst is shifting from task executor to system manager.

Transforming junior associates into supervisors who oversee fleets of specialized agents is "an interesting problem, and I feel the whole industry is grappling with that," observes **Vinit Agrawal, Global Head of Investment AI Solutions at T. Rowe Price**. In his view, overseeing these interconnected systems functions "more like a manager or supervisory job," meaning human professionals will spend less time gathering data and more time rigorously auditing the logic of the machine.

This supervisory imperative extends directly into technology teams, where rapid AI code generation requires quality assurance and testing engineers to act as functional business analysts as opposed to being line-by-line debuggers.

In this TabbFORUM report, Contributing Writer Sean Creamer explores how capital markets institutions are navigating the complex gap between pilot efficiency and enterprise scale. The ensuing sections detail how firms are drawing the boundaries of action risk in the front office, managing the compute unit economics that threaten IT budgets, and building the composable data infrastructure required to execute safely in an increasingly autonomous market.

Section 1: The Compute Crisis, Affordability, and "Sticker Shock"

For the past two years, the capital markets treated generative artificial intelligence as an open-ended efficiency tool. But as institutions move agentic frameworks into live production, the bill has arrived. Executing AI at scale shatters the "free labor" fallacy—the assumption that AI simply eliminates operational headcount. Instead, institutions are discovering that AI transitions those expenses into volatile, usage-based cloud compute and inference bills that strain fixed IT budgets. The industry is facing severe compute "sticker shock," with Gartner predicting that over 40% of enterprise agentic AI projects will be canceled by 2027 due to runaway cloud infrastructure costs.

This financial friction often begins the moment natural language front-ends hit the trading floor. By drastically lowering the barrier to querying complex databases, conversational interfaces create a hidden backend compute multiplier. "Instead of writing a query or going to a particular screen, users can now ask their favorite chatbot 10 questions," observes **Satyam Kancharla, Chief Product Officer and EVP at Numerix**. He explains that "the chatbot can just tunnel through and drive so much more compute than we could imagine previously." This frictionless access transforms a single user interaction into a cascading series of cloud expenses.

To combat this compounding expense, quantitative development teams are abandoning simple end-of-month token tracking in favor of rigorous upfront unit economics.

For systematic asset managers operating under strict low-cost mandates, evaluating cloud inference spend must occur before a developer "writes up a lick of code." **Paul Musella, Head of Quantitative Development within the Quantitative Equity team at Vanguard**, explains that quants must closely marry their investment theses with engineering constraints by forecasting API calls, context engineering, and token burn upfront. Establishing this holistic discipline has fundamentally shifted how firms evaluate research, moving beyond basic tracking toward what Musella describes as a calculation of "cost per idea."

This budget bloat is frequently exacerbated by foundational data architecture. **Irene Galperin, Senior Advisor for Financial Services at InterSystems**, points out that dragging fragmented data across outdated systems multiplies compute costs exponentially.

“The operational benefit of eliminating legacy pipelines is that firms can stop turning every AI use case into a bespoke engineering project,” Galperin explains. Since traditional pipelines require repeated ETL cycles and data duplication before information is accessible, she notes that “all of this increases latency, infrastructure costs, control complexity, manual effort, and complicates governance.”

This paradigm forces tier-one institutions to implement strict governance policies that prevent unproven prototypes from blindly tapping into frontier model APIs. Rather than funding open-ended research sprints, buy-side leadership is demanding proof of return on investment before any workflow is granted production compute.

“On cost discipline, our principle is: scale only what earns it—sustained value, adoption, and disciplined unit economics,” outlines **Robi Krempus, Global Head of Data & AI at Manulife Wealth and Asset Management**. To execute this, Manulife manages costs through shared infrastructure, reusable capabilities, model selection and right-sizing, and continuous consumption monitoring to focus on “reducing ‘compute per outcome’ over defaulting to the largest available model.” By enforcing these hurdles, institutions ensure that autonomous tools are deployed exclusively where operational savings heavily outpace infrastructure costs.

For sell-side and execution venues, surviving this compounding consumption curve requires treating compute metering with the same governance rigor applied to traditional balance sheets. **Andrew Shannon, Head of AI and Cloud at TP ICAP**, reframes the street’s spiraling token bills as an operational breakdown rather than a technological limitation.

“The ‘sticker shock’ is real and it is mostly a governance failure,” Shannon argues, observing that consumption-based AI spend behaves exactly like enterprise cloud did a decade ago: “easy to start, hard to see, and it compounds. If you cannot meter it, you cannot control it.” To maintain cost discipline across continuous workloads, TP ICAP enforces a central “control tower” that meters unit economics while tying engineering funding directly to commercial outcomes—an operating model Shannon summarizes as “control before scale.”

Addressing the high cancellation rates predicted across the sector, Shannon notes that the primary failure mode to avoid “is a long tail of half-funded pilots with no clear owner and no line of sight to value, all quietly consuming compute. Discipline is being selective about what you take on, and just as willing to stop something that is not proving its worth.”

Executing a right-sizing strategy requires asset managers to match model capacity directly to task complexity. This architectural segregation is driven by stark financial realities: a November 2025 Infosys analysis found that running an off-the-shelf Large Language Model (LLM) at scale can cost a bank roughly $3.2 million a year, whereas pivoting to a domain-specific Small Language Model (SLM) drops that operational cost to less than $15,000.

By hosting open-weights architectures inside their own private cloud perimeters, firms can process millions of daily unstructured records without incurring massive third-party API fees. **Andrew Chin, Chief AI Officer at AllianceBernstein**, highlights how this insulates

IT budgets, noting that the firm deploys lightweight architectures like Meta's Llama and Alibaba's Qwen alongside proprietary routing tools. "It comes down to the right tool for the right task," Chin explains. When dealing with high-volume pipelines like overnight news feeds, he avoids premium commercial APIs: "If I want to process that every single night, I don't want to use the best cloud model. I'm going to use an open-source model because that's going to be cheaper for me at scale."

While the buy-side invests heavily in hosting and fine-tuning these smaller models, trading technology providers caution that attempting to build proprietary, domain-specific AI from scratch can become an unsustainable capital expenditure.

"Training our own models isn't where firms like us create value. The value is in how we deliver them," warns Gabbay of Quod Financial. For asset managers and technology vendors alike, attempting to out-engineer foundational AI labs risks being crushed by the sheer velocity of Big Tech's capital investment. "You'll survive until a frontier lab decides to move into your domain, and then you're overrun," Gabbay notes. He argues that capital markets technology is "not about outperforming these global models, it's about building a delivery pipeline that best utilizes them."

In Quod's architecture, that delivery pipeline sits on a Model Context Protocol (MCP) layer that exposes normalized multi-asset data and trading functions to frontier models, allowing the platform to hot-swap between them without vendor lock-in.

Managing compute in 2026 therefore requires a delicate balance across the street: treating raw intelligence as a commoditized public utility while focusing proprietary engineering dollars strictly on the delivery pipelines and data architectures that contain it.

Section 2: The Front Office & Defining "Action Risk"

A core focus of financial AI in 2026 is drawing a hard boundary around "action risk"—the exact point where an automated insight transitions into a client-impacting financial decision. This operational caution is driven by expanding regulatory scrutiny. Specifically, the US Securities and Exchange Commission (SEC) 2026 Examination Priorities enforce strict anti-fraud and fiduciary frameworks to police "AI washing" and algorithmic conflicts of interest.

Despite the enthusiasm surrounding autonomous bots, tier-one institutional quants are explicitly rejecting Large Language Models (LLMs) for direct capital allocation. "We are using a tool trained on information that contains the benefit of hindsight, and then asking it to make predictions about the future," highlights **Judith Gu, Managing Director and U.S. Head of Quantitative Research at UBS Investment Bank**.

Because foundation models are trained on historical snapshots containing post-event knowledge, they poison backtests and risk models with look-ahead bias. "As quant researchers, one of our core principles is to avoid look-ahead bias. Unless carefully controlled, large language models can introduce significant look-ahead bias into research and backtesting," Gu explains, adding that because of this structural flaw, "I would not simply turn on a large language model and say, 'trade for me' or 'put on risk for me.' Today's models are valuable research tools, but they are not yet substitutes for robust

investment and risk controls." As such, current generative models are unsuitable for direct autonomous execution within the investment process.

To bridge the gap between theoretical models and live market execution, trading platforms are forcing generative AI into rigid, rules-based pipelines. Raw foundational models fail in fixed income workflows unless they are stripped of their generative randomness and engineered for speed and precision.

"We needed to prevent any hallucinations and deliver deterministic, 100% accurate results to traders because that's what they expected," notes **Jim Kwiatkowski, CEO of LTX**. To achieve this, Kwiatkowski explains that LTX trained its models to master fixed income vernacular—such as 'par' and 'convexity'—while turning down model temperature and anchoring answers to curated pricing and its proprietary Liquidity Cloud.

When extending this capability to execution, LTX avoids ongoing generative risk by using the LLM only once to interpret trader intent, compiling the workflow into pre-tested software blocks that can stage tickets or launch trades without continuous token exposure.

The architectural division between reasoning and execution serves as the definitive formula for front-office deployment. Chapa at Hitachi Vantara views this as separating the model's predictive capabilities from a firm's hardened transactional core.

"Think of the deterministic application as that domain expertise, and now AI can augment how that application can then deliver more value to the customer," Chapa explains. By maintaining this structural separation, firms ensure erratic model behavior cannot contaminate live trading systems. "We've got traditional deterministic applications all operating against these overlapping data sets, and that's the real power of AI. When we can unshackle it from the traditional agent question and response… and combine that probabilistic reasoning with the deterministic application, IT has been running for decades," he notes.

Coupling the machine's probabilistic speed with the safety of legacy plumbing allows institutions to scale intelligent workflows without absorbing execution risk.

To operationalize these strict execution guardrails across the enterprise, asset managers are establishing progressive taxonomies for agentic authority. Chin at AllianceBernstein enforces a rigid four-tier framework that dictates exactly how much independence a model is granted.

The first tier consists of "insight agents," which surface information but require full human interaction and synthesis to operate. The second tier introduces "interaction agents," which propose a specific response or action but still force a human to explicitly say yes or no before proceeding.

The third tier—"execution agents"—advances to independent action, executing tasks without asking for human feedback, but operating strictly within hard-coded, pre-approved guardrails. However, Chin draws a boundary at the final tier, explicitly barring unsupervised systems from the investment workflow.

“The last layer is truly autonomous AI, and I don’t think that layer is appropriate for our industry,” he states. “That may be okay for driverless cars. I don’t think it’s okay for finance.” By categorically rejecting tier-four autonomy, institutions ensure that human accountability remains legally and operationally intact.

On institutional execution desks, this boundary is enforced with operational finality. At TP ICAP, Shannon establishes that supervisory responsibility remains permanently anchored to the human trader. “In institutional broking, the human stays on the trade,” he asserts, explaining that the firm delineates agentic authority strictly by consequence and reversibility.

While informational, analytical, or reversible tasks—such as surfacing signals, drafting, and summarizing—operate under agentic delegation, live market execution remains strictly bounded. “Where an action carries market or client consequence, commits the firm, moves money or touches an order, that stays deterministic and human-controlled,” Shannon explains. “Autonomy proposes, the qualified human disposes. We do not hand an agent the authority to take market action on its own.”

In regulated wholesale markets, Shannon emphasizes, the underlying control and evaluation framework is as much the product as the model itself.

Rather than granting models unchecked execution authority, leading firms are embedding AI as an active, informed collaborator within existing human workflows. “One of the agents we worked on this year was to help our multi-asset investing team make tactical allocation decisions,” explains Agrawal at T. Rowe Price.

To accelerate the synthesis of proprietary data while preserving the portfolio manager’s voting authority, his team designed a workflow centered entirely on agentic participation. “We thought, can we put an AI agent in this meeting as a committee member? The AI would have access to all our internal research, past decisions, survey data, and macro data, and then it’s chiming in as a committee member along other committee members.”

In the high-stakes environment of quantitative alpha generation, this collaborative philosophy translates into rapid experimentation over unmonitored data mining. Joseph Miller, PhD, Chief Artificial Intelligence Officer at Hudson Bay Capital, notes that relying on AI to blindly search for alpha fundamentally misinterprets the technology’s institutional value.

“Alpha-generation workflows are not data mining or boiling the ocean exercises at Hudson Bay,” Miller explains. “They are intended to more closely resemble rapid experimentation and hypothesis invalidation that anchors on human creativity and ingenuity versus AI-driven exploration.”

By utilizing agents to rapidly falsify human-originated hypotheses, hedge funds multiply their research throughput without exposing their portfolios to the hallucinated correlations of an unsupervised neural network.

Enforcing these strict boundaries across diverse workflows requires a foundation built on both technical architecture and firm-wide governance. From an engineering perspective,

this means programmatically restricting an autonomous agent to the narrowest possible operational permissions.

Citing the industry principle of least agency, Musella at Vanguard notes that “an agent should only possess the minimum authority to do a task.” Instead of outsourcing investment decisions or factual ground truth to a model, he draws a sharp boundary around core decision-making: “We want to retain control over our investment priors. I want to use foundational LLMs for their intelligence, not for their knowledge. There is a distinction there.”

This technical governance perfectly mirrors the buy-side’s cultural stance on fundamental investment judgment. “Our principle is simple: agentic capabilities should speed up the investment workflow while keeping human decision points and accountability intact,” summarizes Krempus at Manulife. As he notes, “We’re a fundamental shop, so AI supports judgment; it doesn’t replace it.” In capital markets, velocity is rewarded, but accountability is mandatory.

Section 3: Middle Office & The Threat of Interconnected Systemic Risk

As artificial intelligence migrates from isolated analytical pilots into interconnected, multi-agent systems, the middle office faces a fundamental operational shift: model risk is rapidly evolving into systemic risk. When thousands of autonomous algorithms interact simultaneously across disparate platforms, they introduce novel systemic vulnerabilities to market stability.

When generative AI functioned primarily as a standalone desktop chatbot, a human reviewer was perpetually stationed at the screen to intercept flawed conclusions. But as capital markets firms link autonomous agents in chained sequences across trade capture, reconciliation, collateral management, and risk reporting, that protective buffer thins dramatically. Since interconnected agents consume each other’s outputs in real time, a single corrupted data record is no longer an isolated defect—it becomes a systemic contagion across downstream workflows.

To prevent erratic buy-side algorithms from destabilizing platform liquidity or triggering cascading market anomalies, execution venues are building rigorous structural controls at the point of trade. Kwiatkowski of LTX points out that platforms must anchor their agents to **rule-based parameters** to prevent rogue model actions and to ensure expected and repeatable behaviors traders can rely upon.

By utilizing a **pre-tested modular architecture**, Kwiatkowski explains that “the large language model is used to understand the intent of the user, and then the agent is created based on a set of building blocks [such as Skills] that are each rigorously tested.” By restricting agentic tools to these verified modules rather than allowing open-ended generative execution, LTX ensures that trading logic remains fully bounded before an order ever reaches dealer liquidity. Consequently, Kwiatkowski notes that “we can be very confident that the agents that users are deploying within the LTX platform will not take unexpected actions and will consistently perform as expected.”

This systemic fragility is exacerbated by the hidden choices generative models make when generating code or processing quantitative inputs. In production environments, standard Large Language Models (LLMs) routinely make unprompted statistical judgments—such as silently dropping missing variables or applying default imputation methods—that distort risk calculations without leaving an obvious error log.

To intercept these invisible hazards, Miller at Hudson Bay Capital explains how multi-strategy hedge funds are re-engineering agentic architectures. To govern these workflows, the firm developed "Hudson," a proprietary internal AI system explicitly designed to catch trace-level decisions before execution. "For example, if Hudson is asked to make a chart of various tickers' average prices, but one of the rows is null, the decision about how to fill or drop that data point would typically be made by the LLM," Miller notes.

Instead of allowing the model to make that assumption autonomously, "Hudson is designed to look for those subtle judgements made in the code and lift them to the human user for sign-off." By programmatically surfacing these micro-assumptions, institutions prevent automated heuristics from silently undermining model integrity.

Beyond silent coding choices, middle-office risk teams must defend against models mistaking statistical noise for genuine market dynamics. While machine learning can process massive volumes of alternative data to identify historical patterns, generative models do not inherently distinguish causation from correlation.

"AI systems can surface spurious correlations faster and at greater scale than humans can," warns Gu from UBS Investment Bank. Without human econometric discipline and peer review to validate the underlying economic thesis, unconstrained models will surface false patterns that collapse under real-world market stress.

For Gu's quantitative research desk, machine-generated discoveries are treated with strict skepticism: "We review the hypotheses and correlations generated by AI, but we do not treat them as investment conclusions. They still require economic intuition, statistical validation, and human review."

Rejecting unverified pattern recognition ensures that generative tools support rigorous research without introducing severe tail risk into portfolio monitoring.

The demand for causal traceability is precisely why the middle office rejects neural network approximations for mission-critical risk and regulatory reporting. While some tech firms market deep learning "surrogate models" as fast, compute-efficient alternatives to Monte Carlo simulations, they introduce a critical audit barrier: neural networks are black boxes that cannot reliably attribute why a specific risk metric moved.

"In risk, or anywhere where there is strong regulatory scrutiny or need for auditability, it is extremely difficult to justify the use of these models," explains Kancharla at Numerix. In an environment governed by strict risk management guidelines, he notes that "reproducibility, auditability, and explainability are all challenges." He concludes that "it is impossible by design to create a proper explanation with a neural network-based model," meaning an approximation that outputs a confident number without mathematical attribution cannot be defended to regulators or internal risk committees.

As institutions confront the reality that single-model governance is insufficient for interacting agent networks, risk management is ascending from individual model validation directly into the multi-agent orchestration fabric.

Chin at AllianceBernstein warns that serial agent workflows introduce compounding vulnerabilities if flawed data cascades downstream. To prevent systemic anomalies, AB deploys parallel architectures rather than relying on a solitary process. “We have multiple agents running the same task to make sure that they all get hopefully a similar answer,” Chin explains, describing them as “ensemble agents” that mirror traditional quantitative modeling techniques.

When outputs diverge or processes reach critical thresholds, independent “validator agents” and automated guardrails flag the anomaly. As Chin explains, “Even if you have multiple agents running in parallel, identify the parts of the process with the greatest risk. Those may be the areas where you interject a human to make sure that things are operating as expected.”

At J.P. Morgan Asset Management (JPMAM), scaling enterprise agentic workflows across proprietary systems like SpectrumIQ requires establishing strict architectural boundaries at the orchestration layer itself.

Kristian West, Managing Director and Head of Investment Platform at J.P. Morgan Asset Management, and **Dillon Edwards, AI Strategist and Executive Director at JPMAM**, jointly detail this governing philosophy: “The governing principle is that autonomy has to be earned: the more consequential the workflow and the more interconnected the agents, the higher the bar for controls, observability, testing, and human oversight before expanding scope,” they explain, enforcing clear rules around what data agents can access and the conditions under which control must revert to a human.

An orchestration framework ensures autonomy is never treated as an open-ended capability, complementing the upfront action boundaries championed across the buy side. Krempus at Manulife frames this operational posture: “Our principle is to define action boundaries up front: what AI can do, what it can access, when it must stop or escalate, and where humans must intervene.” By establishing these hard permissions and automated escalation triggers, firms ensure autonomous assistance never circumvents organizational policy.

Even with rigorous orchestration controls inside the enterprise, the systemic threat compounds significantly when thousands of autonomous algorithms interact across external market venues. If a multi-agent feedback loop malfunctions during a volatile market dislocation, runaway execution requests could overwhelm trading gateways.

To insulate central market plumbing from autonomous buy-side bots, major exchanges are establishing rigid defenses at the ingestion tier. “We have always focused on comprehensive controls, whether or not our clients choose to use AI to model and deploy their trading strategies,” details Sunil Cutinho, Chief Information Officer at CME Group. To provide greater value and security for participants, Cutinho explains that the exchange has “built protections into each API, for example limiting the number of messages that can be submitted to our matching engine.”

By pairing message throttling at the API gateway with centralized defenses—or as Cutinho notes, a suite of 'kill switch' tools that "allow us the flexibility to adjust entry points into our markets if there is ever an issue"—market infrastructure providers ensure that even if buy-side agents cascade out of control, the core matching engine remains completely insulated and predictable.

Section 4: The Real ROI (Operational Leverage & Document Automation)

While front-office trading algorithms command the industry's headlines, the actual balance-sheet return on investment for artificial intelligence in 2026 is overwhelmingly concentrated in post-trade operations and core infrastructure plumbing. An April 2026 report from McKinsey & Company validates this shift, revealing that the most immediate, measurable ROI for agentic AI is a 40% productivity uplift in middle- and back-office data operations, specifically in resolving "last-mile" automation challenges.

In the front office, value remains diffuse and notoriously difficult to separate from broader market betas. "At JPMorgan Asset Management, we're seeing value across the front office and operations, but ROI is usually easiest to measure in the middle and back office," note West and Edwards at J.P. Morgan Asset Management. They explain that this is exactly "where AI can automate repeatable work, increase throughput, and reduce operational risk."

Chin at AllianceBernstein agrees that this is fundamentally a data architecture advantage. "In the back office, many of the tasks are more deterministic. You understand what the workflow looks like, and importantly, you have the data to operate at that level," Chin explains. "That's why it's easy to see the ROI there, because the signal-to-noise ratio is a lot higher, and the data infrastructure is a lot cleaner." By targeting these highly structured environments, institutions can definitively translate theoretical AI capabilities into hard operational leverage.

For institutional asset managers, this operational lift is proven empirically in the automation of unstructured documentation. "For example, our AI Research Assistant has reduced the time investment researchers spend reviewing lengthy documents by 70 to 80 percent," notes Krempus at Manulife. Because this massive time compression is "allowing them to focus on higher-value analysis and decision-making," the firm successfully transforms administrative friction into a direct competitive advantage for its research teams.

Yet unlocking these double-digit efficiency gains requires institutions to confront decades of accumulated architectural debt, and relying on agentic AI to solve it is not the answer. AI models and autonomous agents cannot automate post-trade processing, reconciliation, or document extraction without strict guardrails and strong governance. Galperin at InterSystems emphasizes that "regulated institutions must strictly prohibit autonomous models from accessing raw enterprise source systems directly." She explains that scaling AI safely requires a governed data layer and agentic harnesses that enforce data entitlements, access policies, and audit trails.

Rather than spending millions on bespoke engineering projects for individual pilots, firms need a governed operational data fabric that delivers authoritative context, data entitlements, and policy guardrails to AI systems. By establishing this foundation, organizations can deploy agents safely at scale while achieving measurable business impact across AI initiatives.

When advanced data architecture is successfully paired with rules-based algorithmic optimization, the scale of operational capital efficiency reaches into the billions. In clearing and margin operations, capital markets infrastructure providers are utilizing machine intelligence to mathematically compress counterparty exposure rather than to generate predictive trading signals.

“We leverage capabilities developed for AI to develop models and technology to help our clients optimize their portfolios,” explains Cutinho at CME Group. With their CORE optimizer tool, “firms can model their portfolios to ensure they get the most capital-efficient margin treatment,” which drives staggering balance-sheet value. Cutinho points out that “currently, we are generating over $95B in savings each day across all our cleared markets.”

Freeing up $95 billion in daily capital reliably outpaces the marginal basis-point gains of experimental front-office predictive models, proving that post-trade optimization is one of the highest-yielding applications of advanced mathematics today.

Beyond document processing and capital clearing, another significant driver of quantifiable ROI is the dramatic compression of the traditional Software Development Lifecycle (SDLC).

Generative AI and agentic coding workflows are dismantling multi-month development backlogs. Gabbay of Quod Financial explains how decoupling a customizable AI workspace—such as Quod’s Qanvas front end—from a governed Order and Execution Management System (OEMS) backend transforms this process. “A desk describes what it needs, and our pipeline delivers working software by the end of the week,” Gabbay explains. But because the underlying OEMS remains a deterministic, governed engine, “you cannot breach a limit, trade something that doesn’t exist, or connect to a market that hasn’t been integrated and certified.”

By allowing natural language prompts to generate the presentation chassis while locking down the risk engine, software delivery cycles compress from months into days. This exponential acceleration in code generation introduces a critical new operational bottleneck: verification, security review, and institutional ownership. Gu at UBS Investment Bank cautions that AI coding assistants require continuous expert skepticism.

“AI is helping us improve efficiency and productivity, but it can still make subtle mistakes that require careful review,” Gu warns. Emphasizing that “even with AI coding assistants, developers remain responsible for validating, testing, and understanding the code that is produced,” she notes the machine can reinforce incorrect assumptions or logic if those errors are not identified and challenged by human reviewers. In the era of rapid code generation, line-by-line debugging and human skepticism remain the ultimate firewall against systemic tech risk.

Institutional leaders are realizing that the true financial value of AI lies in expanding operational capacity and research velocity. Agentic infrastructure acts as a force multiplier on human talent instead of an autonomous oracle.

At Vanguard, Musella frames this reality through the lens of research throughput. While AI may not change a quantitative team's baseline hit rate—where roughly one in twenty research ideas makes it into live production—agentic tools effectively double the top of the funnel. As Musella explains, "that batting average might not change, but if I can go from 20 to 40 and double it, just the amount of things that we can do because of the agentic workflows… our investment outcomes to our clients will improve significantly."

By doubling the volume of hypotheses an institution can rigorously test and validate, execution at scale delivers a sustainable competitive advantage while keeping human judgment firmly at the center of the firm.

Section 5: Enterprise Infrastructure, Data Gravity, and Predictable Cloud Migration

Scaling enterprise artificial intelligence across the capital markets is ultimately governed by the physical constraints of network architecture, bandwidth economics, and enterprise storage. While industry attention remains focused on model intelligence and front-office interfaces, production deployments succeed or fail based on where data physically resides and how far it must travel. Institutions attempting to lift and shift decades of accumulated historical records into centralized cloud environments are encountering severe network bottlenecks and cumulative latency.

Chapa at Hitachi Vantara emphasizes that the true cost of scaling AI lies in the physical economics of data gravity, noting that "data does have weight, and where that data sits is where you tend to see compute follow."

When deciding where to position infrastructure, he points out that "moving a model or spinning up compute comparatively can be relatively easy to moving petabytes of enterprise data repeatedly." Evaluating large-scale deployment means recognizing that data movement carries hidden friction: "The cost of AI infrastructure… isn't just the GPU. It's the data movement, which is a hidden cost," Chapa notes, representing a substantial operational expense that often catches business leaders off guard.

Ultimately, latency is cumulative—it encompasses the entire operational timeline: locating distributed records, transporting them across networks, extracting context, and injecting that payload into the model window.

While physical hardware constraints dictate where server infrastructure is deployed, institutions face an equally punishing bottleneck inside their internal software plumbing. Instead of continually duplicating enterprise records into disconnected data stores that rapidly become stale, leading firms are deploying unified data fabrics to query data directly at the source.

Galperin at InterSystems highlights how reducing repeated data replication controls enterprise infrastructure spend. "In many financial services firms, the same data is copied

or moved and reconciled and processed repeatedly across disconnected systems," Galperin explains, pointing out that "every additional copy, every additional pipeline and reconciliation cycle consumes infrastructure capacity, it adds latency and increases energy required to support analytical operational AI workloads."

By establishing a unified data fabric to "create one copy of data that can be used multiple times," she notes that institutions operate far more efficiently while "reducing the load on your overall infrastructure." Operating from an authoritative, connected data estate eliminates the friction of continuous internal data replication, dampening latency while enforcing a disciplined total cost of ownership.

As institutions position compute closer to distributed data estates, they must also ensure their underlying systems do not become captive to a single cloud ecosystem or proprietary model provider. With model capabilities and token pricing shifting rapidly, partnering heavily with hyperscalers introduces the risk of "vendor capture" and operational lock-in.

At J.P. Morgan Asset Management, maintaining technological independence requires establishing strict architectural flexibility. West and Edwards jointly explain how the firm designs its platform infrastructure to safeguard autonomy. "In practice, that means architecting for portability: keeping data and workflow integration patterns as decoupled as possible from any one model or cloud provider," they explain, emphasizing the need for "building the ability to evaluate, compare, and switch between open-source and proprietary models as capabilities, economics, and control requirements evolve."

This architectural discipline extends directly into how execution venues manage the unit economics of agentic deployment. Kwiatkowski at LTX points out that platforms can insulate clients from token expenses by compiling natural language requests into executable workflows upfront.

"We've only used tokens initially to understand the user's command and to build the agent," Kwiatkowski explains. Because the generative lifting is completed before execution begins, "the agent will now run in what's called a deterministic fashion," allowing the platform to monitor live market conditions using standard compute rather than continuous, high-latency LLM API calls.

By structurally separating natural language configuration from live execution logic, institutional platforms ensure that automation scales without exposing trading desks to compounding inference bills or runtime hallucinations.

While buy-side managers prioritize data locality, central market infrastructures face an even more demanding engineering challenge: achieving sub-millisecond execution consistency within public cloud environments. For years, global exchanges resisted migrating core matching and clearing engines to the cloud due to "noisy neighbor" latency jitter and the non-deterministic nature of virtualized hypervisors. To overcome this barrier, CME Group collaborated with Google Cloud to co-engineer specialized bare-metal instances and hardware-level network distribution.

"In capital markets, even small variations in performance and latency can have a significant impact on trade execution," Cutinho at CME Group explains. By deploying

Google Cloud's specialized U4 machine family, CME engineered its cloud stack to "process latency-critical execution tiers in a highly predictable, consistent amount of time, every single time," supported by scalable "hardware-based multicast data distribution for reliable market-data feeds."

This architecture proves that mission-critical exchange matching can operate within modernized environments without sacrificing performance. As Cutinho concludes, "By combining deterministic, ultra-low latency networking and dedicated bare metal compute with the benefits of the cloud, our market participants can ingest market data and execute with the speed of traditional [co]location and the agility of the cloud." Compute and network fabrics must be purpose-built for execution consistency if they are to survive in live production.

Conclusion: The 2026 Mandate

The transition from isolated generative chat interfaces to interconnected, multi-agent workflows definitively marks the conclusion of the capital markets' experimental phase. Moving autonomous tools into live production environments requires institutions to confront the financial, operational, and regulatory realities of AI at scale. Global regulatory bodies—building on frameworks like the SEC's 2026 Examination Priorities, FINRA's 2026 Annual Regulatory Oversight Report, and the International Organization of Securities Commissions' (IOSCO) May 2026 Supervisory Toolkit for AI Use in Capital Markets—have made clear that operational resilience, systemic stability, and governance must keep pace with algorithmic deployment.

Executing at scale is no longer an abstract exercise in model evaluation; it is a rigorous, enterprise-wide test of an institution's data architecture, compute management, and risk controls. The mandate is unequivocal: enterprise plumbing must be capable of safely absorbing the machine's velocity.

The next cohort of institutional leaders will not be defined merely by access to frontier models or experimental trading pilots. Instead, competitive advantage belongs to firms that enforce architectural and budgetary discipline. This requires a measured focus on unit economics—evaluating the "cost per idea" before infrastructure resources are committed—and strategically routing high-volume workflows to domain-specific Small Language Models (SLMs) to manage cloud expenditure. It also demands modernizing legacy ETL pipelines in favor of unified data fabrics that connect directly to source systems, preventing enterprise AI initiatives from devolving into bespoke engineering projects.

Krempus at Manulife establishes this budgetary discipline as a core operational requirement. Warning that technological acceleration risks outpacing measurable enterprise value without strict qualification, he notes that "if a solution can't demonstrate sustained business value relative to its cost and risk, it shouldn't move from experimentation into scaled production." Without this rigorous operational discipline, experimental acceleration easily devolves into wasted capital.

Crucially, as agentic systems gain the capability to trigger actions across workflows, capital markets consistently decline unsupervised autonomy across front- and middle-office operations. From the multi-agent orchestration frameworks deployed at J.P. Morgan

Asset Management to the API message caps and kill switches safeguarding the CME Group matching engine, market participants and infrastructure operators have established clear boundaries around execution. Autonomy in institutional finance is not an open-ended capability; it is a bounded operational permission that must be earned through observability, auditability, and controls. While agentic architectures accelerate research synthesis, support tactical allocation deliberations, and expedite quantitative hypothesis invalidation, regulatory and fiduciary accountability remains explicitly tethered to human decision-makers.

As computational capability becomes commoditized, institutional domain expertise has emerged as the most critical asset across the organization. When software delivery cycles compress from months into hours and multi-agent systems process vast streams of market data in seconds, the operational bottleneck shifts immediately toward verification, causal validation, and technical risk defense.

Drawing an analogy to mathematical proofs, Kancharla at Numerix argues that software velocity without human defense creates enterprise fragility. “At the end of the day, humans have to be able to own it, understand it, and defend it,” he emphasizes. “In the same way, code has to be owned and defended by someone so they can take long-term responsibility for it.” In the era of execution at scale, artificial intelligence provides operational velocity, but human accountability remains the indispensable steering wheel.

• • • •

Sean Creamer *is a writer focusing on AI, financial, and marketing technology. He’s worked as a journalist for HFMTechnology, Wall Street Letter, and Operational Risk and Regulation. As a marketing tech writer, he penned stories for EMARKETER. In addition to writing about real estate, Sean is an outdoor enthusiast who manages publicity for the Kayak and Canoe Club of New York, a whitewater enthusiast club.*

Image via Pexels.

Hitachi Vantara
www.hitachivantara.com

Hitachi Vantara

Hitachi Vantara is transforming the way data fuels innovation. A wholly owned subsidiary of Hitachi Ltd., Hitachi Vantara provides the data foundation the world’s leading innovators rely on. Through data storage, infrastructure systems, cloud management, and digital expertise, the company helps customers build the foundation for sustainable business growth.



David Chapa
Chief AI Strategist
Hitachi Vantara

InterSystems

www.InterSystems.com/Financial

Established in 1978, InterSystems is the leading provider of next-generation solutions for enterprise digital transformations in the finance, healthcare, manufacturing, and supply chain sectors, including production applications at most of the top global banks. Its cloud-first data platforms solve interoperability, speed, and scalability problems for large organizations around the globe. InterSystems is committed to excellence through its award-winning, 24×7 support for customers and partners in more than 80 countries. Privately held and headquartered in Boston, Massachusetts, InterSystems has 36 offices in 25 countries worldwide.

For more information, please visit **InterSystems.com/Financial**



Irene Galperin
Senior Advisor, Financial Services
InterSystems

LTX Trading

www.ltxtrading.com

LTX is the fixed income analytics and e-trading platform that enables corporate bond market participants to trade smarter. We've combined powerful, patented artificial intelligence with innovative trading protocols to improve liquidity, efficiency, and execution.



Jim Kwiatkowski
Chief Executive Officer
LTX, a Broadridge Company

Numerix

www.numerix.com

Since its founding in 1996, Numerix has been at the vanguard of financial technology, providing groundbreaking expertise, quantitative analytics, and software that redefines pricing and risk management in the financial markets. With the strategic acquisitions of FINCAD, Kynex, and PolyPaths, Numerix has further strengthened its leadership position, empowering financial institutions worldwide to transform risk into opportunities with confidence.

Numerix is headquartered in New York City, with offices around the world in major financial centers that allow immediate access to the latest trends, innovations, and thought leaders in the industry. This extensive network across the Americas, EMEA, and Asia-Pacific enables Numerix to offer localized support and services to its diverse clientele while drawing on global insights and expertise.



Satyam Kancharla
EVP, Chief Product Officer
Numerix

Quod Financial

www.quodfinancial.com

Quod Financial is a global provider of multi-asset trading technology, delivering high-performance OMS, EMS, Smart Order Routing, and Algorithmic Trading solutions. Built on a modular, AI-driven architecture, Quod empowers financial institutions to automate, customize, and scale their trading infrastructure seamlessly across the full order lifecycle. With advanced liquidity engines, low-code UI flexibility, and normalized cross-asset workflows, Quod Financial delivers complete trading control without system disruption. Headquartered in London, Quod Financial operates globally with offices in Paris, New York, Dubai, Hong Kong, Singapore, and Toronto.

For more information, visit **www.quodfinancial.com**



Medan Gabbay
Co-Chief Executive Officer
Quod Financial

TabbFORUM is an open community that provides a platform for capital markets professionals to share their ideas and thought leadership with their peers. The views and opinions expressed are solely those of the author(s). They do not necessarily reflect the opinions of TABB Group, its analysts, TabbFORUM and its editors, or their employees, affiliates and partners.

©2009-2026 by the TABB Group, LLC, and contributors