SEC Form 4 Insider Trading Analysis: Large-Scale Signal Extraction, Rule 10b5-1 Detection, and a Framework for LLM Augmentation*

Mohammed Al-Rebh[1], Kush Malhotra[2], Meiting Wu[2]

[1]Saudi Aramco rebhmy@aramco.com

[2]Department of Industrial Engineering and Operations Research Columbia University {km4149, mw3937}@columbia.edu

ABSTRACT

This paper presents a large-scale empirical analysis of insider trading signals extracted from SEC Form 4 filings spanning 22 years (May 2003–September 2025) across 2,515 Russell 3000 companies. We assembled a dataset of 7,456,167 insider transactions—comprising 5,168,131 non-derivative and 2,288,036 derivative records—representing one of the most comprehensive insider trading datasets compiled from public SEC filings. Our primary technical contribution is a hybrid Rule 10b5-1 plan detection methodology that combines SEC's standardized indicator fields (available post-2023) with exhaustive regex pattern matching on historical footnotes and remarks fields, increasing detection coverage from 2.45% (182,544 transactions) to 11.13% (829,525 transactions)—a 354.4% improvement. This approach enables consistent, longitudinal analysis of planned versus discretionary insider trading across the full study period. We document systematic patterns in transaction type composition, insider role stratification, and temporal adoption trends, including a notable increase in Rule 10b5-1 plan adoption from approximately 7% (2003–2008) to approximately 23% (2021). A detailed case study of Salesforce Inc. reveals that 100% of senior executive transactions are conducted under formalized trading plans, illustrating the implications for interpreting insider trading signals at mature technology companies. We propose a five-step framework—feature engineering, quintile group formation, return analysis, robustness validation, and factor integration—as the basis for future predictive signal research.

Keywords: SEC Form 4, insider trading, Rule 10b5-1, EDGAR, signal extraction, ETL pipeline, corporate governance, equity markets

I. INTRODUCTION

Insider trading by company officers, directors, and substantial shareholders represents a unique window into how informed participants evaluate their own companies. The fundamental regulatory framework—Rule 10b-5 of the Securities Exchange Act—prohibits trading on material non-public information, creating a legal constraint that shapes insider behavior. Within this constraint, however, legitimate insider transactions provide valuable signals about insider confidence, risk tolerance, and equity diversification strategies. Understanding these patterns has implications for market efficiency, price discovery, and investor decision-making.

SEC Form 4 is the primary transparency mechanism through which insider transactions are disclosed to the public within two business days of their occurrence. This timing structure creates an analytical window: transaction dates reveal what insiders chose to do, filing dates reveal when markets learned of these choices, and the mandated two-day delay provides a bounded timeframe for examining information flow and market reaction.

Rule 10b5-1, adopted in 2000, created a safe harbor for insiders establishing pre-planned trading arrangements, enabling disciplined equity management while providing legal protection from liability. The adoption of these plans has varied dramatically across companies and time periods, and understanding this variation reveals important information about the evolution of corporate governance practices.

Prior research on insider trading signals has largely been constrained by data availability. SEC only standardized dedicated Rule 10b5-1 indicator fields starting in 2023, making rigorous historical analysis of planned versus discretionary trading difficult. This paper addresses that gap by developing a hybrid detection methodology that achieves comprehensive 10b5-1 identification across 22 years of data.

A. Research Questions

This project is organized around six explicit research questions:

Q1: What insider trading patterns exist within the Russell 3000, and do they vary systematically by transaction type and insider role?

Q2: Can Rule 10b5-1 trading plans be reliably identified from Form 4 filings across all historical periods using comprehensive pattern detection?

Q3: What is the temporal evolution of Rule 10b5-1 plan adoption across the 22-year dataset?

Q4: How do insider trading patterns differ between derivative and non-derivative transactions?

Q5: What are the characteristics and magnitude of data quality issues in historical Form 4 data?

Q6: Can focused company-level analysis reveal actionable trading patterns and insider behavior insights?

II. RELATED WORK AND BACKGROUND

The academic literature on insider trading encompasses two broad streams: legal and compliance studies examining regulation and enforcement, and financial economics research examining the information content of insider transactions. Early work by Jaffe (1974) and Seyhun (1986) established that insider purchases—particularly by company officers—predict positive abnormal returns in subsequent months, consistent with insiders possessing private information about firm prospects. Lakonishok and Lee (2001) extended this finding across a large sample, documenting that the predictive power is concentrated in purchase signals rather than sales, consistent with the hypothesis that sales are driven primarily by diversification needs rather than negative information.

The introduction of Rule 10b5-1 in 2000 significantly altered the landscape. Jagolinzer (2009) documented that executives with 10b5-1 plans achieve superior trading performance relative to non-plan trades, suggesting information advantages persist even within planned trading programs. More recent work by SEC researchers (2022) raised concerns about strategic manipulation of plan entry and modification—precisely the compliance concerns that motivated the 2023 amendments standardizing 10b5-1 disclosure in Form 4.

The Sarbanes-Oxley Act of 2002 accelerated Form 4 filing requirements from ten days to two business days, substantially increasing the informativeness of SEC filings for real-time signal extraction. The 2023 SEC amendments additionally mandated the Rule 10b5-1 indicator field, creating a data discontinuity that motivates our hybrid detection methodology bridging pre- and post-2023 data.

Our work builds on this literature by contributing: (1) a comprehensive 22-year dataset of Russell 3000 insider transactions, (2) a validated hybrid detection methodology enabling longitudinal 10b5-1 analysis, and (3) systematic empirical characterization of transaction type heterogeneity and insider hierarchy composition at scale.

III. DATASET

A. Data Sources and Universe

Our data originates from SEC EDGAR, which provides free public access to all Form 4 filings. The analysis universe is defined by the Russell 3000 index, comprising approximately 3,000 U.S. publicly-traded companies spanning the market capitalization spectrum from large-cap S&P 500 constituents to small-cap firms with market capitalizations above approximately $2 billion. The Russell 3000 was selected because: (1) it is large enough for statistically robust findings; (2) it encompasses diverse industries and company sizes; (3) it represents the primary deployment universe for institutional and algorithmic investors; and (4) insider trading data coverage is comprehensive relative to smaller-cap universes.

The analysis period spans May 5, 2003—the earliest date with complete EDGAR coverage—through September 24, 2025, representing 22 years and 142 days of continuous data. This span captures four distinct market regimes: the pre-financial-crisis period (2003–2007), the financial crisis (2008–2009), post-crisis recovery (2010–2019), and the pandemic era with elevated volatility (2020–2025).

B. Dataset Statistics

The assembled dataset contains 7,456,167 insider transactions across 2,515 unique Russell 3000 companies. Table I provides a comprehensive summary of dataset characteristics.

TABLE I

Dataset Summary Statistics

Metric	Value	Share / Note
Total Insider Transactions	7,456,167	100.0%
Non-Derivative Transactions	5,168,131	69.3%
Derivative Transactions	2,288,036	30.7%
Rule 10b5-1 (Direct Indicator)	182,544	2.45%
Rule 10b5-1 (Hybrid Method)	829,525	11.13%
Unique Russell 3000 Companies	2,515	—
Study Period	May 2003 – Sep 2025	22 years
Non-Derivative 10b5-1 Adoption	756,993	14.65%
Derivative 10b5-1 Adoption	72,532	3.17%
Date Anomalies Identified	7,162	0.12%
Peak Transaction Year (2021)	426,463	5.7%
Total Cleaned Output Size	~4 GB	—

IV. METHODOLOGY

A. Data Extraction Pipeline

The end-to-end ETL pipeline consists of four primary stages, each designed to handle 7.46 million transactions while maintaining data integrity and complying with SEC EDGAR access guidelines.

Stage 1 — CIK Reconciliation. The Russell 3000 holdings list provides company names and ticker symbols; SEC EDGAR requires CIK (Central Index Key) identifiers. We implemented two-stage matching: (1) direct exact-match lookup using the SEC Company Tickers JSON database, successfully resolving approximately 2,300 companies; and (2) fuzzy string matching using the Levenshtein distance algorithm with a 90% similarity threshold for unmatched companies, recovering approximately 150 additional matches.

Stage 2 — Form 4 Acquisition. A dual-path approach was implemented: the SEC EDGAR REST API for targeted retrieval of individual company filings (with exponential backoff retry logic: 1s, 2s, 4s, up to 60s maximum, to comply with the SEC's 10 requests-per-second rate limit), and bulk ZIP archives containing complete submission index files for efficient historical batch processing.

Stage 3 — Link Extraction. Filings are available in HTML, XML, and occasionally PDF formats. The pipeline filters to XML and HTML only, excluding PDFs which would require optical character recognition beyond project scope. This filtering captures the vast majority of processable filings while substantially reducing computational workload.

Stage 4 — Table Parsing. Form 4 documents contain both non-derivative and derivative transaction tables in structurally distinct sections. We developed separate parsing logic for each: BeautifulSoup for HTML documents and lxml for XML documents (navigating the SEC XBRL taxonomy). Critical extracted fields include: reporting person name and CIK, relationship to company, security description, transaction code (P/S/M/X), quantity, price, transaction date, filing date, and associated footnotes/remarks.

Processing Architecture. A batch processing architecture grouped transactions into 50,000-record chunks to prevent Python memory overflow. The complete extraction required 5.5 days of processing—exceeding the initial 2.5-day estimate due to format variations across 22 years of historical data, derivative table structural complexity, and API rate-limit compliance delays.

B. Hybrid Rule 10b5-1 Detection Methodology

Rule 10b5-1 plan detection is the core methodological contribution of this paper, enabling comprehensive longitudinal analysis of planned versus discretionary insider trading.

The challenge stems from regulatory evolution: the SEC standardized a dedicated Rule_10b5_1_Plan indicator field in Form 4 filings only beginning in 2023. Prior to 2023, companies disclosed 10b5-1 plan information textually within transaction footnotes and remarks fields. We address this gap through a two-component hybrid approach:

Method 1 — Direct Indicator (2023+ filings): The SEC's standardized indicator field provides definitive identification when present. Analysis of post-2023 data reveals this method captures 182,544 transactions (2.45% of total dataset; 24.6% of 2024 transactions; 23.5% of 2025 transactions).

Method 2 — Regex Pattern Matching (all historical periods): Comprehensive regular expression patterns applied to footnotes and remarks fields, covering 10+ variations of how companies reference Rule 10b5-1 plans:

- Direct rule citations: 'Rule 10b-1', 'Rule 10b 5-1', 'Rule 10b5-1', 'Rule 10b-5-1'
- Abbreviated forms: '10b-1', '10b 5-1', '10b5-1', '10b-5-1'
- Parenthetical forms: '10b(5)', '10b(5)-1', '10(b)(5)', '10(b)(5)-1'
- Personal plan names: 'Mr./Ms./Dr. [Name]'s 10b(5) Plan', '[Name]'s 10b(5)-1 Plan'
- All spacing and hyphenation permutations of the above

Rule 10b-18 references (pertaining to company share repurchase programs, not insider trading plans) are explicitly excluded to ensure specificity.

Validation was conducted by reviewing 8,615 sample transactions (5,050 non-derivative; 3,565 derivative). Post-2023 data—where both direct indicators and regex matches are available—showed 89% agreement between the two methods, confirming high accuracy of the regex approach for pre-2023 periods where only footnote text is available. Table II summarizes the detection results.

TABLE II

Rule 10b5-1 Detection Method Comparison

Detection Method	Transactions Identified	Coverage	Notes
Direct Indicator Only (2023+)	182,544	2.45%	High confidence; limited to post-2023 data only
Regex Pattern Matching (all periods)	646,981 additional	+8.68pp	Validated via 8,615 sample review; 89% agreement with direct indicators post-2023
Hybrid Method (combined)	829,525	11.13%	354.4% improvement; enables 22-year longitudinal analysis

C. Data Quality and Anomaly Detection

Filing dates are system-generated by the SEC and are highly reliable. Transaction dates, by contrast, are manually entered by filers and are susceptible to data entry errors. We identified and removed transactions whose reported dates fall outside the filing window (May 5, 2003 – September 24, 2025). Dates after the window are definitively invalid (filings cannot precede transactions); dates before the window typically arise from amended filings preserving original transaction dates from earlier periods for which comprehensive filing coverage is unavailable.

This filter removed 9,010 records (0.13% of the dataset), leaving 7,447,157 clean records. Among retained records, 7,162 date anomalies were flagged (0.12% anomaly rate), all investigated and documented. Key identifier fields (CIK and ticker symbol) are fully populated across the dataset. A small number of firms (18 CIKs) are associated with two tickers, reflecting ticker changes over time or multiple share classes—expected behavior in a long-horizon dataset.

D. Insider Title Normalization

The Reporting_Person_Title field in Form 4 filings is free-text and exhibits substantial variation: identical roles may appear under multiple textual representations. We manually categorized insider titles into five standardized hierarchical levels: C_LEVEL (CEO, CFO, CTO, COO, CMO, and equivalents), EXECUTIVE_MANAGEMENT (EVP, SVP, President, General Counsel, Controller), BOARD (Director, Board Member, Independent Director), COMBINED_BOARD_EXEC (dual board–executive roles), and OTHER/NOISE (uncategorized or non-operational entries). Table III summarizes the classification and distribution.

TABLE III

Insider Hierarchy Classification and Transaction Composition

Category	Representative Titles	Deriv. Share	Non-Deriv. Share
C_LEVEL	CEO, CFO, CTO, COO, CMO, CIO, CLO, CHRO, CRO	21.7%	21.6%
EXECUTIVE_MANAGEMENT	EVP, SVP, President, General Counsel, Controller	26.2%	24.1%
BOARD	Director, Board Member, Independent Director	2.6%	8.7%
COMBINED_BOARD_EXEC	Director & Officer dual roles	~1.5%	~2.4%
OTHER / NOISE	Uncategorized, free-text, non-operational	~48%	~43%

V. EMPIRICAL ANALYSIS

A. Transaction Type Stratification

The dataset divides into two fundamental transaction types with structurally distinct characteristics. Non-derivative transactions (5,168,131 records, 69.3%) represent direct equity ownership changes—insider purchases and sales of company stock. These are strategically significant because they reflect deliberate insider decisions about personal equity exposure: a purchase signals confidence in future prospects, while a sale signals diversification needs or valuation concerns. Derivative transactions (2,288,036 records, 30.7%) encompass options, warrants, convertible securities, and other derivatives. The 30.7% proportion is substantially higher than the 10–15% commonly cited in prior literature, indicating that equity compensation-related activity is more significant in the Russell 3000 universe than generally assumed.

Rule 10b5-1 adoption rates reveal a striking 4.6x differential between transaction types: 14.65% for non-derivative versus 3.17% for derivative transactions. This differential reflects the fundamentally different nature of these transactions. Non-derivative transactions—where insiders make deliberate buy or sell decisions—require formal compliance planning because they represent psychologically and legally significant choices about equity positioning. Derivative exercises are predominantly mechanistic responses to predetermined vesting schedules and option expiration dates, requiring less formal planning. The data suggests that insiders intuitively recognize this distinction and apply formal compliance mechanisms primarily to high-stakes equity positioning decisions.

B. Insider Filing Intensity and Identity Concentration

The distribution of non-derivative filing counts per insider is extremely right-skewed. The median insider submits 23 non-derivative filings over the 22-year period, while the mean is approximately 65—reflecting the influence of heavy-tail observations. At the upper end of the distribution, the 95th percentile insider has filed approximately 226 times, the 99th percentile approximately 606 times, and the maximum exceeds 28,000 filings for a single individual. This concentration of activity in a small subset of highly-active insiders suggests that a minority of insiders drives a disproportionate share of observable market activity, with implications for signal design: weighting by transaction count rather than unique-insider count may be more appropriate for predictive applications.

C. Temporal Evolution of Rule 10b5-1 Adoption

Longitudinal analysis of 10b5-1 plan adoption across 22 years reveals distinct phases in insider trading compliance evolution. Adoption was sparse in the early period (2003–2008), approximately 7% of transactions, reflecting the nascent state of Rule 10b5-1 plan infrastructure following the 2000 rule adoption. Adoption accelerated through the 2010s as corporate legal departments and financial advisors standardized 10b5-1 plan offerings, reaching approximately 23% by 2021—the peak transaction year with 426,463 total transactions. Post-2021, elevated adoption rates reflect both organic growth and the anticipation of the 2022 SEC amendments tightening 10b5-1 plan requirements. This temporal continuity—only possible through our hybrid detection methodology—reveals compliance evolution that would be completely invisible in post-2023 data alone.

D. Temporal Composition and Filing Day Distribution

Analysis of filing weekday distributions reveals structural heterogeneity consistent with the coexistence of two distinct behavioral regimes within the same regulatory framework. Discretionary market trades—purchases and sales at prevailing prices—show patterns consistent with active market decision-making, concentrated on trading days with some preference for earlier in the trading week. Compensation-related transactions (RSU vesting, option exercises) exhibit concentrated monthly clustering driven by corporate equity plan vesting schedules, particularly around calendar-quarter boundaries. This structural separation motivates treating the two transaction types differently in downstream empirical work.

VI. CASE STUDY: SALESFORCE INC.

A. Overview and Data Characteristics

We conducted a detailed case study on Salesforce Inc. (NYSE: CRM), selected because: (1) it ranks among the highest-volume insider trading companies in the Russell 3000, with 59,238 total transactions; (2) it is a large-cap technology company with sophisticated governance practices; and (3) Form 4 filing records are essentially complete over the study period. This case study demonstrates how our methodology can yield actionable company-specific insights.

Salesforce's transaction composition diverges markedly from the Russell 3000 average: non-derivative transactions comprise 91.2% of Salesforce activity (versus the 69.3% Russell 3000 average), while derivative transactions represent only 8.8% (versus 30.7%). This suggests that despite a substantial equity compensation program, Salesforce executive activity is driven primarily by deliberate equity positioning rather than routine option exercises.

B. Rule 10b5-1 Adoption Among Senior Executives

The most striking finding from the Salesforce case study is the universal adoption of Rule 10b5-1 plans among senior executives. Analysis of CEO, CTO, CCO, and Board member transactions reveals that 100% of observed trades are executed under Rule 10b5-1 plans. This universal adoption rate—contrasting sharply with the 14.65% non-derivative average across the Russell 3000—indicates that Salesforce has institutionalized Rule 10b5-1 plans as mandatory compliance procedure for executive equity management, rather than treating them as optional safe harbors.

CEO transactions are uniformly pre-planned, typically ranging from six- to seven-figure values—consistent with significant personal wealth positions requiring disciplined diversification management. CTO and CCO transactions similarly show universal plan adoption, confirming that formal trading plans are standard across the entire executive leadership team. Board member transactions also exhibit 100% plan adoption for top board members, reflecting board-level commitment to governance best practices beyond regulatory minimums.

C. Implications for Signal Interpretation

The Salesforce case study illustrates a critical nuance for insider trading signal construction: when 100% of executive transactions are subject to pre-planned arrangements, the signal content of individual trades changes fundamentally. A pre-planned transaction to sell shares on specified future dates reflects strategic portfolio management rather than a response to material information. Planned trading decisions are made weeks or months in advance, removing the contemporaneous information advantage that makes reactive insider trading potentially informative.

This implies that for companies with universal or near-universal plan adoption, insider trading signals should be interpreted as reflecting long-term equity diversification strategies rather than short-term outlook changes. Segmenting the signal universe by plan adoption intensity—differentiating companies where insider trading remains predominantly reactive from those where it is fully formalized—may substantially improve the predictive power of insider trading factors in downstream quantitative applications.

VII. KEY FINDINGS

The following key findings emerge from our analysis:

Finding 1 — The Hybrid Detection Methodology Provides a 354.4% Coverage Improvement. Using only direct SEC indicator fields, 2.45% (182,544) of transactions are identifiable as subject to Rule 10b5-1 plans. The hybrid approach—combining direct indicators with validated regex pattern matching—identifies 11.13% (829,525 transactions). Post-2023 cross-validation confirms 89% agreement between methods, validating the approach for pre-2023 historical analysis.

Finding 2 — Insider Trading Compliance Has Fundamentally Evolved Over 22 Years. Rule 10b5-1 plan adoption increased from approximately 7% (2003–2008) to approximately 23% (2021), a three-fold increase over the study period. This evolution—only quantifiable through longitudinal data enabled by the hybrid detection methodology—reveals how corporate governance practices and legal compliance infrastructure have matured.

Finding 3 — Non-Derivative and Derivative Transactions Are Structurally Distinct. The 4.6x differential in Rule 10b5-1 adoption (14.65% non-derivative vs. 3.17% derivative), combined with the distinct temporal distribution patterns, confirms that these transaction types reflect fundamentally different economic activities and should be treated as separate analytical categories in signal construction.

Finding 4 — Insider Activity Is Highly Concentrated. The median insider submits 23 filings over the 22-year period, but the maximum exceeds 28,000. The 99th percentile insider submits approximately 606 filings. This concentration implies that a small number of highly-active insiders account for a disproportionate share of observable activity, with implications for cross-sectional signal weighting.

Finding 5 — Transaction Direction Is Heavily Sale-Dominated. Approximately 83% of non-derivative transactions are sales versus 17% purchases across the Russell 3000. This asymmetry reflects long-term insider equity diversification rather than concentrated bullish conviction, consistent with the prior literature.

Finding 6 — Salesforce Exhibits Universal Executive Plan Adoption. All observed CEO, CTO, CCO, and Board member transactions at Salesforce are executed under Rule 10b5-1 plans, compared to the 14.65% Russell 3000 non-derivative average. This finding illustrates that signal interpretation must account for firm-level governance sophistication.

VIII. FUTURE WORK

The dataset and methodology developed in this paper provide the foundation for several promising research directions.

A. Predictive Signal Testing Framework

The primary near-term research direction is systematic testing of whether insider trading patterns contain predictive signals for subsequent stock returns. We propose a five-step framework:

Step 1 — Feature Engineering: Construct quantitative features from insider trading data to segment the stock universe. Candidate features include binary indicators (e.g., has_insider_buy over specified time horizons), ratio-based measures (e.g., insider buying as a percentage of total trading volume), role-stratified features comparing officer, director, and major shareholder activity, and the Net Insider Buying Ratio (NIBR) on a [−1, +1] scale. Features should be tested across multiple lookback windows (30, 60, 90, 180, and 360 days).

Step 2 — Group Formation: Segment the stock universe into quintile or decile groups ranked by feature value. Group formation should employ sector-neutral construction—ranking stocks within each sector independently before combining—to isolate insider trading signals from implicit sector exposures.

Step 3 — Return Analysis: Test whether forward returns differ systematically across groups. The primary test is monotonic return gradients from lowest to highest feature quintile, with a long-short portfolio comparing the highest versus lowest quintile providing the core return spread. Multiple forward return horizons (1, 3, 6, and 12 months) should be examined with appropriate statistical significance testing.

Step 4 — Robustness Validation: Stratify results by market capitalization (small-cap, mid-cap, large-cap) and sector to confirm signal generalizability and identify where signal strength is concentrated.

Step 5 — Factor Integration: Test the insider trading signal against established risk factor models (Fama-French, Barra) to determine whether observed return spread constitutes genuine alpha or is attributable to known factor exposures.

B. Extended Research Directions

Additional research directions include: (1) Technology sector comparison across major technology companies (Meta, Alphabet, Amazon, Microsoft) to identify sector-specific patterns; (2) Financial sector analysis given the high insider trading volumes typically observed in banks and financial services companies; (3) Anomalous trading detection to identify companies with unusually high insider purchase rates (relative to the 83% sales-dominant baseline) as potential high-conviction insider confidence signals; and (4) Multi-company comparison across sectors to test whether insider trading patterns vary systematically by industry.

IX. LARGE LANGUAGE MODELS: INTEGRATION, APPLICATIONS, AND FUTURE DIRECTIONS

The analytical pipeline developed in this paper—spanning raw XML extraction from 7.46 million Form 4 filings through hybrid Rule 10b5-1 detection and empirical signal characterization—establishes a structured, clean dataset that is ideally positioned for integration with Large Language Models (LLMs). While the current work is grounded in rule-based pattern matching, regex detection, and quantitative feature engineering, LLMs offer a complementary and transformative layer of intelligence across every stage of this research program. This section articulates the specific roles LLMs can play, the concrete use cases that justify their inclusion, and the broader research agenda they enable.

A. Semantic Enhancement of Rule 10b5-1 Detection

The hybrid Rule 10b5-1 detection methodology described in Section IV.B achieves 89% agreement with direct SEC indicators in post-2023 validation—a strong result, but one that leaves an 11% discordance gap attributable to unconventional disclosure language that eludes regex patterns. LLMs present a direct solution to this ceiling. A fine-tuned or few-shot-prompted LLM applied to the footnotes and remarks fields of pre-2023 Form 4 filings could identify idiosyncratic phrasing (e.g., non-standard abbreviations, company-specific plan nomenclature, or foreign-language disclosures from cross-listed issuers) that no finite regex vocabulary can anticipate. Preliminary experimentation with models such as GPT-4o or Claude Sonnet using structured prompts—asking the model to classify whether a given footnote implies the existence of a pre-established trading plan—suggests classification accuracy competitive with regex approaches while handling novel language patterns with substantially greater robustness. Incorporating LLM-based classification as a third detection layer in the hybrid methodology could plausibly reduce the discordance gap from 11% toward 5% or below, improving the reliability of the 22-year longitudinal compliance record.

B. Automated Insider Title Normalization and Role Classification

Section IV.D describes the manual normalization of the free-text Reporting_Person_Title field into five hierarchical categories. This process, while carefully validated, is inherently labor-intensive, subjective at the margins, and brittle to novel title strings that emerge over time (e.g., “Chief AI Officer” or “Head of ESG

Strategy”). LLMs are naturally suited for zero-shot and few-shot title classification: given a title string and a target taxonomy, modern instruction-following models can assign titles to hierarchical categories with human-level accuracy while scaling to millions of records. More importantly, LLMs can infer organizational seniority from contextual cues in the title itself—distinguishing a “Senior Vice President of Regional Operations” from a “Vice President of Investor Relations” in ways that rule-based mapping cannot. Replacing or augmenting the current manual taxonomy with an LLM classification layer would reduce the approximately 48% OTHER/NOISE share in the non-derivative category (Table III), improving the granularity of role-stratified signal analyses and enabling finer-grained insider hierarchy segmentation in downstream predictive models.

C. Unstructured Signal Extraction from Footnotes and Remarks

Form 4 footnotes and remarks fields contain substantially richer information than the structured transaction fields alone. Insiders sometimes disclose the rationale for transactions (“shares sold to cover tax withholding on RSU vesting”), the existence of previously undisclosed plans (“pursuant to a pre-established Rule 10b5-1 plan adopted on [date]”), or material context about the transaction structure (“transaction effected pursuant to a 10b5-1 plan entered into during an open trading window”). Current analysis treats this text primarily as a signal detection surface for 10b5-1 identification, but the information content extends far beyond plan presence. LLMs can extract structured signals from these unstructured fields at scale: identifying disclosed transaction rationales, flagging transactions where insiders explicitly reference company-specific events, detecting emotional or urgency-signaling language, and extracting implicit timing signals (e.g., plan adoption dates mentioned in footnotes that predate the formal indicator field). Building a systematic LLM-based annotation layer over the footnotes corpus would create a richer feature set for the predictive signal testing framework proposed in Section VIII.A, potentially revealing informational signals that structured fields alone cannot capture.

D. LLM-Augmented Predictive Signal Construction

The five-step predictive signal testing framework outlined in Section VIII.A—feature engineering, group formation, return analysis, robustness validation, and factor integration—is designed around quantitative features derived from structured transaction fields. LLMs introduce a qualitative signal dimension that is orthogonal to these features and potentially highly complementary. Specifically, LLMs can be used to synthesize cross-modal signals: given a cluster of insider transactions alongside contemporaneous SEC filings (10-K, 10-Q, 8-K), earnings call transcripts, and news articles retrieved for the same issuer and time window, an LLM can assess the degree of alignment or divergence between insider behavior and publicly disclosed corporate narratives. An insider purchasing shares while the company’s 10-K discusses deteriorating fundamentals represents a qualitatively different signal than an insider purchasing alongside an optimistic earnings call. Encoding this contextual alignment as an LLM-generated feature—for example, an “insider-narrative divergence score”—and incorporating it into the quintile formation process in Step 2 could substantially improve the signal-to-noise ratio of insider trading factors, particularly for the large-cap universe where pure transaction-based signals tend to be weakest.

E. Retrieval-Augmented Generation for Real-Time Insider Intelligence

A compelling applied extension of this research is the construction of a Retrieval-Augmented Generation (RAG) system that provides real-time, natural-language insider trading intelligence to investment professionals. In this architecture, the 7.46-million-transaction dataset serves as the retrieval corpus: when a portfolio manager queries “what is the recent insider trading pattern for Salesforce executives?” or “which Russell 3000 companies show anomalously high insider purchase rates in Q3 2025?”, the system retrieves relevant transaction records, computes aggregate statistics, and passes them to an LLM for natural language synthesis. The RAG architecture solves a fundamental limitation of standalone LLMs—their lack of proprietary, continuously updated transaction data—while leveraging their strength in natural language generation, contextual reasoning, and multi-document synthesis. Such a system would translate the dataset from a research artifact into a decision-support tool, enabling analysts without quantitative programming backgrounds to query the insider trading database through conversational interfaces. Integration with live EDGAR feeds would further extend the system to near-real-time monitoring, alerting analysts to unusual insider activity within hours of Form 4 filing.

F. Anomaly Detection and Potential Compliance Flagging

Section VIII.B identifies anomalous trading detection—identifying companies with unusually high insider purchase rates relative to the 83% sales-dominant baseline—as a future research direction. LLMs can materially enhance anomaly detection beyond threshold-based statistical flags. When a quantitative model flags an unusual cluster of insider purchases at a given company, an LLM can be tasked with synthesizing the contextual evidence: reviewing the company's recent SEC filings, news coverage, and earnings call transcripts to assess whether the insider activity is consistent with disclosed corporate developments or represents a divergence that warrants further scrutiny. This human-in-the-loop, LLM-assisted anomaly triage pipeline would be valuable both for academic research—filtering false positives from genuine behavioral anomalies—and for regulatory and compliance applications, where identifying potential information asymmetries before they become enforcement actions has significant institutional value. The structured dataset produced by this paper, combined with an LLM reasoning layer, forms the core components of such a system.

G. Technical Considerations and Implementation Pathway

Practical integration of LLMs into this research program requires attention to several technical and operational considerations. Scale is the primary constraint: applying LLM inference to 7.46 million transaction records with associated footnote text is computationally intensive and cost-prohibitive with current API pricing at full corpus scale. The pragmatic solution is tiered application—applying LLM processing selectively to the approximately 650,000 records with non-null footnote text, using smaller open-weight models (e.g., Llama 3 or Mistral-based architectures) for classification tasks where closed-API costs are prohibitive, and reserving frontier model inference for high-value synthesis tasks such as anomaly triage and cross-modal signal generation. Fine-tuning a smaller model on a labeled subset of footnote classifications would further reduce inference costs while preserving accuracy. Prompt engineering considerations specific to financial compliance text—handling technical regulatory language, numerical references, and legal citations—are well-documented in recent finance NLP literature and provide a foundation for implementation. The structured, clean dataset produced by this paper—with its validated fields, normalized insider hierarchies, and hybrid 10b5-1 labels—is precisely the kind of high-quality training and retrieval corpus that LLM-based financial applications require, positioning this work as a natural foundation for the next generation of AI-assisted insider trading research.

X. LIMITATIONS

Several limitations should be noted when interpreting our findings:

Data Quality and Coverage. While our dataset achieves 99.88% temporal data quality (only 0.12% date anomalies), residual data entry errors in transaction prices, quantities, and titles may affect fine-grained analyses. The optional and free-text nature of insider title disclosure results in high missing rates for the Reporting_Person_Title field, requiring manual normalization with associated categorization uncertainty.

Regex Pattern Matching Completeness. While our hybrid detection achieves 89% agreement with direct indicators in post-2023 validation, the 11% discordance represents both false positives (regex matches not confirmed by direct indicator) and false negatives (direct indicators without regex matches). Some company-specific disclosure language may be missed; the true pre-2023 10b5-1 coverage rate remains uncertain.

Scope Constraints. This analysis characterizes patterns in the data without constructing or backtesting predictive models, which would require live market data integration and implementation infrastructure beyond the scope of an academic course project. The Salesforce case study, while illustrative, represents a single large-cap technology company and may not generalize across sectors and size segments.

Forward Return Analysis. Initial quintile-based forward return analysis using the NIBR metric showed strong selling indication in September 2025, but comprehensive predictive testing across the full 22-year dataset remains a direction for future work. Empirical results from full backtesting may reveal strong signals, weak signals, or no statistically significant signals—each outcome is valuable, but uncertainty is inherent.

XI. CONCLUSION

This paper presents a comprehensive, large-scale analysis of insider trading signals extracted from SEC Form 4 filings spanning 22 years and 7,456,167 transactions across the Russell 3000. Our primary technical contribution—a hybrid Rule 10b5-1 plan detection methodology combining direct SEC indicator fields with validated regex pattern matching on historical footnotes—achieves a 354.4% improvement in detection coverage compared to indicator-only methods, enabling for the first time longitudinal analysis of planned versus discretionary insider trading across the full 2003–2025 period.

Our empirical analysis documents systematic patterns in transaction type composition (69.3% non-derivative versus 30.7% derivative), a pronounced 4.6x adoption differential between transaction types, and a three-fold increase in Rule 10b5-1 plan adoption from 2003 to 2021. The Salesforce case study reveals that governance maturity fundamentally changes the signal content of insider transactions: when 100% of executive trades are pre-planned, signals must be interpreted as reflecting strategic equity diversification rather than contemporaneous information advantages.

The 4-gigabyte clean dataset generated by this project, with its comprehensive Rule 10b5-1 identification and validated data quality, provides a robust foundation for downstream machine learning applications, alternative signal construction, and extended academic research. The proposed five-step predictive signal testing framework—feature engineering, sector-neutral group formation, forward return analysis, robustness validation, and factor integration—outlines a rigorous path from descriptive characterization to quantitative signal validation.

XII. GLOSSARY

Beneficial Ownership: Legal or equitable interest in securities, including direct ownership and holdings through family members or trusts.

CIK (Central Index Key): Unique 10-digit identifier assigned by the SEC to every registered company, required for EDGAR access.

Derivative Securities: Securities deriving value from underlying instruments, including options, warrants, convertibles, and rights.

EDGAR: SEC's Electronic Data Gathering, Analysis, and Retrieval system providing public access to corporate filings.

Form 4: SEC form disclosing material insider transactions within two business days of the transaction date.

NIBR (Net Insider Buying Ratio): A composite metric on a [−1, +1] scale measuring whether insiders are net buyers or sellers over a specified lookback window.

Rule 10b-5: SEC regulation prohibiting trading on material non-public information; foundational insider trading law.

Rule 10b5-1 Trading Plan: Pre-planned trading arrangement providing safe harbor from insider trading liability for covered transactions.

Russell 3000: Index including approximately 3,000 U.S. publicly-traded companies; primary analysis universe for this project.

Transaction Code: Standardized letter codes (P = purchase, S = sale, M = exercise, X = expiration) indicating transaction type in Form 4 filings.

REFERENCES.

[1] SEC Form 4 Instructions and Definitions. U.S. Securities and Exchange Commission. https://www.sec.gov/about/forms/form4.pdf

[2] Rule 10b5-1 Trading Plans Overview. U.S. Securities and Exchange Commission Investor Education. https://www.investor.gov/introduction-investing/investing-basics/role-sec/what-rule-10b51-trading-plan

[3] Russell 3000 Index Documentation. FTSE Russell. https://www.ftserussell.com/products/indices/russell3000

[4] SEC EDGAR Database and API Documentation. U.S. Securities and Exchange Commission. https://www.sec.gov/cgi-bin/browse-edgar

[5] SEC Company Tickers Database. U.S. Securities and Exchange Commission. https://www.sec.gov/files/company_tickers.json

[6] Securities Exchange Act of 1934, Rule 10b-5. 17 C.F.R. §240.10b-5. https://www.law.cornell.edu/cfr/text/17/240.10b-5

[7] Sarbanes-Oxley Act of 2002, Sections 16–18. Public Law 107-204.

[8] Seyhun, H.N. (1986). Insiders' profits, costs of trading, and market efficiency. Journal of Financial Economics, 16(2), 189–212.

[9] Lakonishok, J., & Lee, I. (2001). Are insider trades informative? Review of Financial Studies, 14(1), 79–111.

[10] Jagolinzer, A.D. (2009). SEC Rule 10b5-1 and insiders' strategic trade. Management Science, 55(2), 224–239.