Modern Finance

Modern Finance Institute

Article

Measuring agentic AI adoption and control frameworks in finance

Atta Ul Mustafa[1,*] **and Ahmet Faruk Aysan**[2]

1 Hamad Bin Khalifa University, atul88769@hbku.edu.qa
2 Hamad Bin Khalifa University, aaysan@hbku.edu.qa
* Correspondence: Hamad Bin Khalifa University, atul88769@hbku.edu.qa

Abstract: Agentic artificial intelligence (AI) systems can execute actions rather than merely generate content, raising distinct governance and operational risk questions for financial institutions. This study measures how agentic AI is entering U.S. finance firms' annual filings by treating disclosures as text-as-data. We assemble a balanced panel of 2,500 firm–year observations (500 firms per year) from 2021–2025 and implement an auditable dictionary-and-context approach that flags agentic references and then quantifies the surrounding "controls density" (governance and safety language) within the same local disclosure window. Agentic disclosures are absent in 2021–2023, appear in 2024 (0.4% of firm-years), and increase in 2025 (1.6% of firm-years), indicating a late but accelerating diffusion phase. Within the set of agentic-mention filings, autonomy evidence remains rare. However, it focuses on regions with higher control density, consistent with governance maturity serving as a prerequisite for action-taking deployments. The analysis provides a transparent measurement framework and baseline statistics for tracking the emerging shift from AI discussion to action-oriented, agentic deployments in finance.

Keywords: agentic AI; financial institutions; corporate disclosure; governance controls; text-as-data

Jel Classification: C45, G21, M41, O33

1. Introduction

For most of the last decade, AI in finance has largely meant prediction and decision support, such as credit scoring, fraud detection, surveillance, forecasting, and document processing that ultimately feeds into a human decision. What is changing now is not simply model accuracy or the availability of larger language models, but the organizational locus of agency, whether AI systems are allowed to act on the world, not merely advise about it. The emergence of agentic AI in financial institutions, defined as AI-enabled systems that (i) operate as goal-directed agents, (ii) integrate with enterprise tools and workflows, and (iii) exercise some degree of delegated action rights (e.g., initiating processes, completing operational steps, triggering downstream tasks) under specified constraints (Wooldridge & Jennings, 1995; Wang et al., 2024). Importantly, agentic here is not a synonym for generative. Generative models may be a component of an agentic system, but agentic deployment is about execution, the ability to plan, call tools, and move work forward inside a production environment, rather than text generation per se (Wang et al., 2024). In regulated finance, that distinction is decisive as moving from analytics to action changes the risk surface, the control architecture, and the accountability expectations facing banks, insurers, exchanges, and market intermediaries.

The central challenge is that the most consequential part of this transition, such as the real-world deployment of agentic systems inside financial organizations, is difficult to observe systematically. Most existing evidence comes from vendor case studies, selective announcements, or proprietary internal data that is not publicly available. As a

Citation: Ul Mustafa, A., & Aysan, A. F. (2026). Measuring agentic AI adoption and control frameworks in finance. *Modern Finance*, *4*(1), 81-94.

Accepting Editor: Adam Zaremba

Received: 18 February 2026
Accepted: 12 March 2026
Published: 17 March 2026

result, the academic and policy conversation often conflates (a) generic AI enthusiasm and experimentation with (b) genuine operational adoption of agentic capabilities. This is not a trivial measurement problem: the credibility of any claim about "the future being agentic AI" depends on separating symbolic mention from revealed organizational commitment to autonomy and control.

This paper addresses that gap by treating financial institutions' public disclosures as a measurable indicator of adoption, while explicitly recognizing that disclosure language is imperfect and strategic. Empirically, textual disclosures are widely used in economics and finance because they provide repeated, comparable, time-stamped signals about organizational priorities and constraints at scale (Gentzkow, Kelly, & Taddy, 2019). However, standard textual approaches are not designed to isolate the agentic margin: they can detect "AI talk," but struggle to distinguish systems that generate content from those embedded in workflows with action rights and governance gates. Moreover, even when agentic language is detected, a second question remains unanswered. What determines whether organizations move from pilots to autonomy? In financial institutions, a plausible answer is governance capacity. This view is consistent with a growing literature on documentation, auditing, and accountability artifacts, and model reporting, audit trails, dataset documentation, and end-to-end internal control frameworks emphasizing that the bottleneck in high-stakes settings is not only model capability but organizational control maturity (Mitchell et al., 2019; Raji et al., 2020; Gebru et al., 2021). The implication is that agentic diffusion may be constrained less by the availability of models and more by firms' ability to operationalize oversight, including permissions, approvals, monitoring, evidence logging, and fail-safe mechanisms.

Against this backdrop, the paper makes two related contributions. One empirical and one conceptual. Empirically, it provides a transparent, replicable measurement strategy to track agentic AI diffusion in finance from public filings over time, focusing on actionable autonomy rather than generic AI references. The key output is a set of time-series adoption indicators that are conservative: we treat agentic events as rare and require context that is consistent with operational deployment. That conservatism is essential, because over-counting would mechanically generate an agentic boom narrative. Conceptually, the paper advances a governance-bottleneck hypothesis: within the subset of disclosures that credibly reflect agentic implementation contexts, autonomy-positive language should co-occur with stronger, denser governance language, because action requires credible constraints. This channel matters because it reframes the "future = agentic AI" claim in a defensible way: not as an assertion about the current level of autonomy in finance, but as an argument about the trajectory and its constraint that agentic deployment diffuses as organizations build controls.

The evidence produced by this approach is consistent with a late but accelerating diffusion pattern. According to results, explicit autonomy-positive agentic disclosures remain uncommon through the early period and only begin to appear in the later years of the sample. In other words, the data does not support a narrative that agentic AI is already pervasive across financial institutions. Instead, the pattern is better described as emergence; claims of agentic execution are rare, but they rise precisely during the period when firms' disclosures increasingly differentiate between experimentation and governed deployment. This matters for two reasons. First, it suggests that agentic AI is not merely a relabeling of earlier AI adoption but a distinct mode of adoption with separate organizational requirements. Second, it implies that forward-looking claims should focus on the institutional conditions that enable scaling, control frameworks, governance tooling, and auditability, rather than only on model capabilities.

A second set of results speaks directly about why the shift matters. When we restrict attention to anchor contexts and disclosure passages that plausibly correspond to real deployment settings and then examine the relationship between autonomy-positive statements and a measured controls density index (governance language intensity within those same anchored contexts), the direction of association supports the governance-

bottleneck view: autonomy-positive cases are more likely to appear where control language is thicker. This pattern should not be oversold as causal; disclosure language is strategic, and the number of autonomy-positive anchor cases is small by construction. The value of the channel test is different as it shows that the agentic trajectory in finance is not simply "more AI," but "more AI under constraints." That is precisely what makes the transition economically and regulatorily important. If agentic AI were diffusing without governance, one would expect autonomy language to appear in low-control contexts; instead, the association aligns with the notion that controls are a prerequisite for credible autonomy. Put differently, the "future = agentic" claim becomes defensible as a statement about governed autonomy replacing ad hoc experimentation, not about indiscriminate automation.

These findings contribute to the literature in three ways. First, they provide a measurement framework for agentic adoption that is aligned with core agent theory (Wooldridge & Jennings, 1995) while reflecting the modern architecture of LLM-enabled autonomous agents (Wang et al., 2024). Second, they advance a practical empirical distinction between "AI mention" and "agentic deployment," responding to growing concerns that discourse on AI risks and benefits can be distorted when language and implementation are treated as equivalent. Third, they offer a governance-centered interpretation of diffusion: in regulated finance, the limiting factor is plausibly not only technical capability but the ability to produce accountability artifacts such as documentation, monitoring, evidence logging, and auditable controls that allow autonomy to be scaled without violating safety, compliance, or reputational constraints (Mitchell et al., 2019; Raji et al., 2020; Gebru et al., 2021).

The remainder of the paper proceeds as follows. Section 2 describes what has already been done on agentic AI in finance. Section 3 describes the data, corpus construction, and the operationalization of agentic AI measures, including the anchored extraction logic and the construction of the control density index. Section 4 presents the main diffusion evidence and the baseline association between agentic adoption indicators and time. Section 5 evaluates the governance channel within anchor contexts and reports robustness checks designed to address rare events and measure sensitivity. Section 6 concludes by interpreting the results as a trajectory toward governed autonomy, discussing limitations inherent to disclosure-based measurement, and outlining implications for research and policy on safe agentic deployment in financial institutions.

2. Literature Review

The rapid digitization of financial intermediation has long made the sector a natural testbed for data-intensive automation. A large empirical literature examines how machine learning reshapes screening, underwriting, pricing, and monitoring, typically emphasizing predictive accuracy, allocation effects, and distributional consequences in credit markets (Fuster et al., 2022). A related strand highlights how technology-enabled lenders can alter access and pricing, raising renewed concerns about disparate impact and opacity even when models appear operationally effective (Bartlett et al., 2022). What is striking across these contributions is that "AI in finance" is often studied as a set of improved prediction tools embedded into existing workflows. The frontier question for the next wave is different. When systems are designed not merely to predict or recommend, but to plan, coordinate, and execute multi-step actions inside operational environments, the central constraint shifts from prediction quality to governance capacity.

This shift is captured by the contemporary move toward agentic AI. Agentic systems are characterized by their ability to decompose tasks, select tools, interact with external systems, and carry out actions with limited human intervention, often through iterative planning and monitoring loops (Wang et al., 2024). Conceptually, the phenomenon aligns with a broader perspective in which autonomous computational entities can be treated as decision-making "actors" whose behavior must be anticipated and constrained, much like other complex socio-technical agents (Rahwan et al., 2019). This matters acutely in finance

because the operational surface area is huge: customer interaction, transaction processing, compliance screening, fraud operations, incident response, and internal controls are all environments where action-taking software can generate real-world effects at scale. In such settings, the key empirical challenge is no longer whether algorithms can produce accurate signals, but whether organizations are sufficiently prepared to grant those systems execution rights.

A substantial governance and accountability literature explains why autonomy creates qualitatively new risks. First, increased capability can amplify harm through scale and speed, while also increasing the difficulty of tracing responsibility across people, models, data pipelines, and downstream systems (Bender et al., 2021). Second, the practical challenge is not only about ethical intent but also about implementable documentation, auditing, and oversight mechanisms that are compatible with real organizational constraints. The literature has therefore proposed operational artifacts designed to make AI systems legible and reviewable. Model card's structure standardized reporting of intended use, evaluation, and limitations for deployed models (Mitchell et al., 2019). Datasheets for datasets articulate provenance, collection decisions, and known risks in training data, making failures and biases more diagnosable (Gebru et al., 2021). Complementing these documentation frameworks is an auditing tradition that treats accountability as an end-to-end process requiring measurable governance checkpoints and evidence trails, rather than a one-time compliance statement (Raji et al., 2020). Together, these strands suggest a common implication for agentic AI, that autonomy becomes credible only when it is paired with demonstrable control infrastructure capable of bounding behavior and enabling post-hoc verification.

Human oversight is not merely a normative preference in this literature, but it is an engineering and organizational design problem. Work on human–AI interaction provides concrete guidance for designing systems that maintain appropriate human control, set expectations, and support error recovery, features that become more important as systems take actions rather than offer suggestions (Amershi et al., 2019). The fairness and bias literature reinforces this point by showing that harmful outcomes can arise even in high-performing systems, particularly when data reflect structural inequities or when deployment contexts differ from training conditions (Mehrabi et al., 2021). In financial institutions, these insights imply that the transition from AI-as-assistant to AI-as-operator is not primarily a question of more advanced models, but of whether organizations can operationalize monitoring, approvals, traceability, escalation, and accountability in a way that can keep pace with autonomous action.

Despite the relevance of these governance frameworks, empirical finance has limited evidence on agentic deployment because the most informative operational data are proprietary. Internal incident logs, workflow-level permissions, audit trails, and performance dashboards are rarely observable to researchers, creating a gap between fast-moving practice and what can be credibly measured in public data. This motivates a complementary approach grounded in disclosure analysis. The text-as-data tradition in economics demonstrates that corporate communications can be systematically measured to infer organizational priorities, risk perceptions, and strategic posture, often with predictive and explanatory value (Gentzkow et al., 2019). In finance, disclosure-based measures have been used to map firm-level exposures to political and regulatory risk using large-scale text, illustrating how language can serve as a measurable signal of attention and salience (Hassan et al., 2019). These methods are especially useful when the underlying operational reality is hard to observe directly, but where organizations have incentives to communicate strategically about technology, governance, and risk management to investors, regulators, and counterparties.

However, disclosure analysis also has limitations that are particularly salient for agentic AI. Corporate language can reflect aspiration, marketing, or experimentation rather than genuine production deployment. Classical dictionary approaches are attractive because they are interpretable and auditable, qualities valued in accounting and

finance research that relies on textual classification (Loughran & McDonald, 2011); however, they are also vulnerable to false positives when terms are used in broad or generic ways. These concerns suggest that credible measurement of agentic AI adoption must go beyond counting general AI mentions. The literature suggests that the more defensible signal is language that couples autonomy with institutional controls, because autonomy without controls is not only risky but also difficult to legitimize within regulated intermediaries. In other words, governance language is not an accessory to the phenomenon, but it is part of what makes autonomy plausible as a genuine organizational capability rather than a rhetorical flourish.

This leads to a synthesis across the three literatures: (i) empirical finance documents significant impacts from algorithmic decision tools but often treats them as prediction upgrades embedded in stable workflows (Fuster et al., 2022); (ii) computer science and socio-technical governance research shows that autonomy changes the risk profile and therefore the required accountability apparatus (Mitchell et al., 2019; Raji et al., 2020; Gebru et al., 2021; Bender et al., 2021); and (iii) text-as-data methods provide a scalable path for observing how firms communicate about emerging capabilities when operational data are not public (Gentzkow et al., 2019; Hassan et al., 2019). The resulting research gap is not simply whether finance firms talk about AI, but whether the public record contains systematic evidence consistent with a transition toward agentic deployment, and whether that transition is accompanied by a rise in control-oriented language that would be expected if governance is the binding constraint. Addressing this gap clarifies why the shift to agentic AI matters. It speaks to operational readiness and institutional constraints, not merely technological enthusiasm. It also provides a more defensible foundation for interpreting diffusion patterns as capability formation rather than as hype cycles, because credible autonomy in finance must be explainable, monitorable, and auditable under regulatory scrutiny.

3. Methodology

This study measures the emergence and diffusion of agentic AI in financial institutions using a reproducible, text-as-data design applied to annual SEC Form 10-K filings. We conceptualize agentic AI as organizational deployment of AI systems that (i) can execute tasks or decisions beyond text generation (autonomy/action rights) and (ii) require governance safeguards (controls) such as approvals, audit trails, monitoring, least-privilege access, and escalation/kill mechanisms. This framing aligns with the growing engineering and governance view that the binding constraint for agentic deployment is not model capability alone. However, the ability to operationalize reliable tool-use with appropriate control structures (Qu et al., 2024), and that credible AI deployment claims are increasingly evaluated in audit-ready disclosure terms (Mitchell et al., 2019; Raji et al., 2020; Gebru et al., 2021).

3.1. Data construction and preprocessing

We construct a strictly balanced panel of 500 publicly listed financial institutions, observed annually from 2021 to 2025, yielding 2,500 firm-year observations. Firms are harmonized across years using the SEC registrant identifier CIK. For each firm and fiscal year, we retain one annual report by selecting the primary Form 10 K associated with that fiscal year, and we exclude amendments unless an amended filing is the only available annual report for that firm year. Eligibility requires that the same firm files an annual report every year of the sample window, so firms that merge, exit public markets, or otherwise stop filing during the period are not included. If a corporate action produces a new registrant CIK, we treat the post-event registrant as a distinct entity and do not combine it with the predecessor. Industry membership is determined using the SEC-reported SIC code, and we apply the finance SIC screen consistently across the full window, excluding firms whose SIC classification moves outside the finance range during

the sample period. Each filing is parsed into plain text by removing HTML, embedded tables, and non-substantive boilerplate to reduce spurious keyword matches. Because agentic disclosures are sparse and often embedded in operational narrative (e.g., risk management, technology strategy, controls, and compliance sections), we retain full narrative text rather than restricting to a single item (e.g., Risk Factors), consistent with modern disclosure research that treats language as a measurable economic object and emphasizes careful preprocessing to avoid mechanical artifacts (Gentzkow et al., 2019; Loughran & McDonald, 2011).

3.2. Dictionary-based measurement of agentic AI constructs

We operationalize agentic AI using three keyword dictionaries (see Table 1) that map directly to the conceptual components of agentic deployment. First, an anchor dictionary identifies filing passages plausibly discussing AI systems relevant to agentic workflows. Second, an autonomy dictionary captures language indicating execution rights, orchestration, agent-based automation, or system-triggered actions (i.e., "doing" rather than "generating"). Third, a controls dictionary captures governance and safeguards (e.g., approvals, audit logs, access controls, monitoring, model risk governance). The dictionaries are applied with case-insensitive matching and conservative boundary rules (word/phrase boundaries and common plural/tense variants) to reduce false positives. The measurement approach follows best-practice guidance in finance and economics to use transparent, domain-adapted vocabularies rather than generic sentiment lexicons (Loughran & McDonald, 2011; Gentzkow et al., 2019).

Table 1. Keyword dictionaries used to operationalize agentic AI constructs in Form 10-K disclosures.

Construct	Role in measurement	Terms (exact)
Anchor (agentic-system vocabulary)	Locates candidate passages where agentic/AI-system discussion is likely to occur; used for gating and within-anchor analyses	agentic; autonomous agent; AI agent; multi-agent; agentic workflow; agent orchestration; tool use; tool-using; function calling; action execution; workflow automation; autonomous workflow; copilots; copilot; LLM; large language model; foundation model; generative AI; genAI; ChatGPT; GPT; retrieval augmented generation; RAG
Autonomy (execution rights)	Primary indicator of agentic adoption evidence (action/orchestration/execution)	execute; execution; trigger; triggered; automate; automation; orchestrate; orchestration; agent; agents; multi-agent; autonomous; autonomy; tool call; tool-calling; function call; function-calling; action; actions; decisioning; straight-through processing; STP; closed-loop; self-serve resolution; auto-resolve
Controls (governance safeguards)	Measures governance emphasis around agentic deployment; used to build "controls density" and channel tests	approval; approvals; human-in-the-loop; HITL; sign-off; oversight; audit trail; audit log; logging; monitoring; guardrail; guardrails; kill switch; rollback; escalation; exception handling; segregation of duties; SoD; least privilege; access control; permissions; authentication; authorization; model risk management; MRM; validation; governance; compliance; controls; risk controls

Notes. This table lists the exact anchor, autonomy, and controls vocabularies used in the dictionary-and-context measurement. Terms are matched case-insensitively with conservative boundary rules; multiword phrases are matched as phrases. The anchor dictionary is used to locate candidate AI/agentic passages and define "gated text" windows (±220 characters around each anchor hit; overlapping windows are merged). Autonomy and control hits are counted only within the gated text. Abbreviations: LLM = large language model; RAG = retrieval-augmented generation; STP = straight-through processing; HITL = human-in-the-loop; SoD = segregation of duties; MRM = model risk management.

3.3. Gated text windows and controls density channel variable

A central identification challenge is that control language is ubiquitous in financial filings (e.g., generic compliance wording), while autonomy language can appear in unrelated operational contexts. To make the controls–autonomy relationship defensible, we compute control measures only within gated windows around anchor hits. Concretely, each anchor match defines a symmetric character window (±220 characters) in the file. Overlapping windows are merged, producing a gated text subset intended to approximate the local context in which the firm is discussing relevant AI agentic systems. We then count autonomy and controls keyword hits only inside gated text, producing binary indicators (e.g., AutonomyAny = 1 if any autonomy hit occurs in gated text) and continuous intensity measures.

The key channel variable is control density, defined as:

$$\text{ControlsDensity}_{it} = 1000 \times \frac{\#\text{ControlsHits in gated text}_{it}}{\text{GatedCharacters}_{it}}$$

This normalizes governance language by the amount of locally relevant text, mitigating mechanical correlations driven by filing length or generic risk/compliance boilerplate. The logic is directly tied to the governance-bottleneck hypothesis, that if autonomy evidence reflects genuine movement toward agentic execution, it should appear disproportionately where governance language is dense within the same local AI-system context, rather than merely co-varying with overall compliance verbosity.

3.4. Diffusion estimation and inference

To characterize diffusion over time, we estimate panel logit specifications where the dependent variable is the binary agentic indicator derived from gated autonomy evidence at the firm-year level. Year fixed effects capture time variation in the prevalence of agentic disclosure, and predicted probabilities are computed for each year from the estimated model. Because the outcome is sparse, we use regularized/bias-reducing logistic estimation as needed to avoid separation and inflated coefficients in rare-event settings, following established bias-reduction principles for logistic models (Firth, 1993). Uncertainty is quantified using firm-clustered resampling (cluster bootstrap by firm across years) to respect within-firm dependence in disclosure style and technology strategy language (Gentzkow et al., 2019).

3.5. Testing governance as a bottleneck within anchor cases

To directly test whether governance intensity is associated with autonomy evidence conditional on being in an AI-system discussion, we restrict attention to anchor-positive firm-years and estimate within-anchor models of the form:

$$\Pr\left(\text{AutonomyAny}_{it} = 1 \mid \text{Anchor}_{it} = 1\right) = \text{logit}^{-1}\left(\alpha + \beta \cdot \text{ControlsDensity}_{it} + \sum_{y} \delta_y \mathbb{1}\{t = y\}\right)$$

We complement this parametric approach with a nonparametric association test by discretizing controls density into quantile bins and comparing autonomy incidence across low- vs high-density strata (Fisher-type exact association tests), then reporting uncertainty using cluster bootstrap procedures aligned with the firm-year structure. This combination is designed to be defensible under small-sample and rare-event conditions: the exact test does not rely on asymptotic normality, while the regularized logit provides an interpretable marginal association conditional on time effects and within-panel dependence (Firth, 1993). The overall approach follows the transparency principle emphasized in modern text-as-data and AI accountability work: measurement choices are

explicit, interpretable, and inspectable rather than purely latent (Gentzkow et al., 2019; Mitchell et al., 2019; Gebru et al., 2021).

3.6. Results

Table 2 reports on the time profile of agentic AI disclosures in the finance sample. The pattern is stark: agentic adoption is not a smooth trend but a late-period emergence. Across 2021–2023, the measured share of firm-year reports containing agentic AI deployment language is exactly zero. The first observable non-zero prevalence appears in 2024 (0.4%, 2/500) and then rises further in 2025 (1.6%, 8/500). While the levels remain low in absolute terms, consistent with the idea that operational agentic deployment is still early-stage, the break from a multi-year zero baseline to a sustained non-zero prevalence is economically meaningful for a disclosure-based measure. In text-as-data settings, the appearance of a new operational vocabulary in regulated, carefully drafted reports typically reflects an underlying transition from experimentation to institutionally sanctioned implementation rather than random noise (Gentzkow, Kelly, & Taddy, 2019; Loughran & McDonald, 2011).

Table 2. Agentic AI diffusion over time in financial institutions' 10-K filings (2021–2025).

Year	Firm-years (N)	Agentic mentions (n)	Share	95% CI (low)	95% CI (high)
2021	500	0	0.000%	0.000%	0.735%
2022	500	0	0.000%	0.000%	0.735%
2023	500	0	0.000%	0.000%	0.735%
2024	500	2	0.400%	0.110%	1.453%
2025	500	8	1.600%	0.819%	3.108%

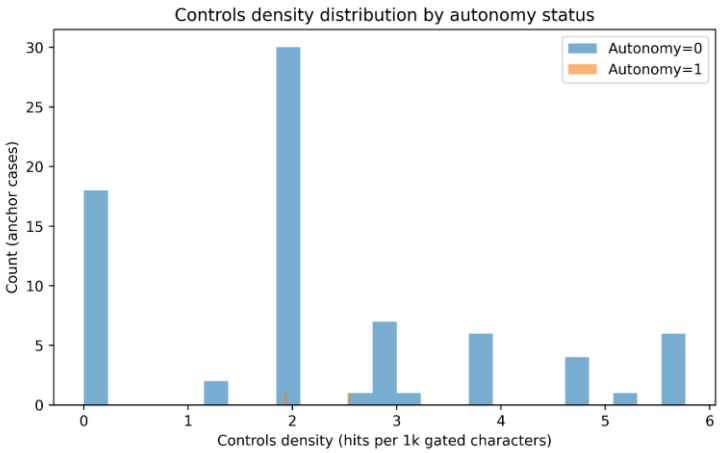
Notes. The unit of observation is a firm-year. The sample is a balanced panel of 2,500 firm-years (500 per year). "Agentic mentions" count firm-years classified as agentic under the paper's text-as-data procedure (agentic/autonomy evidence identified within gated anchor contexts). "Share" is the fraction of firm-years in each year with agentic classification (n/N), reported in percent. 95% confidence intervals are exact binomial (Clopper–Pearson) intervals for the yearly share.

A useful way to formalize the shift (without forcing unstable parametric structure onto rare events) is to treat 2021–2023 as a pre-emergence regime and 2024–2025 as a post-emergence regime. Under that comparison, the data reject the null that the post period is no different from the pre period: with 10 agentic positives in 2024–2025 versus 0 in 2021–2023, an exact test strongly supports a regime change (Fisher's exact $p \approx = 0.0001$, one-sided). This matters because it converts narrative from what we see some mentions lately into the agentic AI vocabulary, becoming present only after a specific point in time, consistent with diffusion rather than idiosyncratic drafting (King & Zeng, 2001).

However, diffusion in agentic AI should not be interpreted as a simple increase in hype. The channel results show that what differentiates the rarer "autonomy" disclosures (systems described as acting, executing, or completing steps with reduced human handling) is not merely the presence of agentic framing, but the co-presence of governance and control language. This is the central reason the shift matters: in finance, autonomy is not adopted in isolation; it is adopted when it is adopted alongside explicit controls. That is exactly the "governance-bottleneck" story.

To test this channel, we restrict attention to anchor cases: firm-years where there is at least some baseline agentic AI context (i.e., where it is meaningful to ask whether the institution is also describing controls and/or autonomy within that context). Within this anchor set ($N = 78$), autonomy disclosures are extremely rare ($k = 2$), which is itself an important substantive result: even among institutions that talk in agentic terms, only a small minority go as far as describing autonomous execution-like capabilities. This rarity is consistent with the idea that finance treats autonomy as a high-friction organizational step, requiring governance readiness (Raji et al., 2020; Mitchell et al., 2019).

Figure 1. Autonomy incidence by controls-density quantile (within anchor cases).



Notes. This figure plots the autonomy rate within anchor-positive firm-years across quartiles of control density. Controls density is defined as the number of controls/governance keyword hits per 1,000 characters of gated text (gated text defined around anchor hits; see Methodology). Points show the share of anchor cases with Autonomy = 1 in each bin; vertical bars show 95% exact binomial confidence intervals for the bin-level rate. Bin counts correspond to the companion table reporting the same quantile breakdown.

Figure 1 and Table 3 then connect autonomy to control density, measured as the intensity of control/governance language within the gated anchor text (hits per 1,000 gated characters). When anchor cases are sorted into control density quantiles, the autonomy rate rises from 0% in the lowest bin to 3.45% in bin 2 and peaks at 11.11% in bin 3, before returning to 0% in bin 4. The non-monotonicity at the top bin is interpreted as small-sample instability: with only two autonomy-positive cases in total, any binning necessarily produces wide confidence intervals and sensitivity to where those two cases fall. What is robust, and visually evident in Figure 1, is that autonomy does not appear in the "controls-sparse" region; it appears only in the mid-to-higher controls density region.

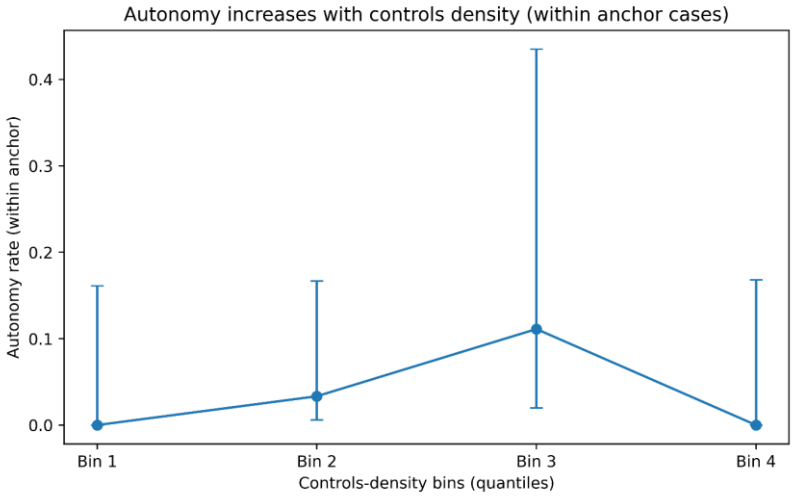
Table 3. Autonomy rate by controls density quantiles (anchor-positive firm-years).

Controls density bin (quantiles)	Anchor cases (n)	Autonomy present (k)	Autonomy rate (k/n)	95% CI (low)	95% CI (high)
1	25	0	0.000%	0.000%	13.682%
2	29	1	3.448%	0.611%	16.759%
3	9	1	11.111%	1.999%	43.481%
4	15	0	0.000%	0.000%	20.361%

Notes. The sample is restricted to anchor-positive firm-years (firm-years with at least one anchor hit defining gated text). "Controls density bin" groups anchor cases into quartiles based on controls density, measured as controls hits per 1,000 gated characters. "Autonomy present" indicates at least one autonomy keyword hit within gated text. "Autonomy rate" is k/n in each bin; 95% confidence intervals are exact binomial (Clopper–Pearson) intervals.

Figure 2 complements this by showing the full distribution of control density by autonomy status. Most anchor cases cluster at low-to-moderate controls density, but the autonomy-positive cases appear in a higher-density region relative to the mass of autonomy-negative cases. With only two positive observations, this cannot deliver tight inference on its own; its contribution is interpretive: autonomy is not randomly distributed across the control density support. In other words, the pattern supports the mechanism that autonomy emerges only where control language is salient, consistent with the "governance-bottleneck hypothesis."

Figure 2. Distribution of control density by autonomy status (within anchor cases).



Note. This histogram shows the distribution of controls density (controls/governance hits per 1,000 gated characters) among anchor-positive firm-years, split by autonomy status (Autonomy = 0 vs Autonomy = 1). The figure is descriptive and intended to illustrate where the (rare) autonomy-positive cases fall relative to the overall anchor-case distribution.

Table 4 provides the inferential statement of this mechanism using a median split and a Fisher's exact test. Anchor cases are classified into "high-controls" and "low-controls" based on whether controls density is above or below the sample median. The autonomy rate in the high-controls group is 3.57% (1/28), compared to 2.00% (1/50) in the low-controls group. The point estimates imply a risk ratio of 1.79 and an odds ratio of 1.80 in favor of autonomy, being more likely when controls density is higher. Yet the confidence intervals are wide, and the Fisher exact p-value is not small ($p = 0.592$, one-sided). The interpretation is therefore not "statistical proof," but "directional evidence under severe power constraints." This is exactly what one expects in rare-events governance settings: the mechanism can be real, but the available number of autonomy disclosures is the limiting factor for conventional significance (King & Zeng, 2001; Heinze & Schemper, 2002).

A ridge-logit specification estimated within the anchor sample (Table 5) is consistent with the same bottom line: the coefficient on standardized controls density is not precisely estimated and does not overturn the directional story implied by the nonparametric results. This is not surprising. With extremely sparse autonomy positives, logit-based estimators are known to become unstable (including quasi-separation), making effect sizes sensitive to small perturbations and encouraging reliance on exact tests and transparent descriptive gradients (Heinze & Schemper, 2002; King & Zeng, 2001). For that reason, the most reasonable way to present the channel is exactly to (i) show the binned autonomy gradient and distributional plot, (ii) report the median-split Fisher test with effect-size intervals, and (iii) explicitly acknowledge limited power while emphasizing that all evidence points in the same qualitative direction.

Table 4. Controls density and autonomy within anchor cases: median split, effect sizes, and Fisher's exact test.

Anchor cases (N)	Autonomy present (k)	Median controls density	High controls: autonomy=1	High controls: autonomy=0	Low controls: autonomy=1	Low controls: autonomy=0	Autonomy rate (high controls)	Autonomy rate (low controls)	Risk ratio (high/low)	RR 95% CI (low)	RR 95% CI (high)	Odds ratio (high/low)	OR 95% CI (low)	OR 95% CI (high)	Fisher's exact p (greater)
78	2	1.923	1	27	1	49	3.571%	2.000%	1.786	0.000	5.357	1.800	0.238	13.863	0.592

Notes. The sample includes N = 78 anchor-positive firm years. The anchor sample is split into "high controls" vs "low controls" using the sample median control density (1.923 control hits per 1,000 gated characters). The table reports the 2×2 counts (autonomy present/absent by high/low controls), autonomy rates within each group, and effect sizes comparing high vs low controls (risk ratio and odds ratio) with 95% confidence intervals. The reported p-value is from a one-sided Fisher's exact test ("greater") testing whether autonomy is more frequent in the high-controls group.

Table 5. Ridge-penalized logistic **regression** of autonomy on controls density (within anchor cases).

Term	Coef.	Odds ratio	95% CI (OR)	p (boot, 2-sided)
controls_density_z	-0.007	0.993	[0.331, 2.410]	0.798
Y2022	-0.226	0.798	[0.298, 2.663]	0.588
Y2023	-0.302	0.740	[0.310, 3.929]	0.591
Y2024	-0.419	0.658	[0.076, 1.760]	0.471
Y2025	0.737	2.089	[0.528, 12.376]	0.414

Notes. The dependent variable is Autonomy = 1 for anchor-positive firm-years with at least one autonomy hit in the gated text. The main predictor controls_density_z is controls density standardized to a z-score (mean 0, SD 1) within the estimation sample. Year indicators (Y2022–Y2025) are included with 2021 as the omitted reference year. Estimates use ridge-penalized logit to stabilize inference under rare events. Reported "Odds ratio" values are exp(coefficient). Confidence intervals and bootstrap p-values follow the resampling procedure described in the text (firm-level dependence respected via clustering).

Substantively, these results support a coherent interpretation of why the "shift toward agentic AI" matters in finance. The diffusion result (Table 1) shows a transition from absence (2021–2023) to observable presence (2024–2025) in formal disclosures, consistent with early-stage institutionalization of agentic deployment. The channel result (Tables 2–3; Figures 1–2) clarifies what that transition means: the frontier is not "more AI talk," but movement toward autonomy under governance constraints. In a regulated domain, adoption is not constrained primarily by model capability; it is constrained by the organization's ability to evidence control human gates, approvals, audit trails, monitoring, and accountability language before autonomy can be credibly described. That is why the future-facing claim "the trajectory points toward agentic AI" is best framed as a statement about the path of diffusion conditional on controls: autonomy remains rare today, but when it appears it is disproportionately located where controls density is higher, implying that the speed of diffusion is plausibly governed by governance maturity rather than by generic AI enthusiasm (Raji et al., 2020; Gebru et al., 2021).

4. Conclusion

This study documents an early transition in how financial institutions discuss advanced AI capabilities in formal annual disclosures. The central empirical pattern is temporal: agentic AI references are essentially absent through 2023 and then emerge in 2024 (0.4% of firm-years) before rising in 2025 (1.6% of firm-years). While the levels remain small, the time profile is informative: the adoption signal is not gradual across the full window but instead appears as a late takeoff consistent with an emergence phase in which a technology vocabulary begins to stabilize inside regulatory reporting. Within the subsample of filings that contain agentic language, autonomy signals are rare, but they appear disproportionately where governance language is denser in the same local context. In the split-sample channel test, autonomy is observed only in the higher controls density region (4% versus 0% in the lower region). The corresponding association is directionally consistent with the "governance-bottleneck" view: institutions tend to talk about action-taking systems alongside language that signals monitoring, approvals, auditability, and safeguards. At the same time, the autonomy events are extremely sparse, which makes the statistical power limited and implies that the mechanism should be interpreted as a trajectory-consistent pattern rather than a definitive causal claim.

Policy implications follow directly from the distinction between assistance and execution. Our evidence is disclosure-based and, therefore, best interpreted as a screening signal that helps identify where firms are beginning to describe systems with the capability to take actions and where they also emphasize governance language in the same passages. A practical implication is that controls density or a similar governance language measure could help prioritize which filings and which passages merit closer review. If a

filing describes systems that initiate workflow steps in credit and underwriting, fraud operations, customer communications, or AML and sanctions triage, reviewers could look for clearer disclosure on what the system is allowed to do, what actions require human approval, and what monitoring and auditability language accompanies those action rights. In this framing, the relevant policy question is not whether a firm mentions AI, but whether it describes execution scope and corresponding constraints. A second implication is for disclosure practice. When firms choose to discuss agentic capabilities, more specific reporting on action scope, permissioning, approval gates, escalation, and logging would reduce ambiguity and make it easier to distinguish assistance from execution across institutions. These implications are consistent with the paper's core finding that autonomy language is rare and appears only where governance language is relatively dense, suggesting that disclosure about controls is part of how action-taking systems are made credible in regulated finance.

The limitations of this paper are clear and define its research value. The measures are disclosure-based and therefore capture what firms choose to report, not necessarily what they deploy. The approach is intentionally interpretable (dictionary plus local gating), but any keyword-driven system can miss implicit descriptions or capture broad language that is not truly operational. The time window is short, and the key autonomy signal is rare, making fine-grained inference fragile and warranting caution when interpreting cross-sectional heterogeneity. Finally, the sample is restricted to a finance-SIC universe and to annual filings; faster-moving disclosure channels (earnings calls, investor presentations, press releases, risk disclosures outside 10-Ks) may reveal earlier signals and richer variation.

Future research can build directly on these constraints. The first step is horizon expansion (earlier years and more recent filings) to convert emergence-phase patterns into diffusion curves. A second direction is triangulation across disclosure types to test whether agentic language appears first in high-frequency communications and only later in annual reporting. Third, linking disclosure measures to observable outcomes would sharpen the "why it matters" argument: operational incidents, remediation costs, compliance actions, cyber events, efficiency metrics, or risk-weighted asset dynamics. Finally, measurement can be strengthened with hybrid methods that preserve auditability while improving recall, e.g., supervised classifiers trained on verified snippets, or embedding-based retrieval to detect agentic descriptions that do not use canonical keywords, while maintaining the transparency required for finance and policy audiences.

Author Contributions: Conceptualization, Atta ul Mustafa; methodology, Atta ul Mustafa; formal analysis, Atta ul Mustafa; investigation, Atta ul Mustafa; writing and original draft preparation, Atta ul Mustafa; writing, review and editing, Ahmet Faruk Aysan; Project Administration, Ahmet Faruk Aysan; Validation, Ahmet Faruk Aysan. All authors have read and agreed to the published version of the manuscript

Funding: This research received no external funding.

Data Availability Statement: Data available upon request from the corresponding author.

AI Use Statement: The authors used Grammarly for grammar and language refinement. All content was carefully reviewed and verified by the authors.

Conflicts of Interest: The authors declare no conflict of interest.

References

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human–AI interaction. Proc*eedings of the 2019 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3290605.3300233

Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech era. SSRN Working Paper. https://doi.org/10.2139/ssrn.3063448

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3442188.3445922

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. *The Journal of Finance*, 77(1), 5–47. https://doi.org/10.1111/jofi.13090

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1), 27–38. https://doi.org/10.1093/biomet/80.1.27

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92. https://doi.org/10.1145/3458723

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57(3), 535–574. https://doi.org/10.1257/jel.20181020

Hassan, T. A., Hollander, S., van Lent, L., & Tahoun, A. (2019). Firm-level political risk: Measurement and effects. *The Quarterly Journal of Economics*, 134(4), 2135–2202. https://doi.org/10.1093/qje/qjz021

Heinze, G., & Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine*, 21(16), 2409–2419. https://doi.org/10.1002/sim.1047

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, 9(2), 137–163. https://doi.org/10.1093/oxfordjournals.pan.a004868

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1), 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), Article 115. https://doi.org/10.1145/3457607

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency* (FAccT). https://doi.org/10.1145/3287560.3287596

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A., Roberts, M. E., Shariff, A., Tenenbaum, J. B., & Wellman, M. (2019). Machine behaviour. *Nature*, 568(7753), 477–486. https://doi.org/10.1038/s41586-019-1138-y

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency* (FAccT). https://doi.org/10.1145/3351095.3372873

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., & Wen, J.-R. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6), 186345. https://doi.org/10.1007/s11704-024-40231-1

Wooldridge, M., & Jennings, N. R. (1995). Intelligent agents: Theory and practice. *The Knowledge Engineering Review*, 10(2), 115–152. https://doi.org/10.1017/S0269888900008122

Qu, C., Dai, S., Wei, X., Cai, H., Wang, S., Yin, D., Xu, J., & Wen, J.-R. (2025). Learning tools with large language models: A survey. *Frontiers of Computer Science*, 19(8), 198343. https://doi.org/10.1007/s11704-024-40678-2