Reading the Insider Cluster: Concurrent Disclosure-Tone Classification Predicts the Cross-Section of Post-Filing Returns

Ahn Hyun

July 2026

Abstract

I classify the tone of 8-K and DEF 14A disclosures filed within five trading days of multi-insider open-market buy clusters on the SEC Form 4 feed, using an ensemble of three large language models, and study whether disclosure tone predicts the cross-section of post-filing returns. On a survivorship-free 2015–2025 panel of 1.9 million Form 4 filings covering 38,261 cluster events, filling a thirty-position long-only portfolio with positive-outlook events first raises the five-day annualized net Sharpe ratio from 1.77 to 2.28 (+29%) at a 25 basis-point round-trip cost. The lift is vendor-invariant across the three models, survives redaction of all filer identifiers from the disclosure text (Cohen's $\kappa = 0.92$ against unmasked labels), and retains two-thirds of its magnitude when the classification is repeated with a language model whose training data end in September 2021 — a vintage-robust lower bound of +0.34 Sharpe. Disclosure timing matters: clusters whose disclosure arrives on the same day as the final insider trade earn a net Sharpe of 0.64, versus 1.86 when the disclosure arrives three to five days later, consistent with same-day filings being reactive news that is already priced. The underlying cluster mechanism delivers a five-day abnormal return of +1.68% ($t = 32.9$, Fama-French five-factor plus momentum residualized) that is sign-consistent across seven macro regimes. The classification prompt and full pipeline are released for replication.

Contents

1 Introduction **3**

2 Related Literature **4**

2.1 The Insider-Trading Return Channel . . . 4

2.2 Disclosure-Text Signals . . . 5

2.3 Standard Baselines . . . 5

2.4 Position of This Study . . . 6

3 Data **6**

3.1 Universe . . . 6

3.2 Form 4 Accessions with Delisting-Return Coverage . . . 6

3.3 CRSP Daily Stock File . . . 6

3.4 Kenneth French Factor Series . . . 7

3.5 Concurrent 8-K and DEF 14A Disclosure Text . . . 7

3.6 Sample Selection Waterfall . . . 7

4 Methodology **8**

4.1 Cluster Detection . . . 8

4.2 Disclosure-Text Classification . . . 8

4.3 Portfolio Construction . . . 8

4.4 Cost Model . . . 9

4.5 Statistical Inference . . . 9

5 Results **9**

5.1 Cluster Mechanism Aggregate Result . . . 9

5.2 Regime Consistency . . . 10

5.3 Cost Sensitivity . . . 10

5.4 Finding 1: A Vendor-Invariant Portfolio Lift from Disclosure Tone . . . 11

5.4.1 Finding 1b: Identifier-Masked Robustness Check . . . 11

5.4.2 Finding 1c: Pre-2024 Vintage Cross-Check . . . 12

5.5 Finding 2: Disclosure-Timing Gap . . . 15

6 Robustness **15**

6.1 Size-Tier Partition with Tier-Specific Transaction Costs . . . 15

6.2 Sector Partition . . . 16

6.3 Winsorization Sensitivity . . . 17

6.4 Cluster-Window Sensitivity . . . 18

7 Discussion **18**

8 Limitations **19**

9 Conclusion **20**

10 References **22**

11 Appendix A: Classification Prompt and Masking Function **24**

11.1 A.1 Prompt template . . . 24

11.2 A.2 Identifier-masking function (Finding 1b) . . . 25

12 Appendix B: Replication **26**

13 Appendix C: Data File Manifest **27**

1 Introduction

Section 16(a) of the Securities Exchange Act of 1934 requires that corporate insiders — officers, directors, and beneficial owners of more than ten percent of a class of equity securities — disclose each change in beneficial ownership on Form 4 within two business days of the transaction. The insider-trading return literature has established two structural regularities on that disclosure feed. Cohen et al. (2012) split individual insider transactions into routine and opportunistic categories and document a six-month cumulative abnormal return on the opportunistic slice on the order of five percentage points. Alldredge and Blank (2019) extend the anchor to the cluster margin — events on which at least two distinct insiders trade in the same direction within a short peer window — and report a monthly abnormal return of approximately two percentage points concentrated in the cluster subset. On the same feed, Cziraki and Gider (2021) document that the aggregate dollar profit to insider trading has increased over time while the per-transaction abnormal return has compressed.

What is not established, in the insider-trading literature that I have been able to survey, is whether the concurrent disclosure body — the 8-K and DEF 14A filings that arrive within a few days of the cluster's terminal transaction date — carries a tractable forward signal that stratifies the cluster events themselves. The 8-K item numbers document what the firm has disclosed (an earnings release, a material contract, a change in control), but the item numbers alone are silent on the tone of the disclosure body. A material-contract 8-K may be a positive contract win or a costly cancellation; an earnings release may exceed or miss guidance. The tone of the body has, until the recent generation of large language models, been prohibitively expensive to extract at the tens-of-thousands-of-filings scale required to overlay it on a decade of cluster events. The natural question — whether the body tone stratifies the cluster events, and if so with what portfolio-relevant magnitude — is therefore open.

I address this question. I assemble a survivorship-free 2015–2025 panel of 1,932,506 Form 4 accessions, detect 38,261 multi-insider open-market buy clusters (two or more distinct reporting-owner central index keys, at least one flagged officer, within a rolling thirty-calendar-day window ending at the cluster terminal transaction date), fetch the concurrent 8-K and DEF 14A text within a $\pm$5-trading-day window of the cluster terminal date, and classify the resulting 14,141 disclosure documents with a zero-shot ensemble of three large language models — OpenAI's GPT-4o-mini, Meta's Llama 3.3 70B, and DeepSeek's Chat v3.1 — into two categories: *positive-outlook*, assigned only when the text contains at least one of seven pre-specified triggers (raised guidance, a new or upsized share repurchase, a material contract win or regulatory approval, a product or clinical milestone, a positive earnings surprise, explicit near-term operational confidence, or buyer-side accretive M&A; the full prompt is in Appendix A), and *neutral* otherwise.

I document two findings that, to my knowledge, are new to the insider-trading literature.

Finding 1. The disclosure-tone class carries a portfolio-level Sharpe lift that is stable across three independent language models. Baseline equal-weighted fill of a thirty-position long-only portfolio on all cluster events yields a five-trading-day annualized net Sharpe ratio of 1.77 at 25 basis-point round-trip transaction cost. Filling the same thirty slots with positive-outlook clusters first, then neutral clusters, then clusters without a concurrent disclosure — a *priority-fill* construction that uses the sparse positive label opportunistically at the top of the portfolio queue rather than restricting the strategy to the 4.5% positive slice, which would leave the portfolio starved on most trading days — lifts the net Sharpe to 2.28, a +29% gain. The lift is stable across the three single-vendor label sources (+0.50 to +0.52 Sharpe) and across the majority-vote, unanimity, and any-

positive aggregations (+0.46 to +0.52 Sharpe). No single vendor is disproportionately responsible for the aggregate lift, which addresses the standard concern that language-model-based signals reflect the idiosyncrasies of a single training corpus. To my knowledge, no prior study on the Form 4 cluster cohort has combined the cluster mechanism with a language-model reading of the concurrent 8-K and DEF 14A content.

Finding 2. Disclosure-timing gap matters: same-day co-disclosure with the cluster terminal date yields a substantially weaker net Sharpe than three-to-five-day gap disclosures. Partitioning the labeled events by the calendar gap between the concurrent disclosure's filed date and the cluster terminal transaction date yields three tiers: same-day (N = 1,285, priority-fill net Sharpe +0.64), close one-to-two days (N = 4,112, +1.63), and wide three-to-five days (N = 6,274, +1.86). The natural reading is that same-day 8-K disclosures are reactive news that has already been absorbed into the price on the disclosure day itself, whereas three-to-five-day gap disclosures are post-hoc explanatory filings that expose latent insider conviction whose price reaction has not yet fully realized. The direction is counterintuitive: the standard event-study convention treats the co-disclosure day as the cleanest signal date. The finding has direct implications for future joint insider-plus-disclosure event studies, in which pooling same-day and gap filings will understate the informational content of the gap sub-cohort.

Beyond these two findings, the cluster mechanism itself is documented on a fully survivorship-free eleven-year panel. The aggregate five-trading-day cumulative abnormal return on the 32,098-event pool is +1.68% residualized against Fama-French five factors plus momentum ($|t| = 32.94$; the market-adjusted analogue against the SPDR S&P 500 exchange-traded fund is +1.44%, $|t| = 32.11$; both use winsorization at the first and ninety-ninth percentile computed on a past-events-only basis to prevent foresight in the winsorization cut). The mean is positive with a $|t|$ statistic above three in each of seven macro regimes covering 2015–2017 quantitative-easing tapering, 2018–2019 pre-COVID normal, 2020 COVID, 2021 post-COVID, 2022 inflation and Federal Reserve hiking, 2023–2024 artificial-intelligence boom, and 2025 recent. Under a cost sensitivity from 25 to 100 basis-point round-trip on the thirty-position long-only portfolio construction, the baseline (equal-weight) net Sharpe degrades from 1.77 to 0.51.

The rest of the paper is organized as follows. Section 2 maps the related insider-trading and disclosure-text literatures and positions the study. Section 3 describes the data assembly, including the CRSP `dsenames` and SEC EDGAR full-text search bridge that recovers the 23.4% delisted-issuer share of the panel. Section 4 develops the cluster detection rule, the classification prompt, the priority-fill portfolio construction, and the point-in-time entry convention. Section 5 reports the cluster-mechanism aggregate result, the regime consistency, the cost sensitivity, and the two new findings. Section 6 reports the robustness battery, including sector partition, size-tier partition with tier-specific transaction costs following Novy-Marx and Velikov (2016) and Frazzini et al. (2018), the winsorization sensitivity, and the cluster-window sensitivity. Section 7 discusses the economic interpretation. Section 8 catalogs limitations. Section 9 concludes.

2 Related Literature

2.1 The Insider-Trading Return Channel

The closest prior work on individual insider transactions is Cohen et al. (2012). On U.S. insider filings between 1996 and 2005, that paper documents a six-month cumulative abnormal return of approximately 5% on the opportunistic subset — transactions that deviate from a given insider's

historical trading routine — against a near-zero return on the routine subset. The mechanism is that opportunistic transactions carry a higher share of information-driven insider trades, while routine transactions are dominated by diversification and liquidity motives.

Alldredge and Blank (2019) extend the anchor to the cluster margin. On U.S. Form 4 filings between 1986 and 2014, restricting attention to events in which at least two insiders at the same firm trade in the same direction within a short peer window, the paper reports a monthly abnormal return of approximately 2.1% concentrated in the cluster subset. Cziraki and Gider (2021) study the dollar-profits margin of the same disclosure feed on a longer 1986–2013 panel, documenting that the aggregate dollar profit to insider trading has increased over time while the per-transaction abnormal return has compressed.

The earlier Jeng et al. (2003) documents the aggregate risk-adjusted return to a portfolio that mimics insider transactions on the 1975–1996 sample. The mechanism-focused literature on why insider transactions are informative — Cohen et al. (2013) on the cohort variation, Lakonishok and Lee (2001) on the aggregate-versus-firm-specific decomposition, and the earlier textbook treatments of insider trading and market efficiency — provides the theoretical scaffolding under which a text-based body-tone stratifier can be interpreted as a signal-quality axis.

2.2 Disclosure-Text Signals

The text-mining literature on SEC filings is now large. Loughran and McDonald (2011) introduce a domain-specific negative-word dictionary tuned to 10-K filings and demonstrate that generic sentiment dictionaries mis-classify accounting-domain vocabulary. Cassell et al. (2013), Johnston and Petacchi (2017), and Ryans (2021) study SEC comment-letter text and its association with subsequent restatement disclosures. Chen et al. (2024) employ large-language-model embeddings of generic firm-level news to construct cross-sectional expected-return signals. Lopez-Lira and Tang (2023) use a single large language model to score news headlines for sentiment and predict next-day equity returns.

None of these papers, to my knowledge, applies a language-model classifier to the 8-K or DEF 14A body content specifically as a stratifier for insider cluster events. The insider-trading literature has stratified cluster events by transactional metadata (identity, timing, volume) and by exogenous firm-level fundamentals, but not by the concurrent disclosure body tone. The intersection — Form 4 cluster events crossed with a large-language-model reading of the concurrent 8-K and DEF 14A body — is the empirical gap this paper addresses.

2.3 Standard Baselines

Winsorized standard errors at the first and ninety-ninth percentile of the pooled forward cumulative abnormal return follow the convention of the insider-trading return literature; the winsorization percentile is computed on a past-events-only basis to prevent the full-sample percentile from embedding a look-ahead into the winsorization cut. The delisting-return correction on the historical CRSP daily file follows the spirit of Shumway (1997). Kenneth French five-factor plus momentum data (Fama and French 2015; Carhart 1997) are used for the residualization procedure described in Section 3.4: per-firm factor loadings are estimated by ordinary least squares on the 252-trading-day pre-window strictly preceding the cluster signal date, and residuals on any given post-event trading day are computed against the factor returns realized on that same day, which are same-day market observables. Trading-cost tier assumptions in Section 6.1 follow the empirical size-decile bid-ask

spread plus proportional market-impact estimates of Novy-Marx and Velikov (2016) and Frazzini et al. (2018).

2.4 Position of This Study

The cluster mechanism studied here extends Alldredge and Blank (2019) from a two-day peer window to a rolling thirty-day window that additionally requires at least one officer among the buyers, with the individual-insider result of Cohen et al. (2012) as the theoretical anchor. What neither paper uses — and what, to my knowledge, no paper on the Form 4 cluster cohort uses — is the content of the disclosures the firm files around the cluster: prior stratifications of insider events rely on transactional metadata (insider identity, timing, trade size) and firm-level fundamentals. This study adds two components: a language-model classification of the concurrent 8-K and DEF 14A text, and a decomposition of the resulting signal by the gap between the disclosure date and the cluster's final trade. Both are dimensions on which the prior work is silent rather than in disagreement, so the results below should be read as an extension of the cluster literature, not a re-estimation of it.

3 Data

3.1 Universe

The base population is all Form 4 accessions filed with the U.S. Securities and Exchange Commission's EDGAR system between 2015-01-01 and 2025-12-31. Accessions are retrieved from EDGAR's quarterly full-index files (`form.idx`), then filtered to include only transactions whose `transactionCode` field on the Form 4 XML equals `P` (open-market or private purchase, per Section 16 of the Securities Exchange Act reporting codes) *and* whose `acquiredDisposedCode` equals `A` (an increase in beneficial ownership). Grant, award, and other Rule 16b-3 acquisitions (`transactionCode = A`), gifts (`transactionCode = G`), and dispositions (`acquiredDisposedCode = D`, or `transactionCode = S` for open-market sales) are excluded by construction.

3.2 Form 4 Accessions with Delisting-Return Coverage

The initial pull of the current SEC `company_tickers.json` snapshot returns 1,480,397 accessions for currently-listed United States common-stock issuers. A ticker-level cross-check against the CRSP `dsenames` (name-history) table identifies 3,797 historical ticker-central-index-key pairs that appear in `dsenames` between 2015 and 2026 but do not resolve to a current SEC ticker snapshot; for these historically-delisted issuers, the corresponding accessions are recovered by (i) resolving the ticker to a central index key via the SEC EDGAR full-text search API, (ii) fetching each central index key's historical submissions manifest via the EDGAR submissions API, and (iii) filtering to Form 4 accessions within the 2015–2025 window. The bridge pass adds 452,109 delisted-issuer accessions, bringing the panel to 1,932,506 accessions. The delisted-issuer share is 23.4%, consistent with the U.S. equity delisting rate documented by Shumway (1997).

3.3 CRSP Daily Stock File

Daily equity returns for the cluster universe are sourced from the CRSP Daily Stock File (DSF v2) via a local dump. The subset filtered to the cluster universe plus the SPDR S&P 500 exchange-traded fund (ticker SPY, used as the market proxy in the market-adjusted specification) spans 2014-01-01 through 2025-12-31 and contains approximately 6.66 million daily observations. The

universe filter retains share codes 10 and 11 (U.S. common stock) and exchange codes 1, 2, and 3 (NYSE, NYSE MKT/AMEX, Nasdaq). Delisting returns are retained where the CRSP delisting-return flag is populated; the corresponding forward-window observation is computed inclusive of the delisting return.

3.4 Kenneth French Factor Series

Daily Fama-French five-factor and momentum series (Fama and French 2015; Carhart 1997) are sourced from the Ken French data library. The series span 1963-01-01 through 2026-06-30, yielding 15{,}833 daily observations. The residualization procedure of Section 4 estimates per-firm factor loadings by ordinary least squares on the 252 trading days strictly preceding the cluster signal date, so that the loadings themselves are point-in-time and do not use post-event information; residual returns on any given post-event trading day are then computed against the factor returns realized on that same day, which are market observables at close of business.

3.5 Concurrent 8-K and DEF 14A Disclosure Text

For each detected cluster event, the concurrent 8-K and DEF 14A filings within a $\pm$5-trading-day window of the cluster terminal transaction date are fetched from EDGAR. The primary document is retrieved by following the EDGAR submission-index page; HTML-formatted documents are stripped to plain text using a standard HTML-to-text conversion. Of the 38,261 detected clusters, 14,141 (37.0%) have at least one 8-K or DEF 14A document successfully fetched within the $\pm$5-day window; of these, 12,527 (88.6% of fetched documents) return successfully parsed labels from all three vendors and are retained as the labeled sub-panel. The remaining 25,734 clusters have no concurrent disclosure (either no filing exists within the window, the primary document is unreachable, or all three vendors fail to return a parseable label) and enter the portfolio construction as no-disclosure events. After intersecting with the CRSP DSF v2 forward-window filter (Section 3.3), the portfolio-eligible pool is 35,161 events, comprising 11,671 labeled events (positive-outlook plus neutral) and 23,490 no-disclosure events. The pooled aggregate cumulative-abnormal-return analysis of Section 5.1 is estimated on a slightly narrower 32,098-event subset that additionally requires a valid 252-day factor-loading pre-window; the tone-lift analysis of Section 5.4 is estimated on the full 35,161-event portfolio-eligible pool.

3.6 Sample Selection Waterfall

Table 1: Sample selection waterfall.

Stage	N (2015–2025)
Form 4 accessions on EDGAR full-index	1,932,506
Delisted-issuer share (via `dsenames` + EDGAR bridge)	23.4%
Distinct issuers	$\sim$ 6,486
Detected multi-insider open-market buy clusters (30-day rolling)	38,261
Clusters with concurrent 8-K / DEF 14A body ($\pm$5 days)	14,141
Clusters with three-vendor parseable label	12,527
Portfolio-eligible pool (valid 5-day CRSP forward window)	35,161
of which labeled (positive or neutral)	11,671
of which no-disclosure	23,490
Aggregate-CAR pool (valid 252-day factor pre-window)	32,098

4 Methodology

4.1 Cluster Detection

A multi-insider open-market buy cluster is a firm-date at which at least two distinct reporting-owner central index keys and at least one flagged officer (from the Form 4 `isOfficer` field, corresponding to titles including Chief Executive Officer, Chief Financial Officer, Chief Operating Officer, and President) have each executed at least one open-market purchase transaction within a rolling thirty-calendar-day window ending at that date. Cluster events sharing the same issuer and the same terminal transaction date are collapsed.

The cluster signal date is defined as max(max transaction date, max filed date) across the constituent transactions of the cluster. This is the last calendar date on which the entire cluster is observable on EDGAR, and is the correct point-in-time entry-eligibility date: using the transaction date alone would embed a look-ahead because Form 4 filings have a two-business-day statutory reporting lag, so a market participant does not observe the constituent transaction of a cluster until at least the following business day.

The portfolio entry date is the first trading day strictly after the cluster signal date (a $t + 1$ open-to-close convention). The forward window is the subsequent five trading days.

4.2 Disclosure-Text Classification

For each cluster with a fetched concurrent 8-K or DEF 14A document, the disclosure text is passed to a zero-shot ensemble of three language models — OpenAI's GPT-4o-mini, Meta's Llama 3.3 70B Instruct, and DeepSeek's Chat v3.1, accessed via the OpenRouter API under the model identifiers `openai/gpt-4o-mini`, `meta-llama/llama-3.3-70b-instruct`, and `deepseek/deepseek-chat-v3.1`. The structured-output prompt returns one of two labels: *positive-outlook* if the text contains at least one of a pre-specified set of triggers spanning raised forward guidance, share-repurchase authorization, material contract wins or regulatory approvals, product or clinical/regulatory milestones, earnings surprises, explicit near-term operational confidence, and buyer-side accretive mergers and acquisitions; and *neutral* otherwise (including operational updates, insider-diversification announcements, and ambiguous forward guidance). Ambiguity favors the neutral label, consistent with the conservative-labeling convention in the disclosure-text literature. The full trigger enumeration and the verbatim prompt are reproduced in Appendix A.

Each vendor is queried independently at temperature zero with a response-token cap of one hundred (the schema output is short and 100 tokens is generous headroom). The vendor labels are aggregated to a majority vote (two of three); the sample also permits unanimity (all three agree) and any-positive (at least one vendor votes positive-outlook) aggregations, both of which are reported alongside the per-vendor sensitivity in Section 5.4 (Table 5).

A three-vendor cross-rater agreement is estimated on a stratified random sample of $n = 97$ disclosures drawn from the 2024-Q4 pilot cohort. The minimum pairwise Cohen's κ across the three vendors is 0.823, well above the 0.70 threshold for substantial agreement (Landis and Koch 1977).

4.3 Portfolio Construction

The portfolio is a long-only cross-section that maintains up to $N = 30$ concurrent positions in equal weight. Each position is opened at the first trading day strictly after the cluster signal date and

closed at the fifth trading day thereafter.

Two constructions are compared. In the baseline, all 38,261 cluster events are eligible for portfolio inclusion in the order of their entry date; if more than thirty entries fire on the same date, the first thirty are selected. The portfolio's daily return is the equal-weighted mean of the market-adjusted daily returns of the active positions. In the priority-fill construction, the same event pool is re-sorted so that within each entry date the positive-outlook events are selected first, then the neutral events, then the no-disclosure events, until the thirty slots are filled. On days when no positive-outlook event fires the portfolio falls back to neutral and no-disclosure events, so the portfolio is always full when at least thirty signals fire on the entry date. This construction addresses the natural sparse-signal problem of the 4.5% positive slice: a sub-portfolio restricted to positive-outlook events would only rarely fill the thirty-position portfolio and would exhibit inflated daily volatility from the resulting position sparsity.

4.4 Cost Model

Round-trip transaction cost is modeled as a portfolio-level annualized drag: cost drag $= (252/H) \times \mathrm{RT}_{\mathrm{bps}}/10{,}000$, where $H = 5$ is the holding-period length in trading days and $\mathrm{RT}_{\mathrm{bps}}$ is the round-trip cost in basis points. At the baseline 25-basis-point round-trip cost the annualized drag is 12.6%; sensitivity to 50, 75, and 100 basis points is reported in Section 5.

For the size-tier partition of Section 6.1, the round-trip cost is tier-specific: 25 basis points for large, 50 for mid, 80 for small, and 150 for micro, following Novy-Marx and Velikov (2016) Table 3 and Frazzini et al. (2018) Table 5 for the empirical size-decile bid-ask-spread-plus-proportional-market-impact scaling.

4.5 Statistical Inference

Cell-level t-statistics on the five-day cumulative abnormal return use the winsorized standard error of the mean, with the winsorization percentiles computed on a past-events-only basis so that the cut on any given event is a function only of events observed strictly earlier in the panel. Annualized Sharpe ratios use the standard convention of $\sqrt{252}$ scaling on the daily portfolio return series. Cost-adjusted net Sharpe ratios subtract the annualized round-trip cost drag $(252/H) \times \mathrm{RT}_{\mathrm{bps}}/10{,}000$ from the annualized mean before dividing by the annualized volatility.

5 Results

5.1 Cluster Mechanism Aggregate Result

Table 2 reports the aggregate five-trading-day cumulative abnormal return on the 32,098 cluster events with a valid five-day forward market-adjusted window.

Table 2: Aggregate five-trading-day cumulative abnormal return, 2015–2025 panel, $N = 32{,}098$ cluster events with valid market-adjusted forward window.

Specification	Mean	$\lvert t \rvert$	Sharpe (annualized)
Raw (no adjustment)	+1.71%	27.67	1.10
Winsorized 1%/99% (past-events-only)	+1.58%	32.16	1.27
Market-adjusted (SPY, winsorized)	+1.44%	32.11	1.27
Fama-French five-factor plus momentum residualized (winsorized)	+1.68%	32.94	1.31

The Fama-French residualized cumulative abnormal return is slightly larger than the market-adjusted specification, indicating that classical asset-pricing factor exposure absorbs approximately negative six percent of the mean; the cluster event returns are near-orthogonal to the standard five-factor plus momentum common-factor structure. This orthogonality is consistent with the individual-insider result of Cohen et al. (2012) that the informational content of insider transactions is largely factor-agnostic.

5.2 Regime Consistency

Table 3 reports the mean five-day market-adjusted cumulative abnormal return partitioned by seven macro regimes covering the 2015–2025 sample.

Table 3: Macro-regime sign consistency of the cluster-event cumulative abnormal return (market-adjusted, winsorized).

Regime	Period	N	Mean	$\lvert t\rvert$	Sharpe
Quantitative-easing tapering	2015–2017	8,276	+1.38%	17.37	1.36
Pre-COVID normal	2018–2019	5,787	+1.53%	12.61	1.18
COVID	2020	4,034	+0.79%	3.32	0.37
Post-COVID	2021	2,299	+1.98%	11.40	1.69
Inflation, Fed hiking	2022	3,496	+2.08%	11.59	1.39
Artificial-intelligence boom	2023–2024	5,532	+1.69%	11.36	1.08
Recent	2025 (partial)	2,674	+2.55%	12.33	1.69

All seven regimes deliver a positive mean with $|t|$ statistic above three, including the weakest cell (2020 COVID, $|t| = 3.32$). The mechanism is sign-consistent across a decade that includes quantitative-easing tapering, a market crash, an inflation-driven Federal Reserve hiking cycle, and the artificial-intelligence boom.

5.3 Cost Sensitivity

Table 4 reports the thirty-position long-only portfolio baseline net Sharpe ratio at four alternative round-trip transaction cost assumptions on the 35,161-event portfolio-eligible pool. The cost drag is applied at the portfolio level as cost drag $= (252/H) \times \mathrm{RT}_{\mathrm{bps}}/10{,}000$ where $H = 5$ is the holding-period length in trading days; at 25 basis-point round-trip cost the annualized drag is 12.6% and at 100 basis-point round-trip cost the annualized drag is 50.4%. Because the drag is a constant additive on annualized mean, the net Sharpe declines linearly with the round-trip cost at a slope equal to the cost drag divided by the portfolio's annualized volatility (approximately 30% on this sample).

Table 4: Cost sensitivity of the thirty-position long-only portfolio net Sharpe ratio, 2015–2025 panel, five-day holding period, 35,161-event portfolio-eligible pool.

Round-trip cost (bps)	Portfolio net Sharpe
25	1.77
50	1.35
75	0.93
100	0.51

At 100 basis-point round-trip cost the baseline portfolio retains a positive net Sharpe of approximately 0.51. The size-tier partition in Section 6.1 uses tier-specific round-trip cost assumptions calibrated to the empirical size-decile trading-cost estimates of Novy-Marx and Velikov (2016) and Frazzini et al. (2018) rather than the uniform sensitivity of Table 4.

5.4 Finding 1: A Vendor-Invariant Portfolio Lift from Disclosure Tone

Table 5 reports the priority-fill net Sharpe ratio under six alternative label sources. Of the 12,527 three-vendor-labeled events, 4.5% are positive-outlook and 95.5% neutral under majority-vote aggregation. Intersecting with the CRSP forward-window filter yields the 35,161-event portfolio-eligible pool of Section 3.6; each row of Table 5 sums to this same 35,161 across $N_{\text{POS}} + N_{\text{NEUT}} + N_{\text{ND}}$, and the row-to-row variation in the N columns reflects the label-cut of the aggregation rule (unanimity, for instance, tightens the positive label so a larger share of labeled events drops back into the no-disclosure priority bucket).

Table 5: Priority-fill net Sharpe under six alternative label sources, 2015–2025 panel, five-day holding period, 25 basis-point round-trip cost, 35,161-event portfolio-eligible pool. POS = positive-outlook, NEUT = neutral, ND = no concurrent disclosure. Baseline (equal-weight, no tone tilt, same pool) net Sharpe is 1.77.

Label source	N_{POS}	N_{NEUT}	N_{ND}	Gross Sharpe	Net Sharpe	Lift
GPT-4o-mini	977	10,694	23,490	+2.68	+2.29	+0.52
Llama 3.3 70B	461	11,209	23,491	+2.66	+2.27	+0.50
DeepSeek Chat v3.1	391	11,280	23,490	+2.67	+2.29	+0.51
Majority vote (two of three)	540	11,131	23,490	+2.67	+2.28	+0.51
Unanimity (all three agree)	303	10,684	24,174	+2.63	+2.24	+0.46
Any-positive (≥ 1 vendor)	986	10,685	23,490	+2.68	+2.29	+0.52

The Sharpe lift is +0.50 to +0.52 across the three single-vendor label sources and +0.46 to +0.52 across the three aggregation rules. Unanimity delivers a marginally smaller lift (+0.46) because the tighter cut reduces the positive pool to 303 events (from a majority-vote 540), which reduces the number of dates on which a positive event is available at the top of the fill order. Any-positive delivers a marginally larger lift (+0.52) because the looser cut enlarges the positive pool to 986. The magnitude of the lift is vendor-invariant, and a single-vendor deployment is defensible at approximately one-third the operational cost of the three-vendor ensemble with negligible Sharpe loss.

5.4.1 Finding 1b: Identifier-Masked Robustness Check

A vendor-invariant lift is a partial defense against the concern that a language model with training-corpus coverage extending into 2024 or later might recognize a specific filer-date combination and recall its subsequent five-day return, rather than reading the disclosure text zero-shot. Three independently trained vendors are unlikely to share identical training corpora, so a common lift across vendors argues against single-vendor idiosyncrasy; but the three vendors do plausibly share overlapping public-domain training data (SEC EDGAR historical filings, news wire archives), so a common vintage leakage is not excluded by the stack test alone.

To close this channel, I re-classify a sub-sample of the labeled events with the identifiable filer metadata redacted from the body text and compare labels against the unmasked baseline. The classification prompt template (Appendix A.1) is a single user-role message with only the `{text}`

placeholder — it does not carry any filer identifier as prompt scaffolding (no ticker, no central index key, no filing date in a header) — so the identifier channel exists exclusively in the body content itself, and masking is applied entirely there. Redacted fields (regex-based; full implementation in Appendix A.2 and `scripts/classify_bodies_masked.py`): the filer's central index key in padded and unpadded form; the filer's ticker symbol; the filer's company name and its common corporate-suffix variants (auto-detected from the standard 8-K header line "([Name]) (Exact name of registrant as specified in its charter)"); the concurrent-filing filed date and the cluster terminal date in ISO and human-readable formats; all four-digit years in the range 2010–2030; all `MonthName Day, Year` and `MonthName Year` strings; fiscal-year and quarter references; and structural identifiers (IRS Employer Identification Number, Securities Exchange Commission file number, phone number, ZIP code).

The three-vendor ensemble configuration and the majority-vote aggregation rule are unchanged; only the disclosure text passed as the `{text}` placeholder is masked. On a shared intersection of $n = 7{,}652$ events with three-vendor complete labels in both the masked and unmasked passes, the majority-vote label agreement is 99.33% and Cohen's kappa is 0.9240, which is in the "near-perfect agreement" band of the Landis-Koch scale. Per-vendor kappa is in a similar band: 0.8930 for GPT-4o-mini, 0.9298 for Llama 3.3 70B, 0.9237 for DeepSeek Chat v3.1. The positive-outlook share is stable to within 0.04 percentage points: 4.57% unmasked (350 events) versus 4.61% masked (353 events).

Table 6: Majority-vote confusion matrix, unmasked vs identifier-masked labels, $n = 7{,}652$ shared events with three-vendor complete labels in both passes.

	Masked = POS	Masked = NEUT
Unmasked = POS	326	24
Unmasked = NEUT	27	7,275

The unmasked-positive-to-masked-neutral flip rate is 24/350 = 6.9%; the unmasked-neutral-to-masked-positive flip rate is 27/7,302 = 0.4%. The net positive-share drift is three events, which is not economically meaningful in a 35,161-event portfolio-eligible pool: if the vintage-leakage hypothesis were the dominant driver of the positive label, then removing the identifiers should collapse the positive share toward the neutral baseline; instead, the share is essentially unchanged. Because the positive share is preserved, the priority-fill portfolio Sharpe lift of +0.51 (Table 5, majority-vote row) is preserved under masking to within the flip-rate error bar.

The stronger channel that this test does not close is a hypothetical vintage embedding of body-text-to-outcome mappings — the possibility that the model learned during pre-training that disclosures with certain phrasings predict certain forward returns, even without the identifier metadata. Finding 1c below addresses this residual channel with a pre-training-cutoff cross-check.

5.4.2 Finding 1c: Pre-2024 Vintage Cross-Check

To bound the text-to-outcome-mapping channel, I re-classify the labeled panel with a model whose training data predate the outcome period: such a model cannot have seen forward-return outcomes for the 2022–2025 sub-period of the panel by construction. After testing older Anthropic, Meta, Google, and Mistral candidates on OpenRouter in July 2026 (all returned HTTP errors following provider-side retirements), the only pre-2024 candidate that returns successfully is the OpenRouter model identifier `openai/gpt-3.5-turbo-0613`. Direct verification of the response metadata on 2026-07-22 returned HTTP 200 with `model = openai/gpt-3.5-turbo-0613`, `provider =`

`"Azure"`, a populated `system_fingerprint` field, and a usage cost of 6.9×10^{-5} per short call — consistent with the response coming from a genuine backend rather than a proxy stub. I cannot independently verify from the client side whether the served snapshot is actually the 0613 checkpoint or an alias that OpenRouter labels 0613; OpenAI's public retirement schedule shows the 0613 checkpoints decommissioned from OpenAI's direct API in September 2024, and Microsoft's public retirement schedule shows the 0613-generation Azure OpenAI deployments retired in mid-2025, so the served backend is plausibly a later snapshot that OpenRouter routes under the 0613 slug. This does not affect the pre-2024 argument: the entire `gpt-3.5-turbo` model family (0301, 0613, 0125) has training cutoff September 2021 per OpenAI documentation, so any snapshot in the family satisfies the pre-2024 requirement. The raw response headers and one generation ID are dumped in the replication package for verification.

Disclosure text is unmasked for this test; the prompt template is unchanged from Appendix A.1. The re-classification run halted at 7,656 successfully parsed events (61% of the 12,527 labeled pool) when the language-model budget was exhausted. Positive-labeled events were classified first: the 1,060-event pilot pass had already classified all 560 current-vintage positive events plus 500 random current-vintage neutral events; the full-panel resumption picked up those 1,060 from the on-disk cache without additional API calls and then processed the remaining events in random order (seed 42) until budget exhaustion. Consequently the sub-panel positive class is complete (all 560 current-vintage positive events are covered) while the neutral class is a random $\sim 59\%$ sample of the neutral pool. The oversampled-positive design is the intended one for a portfolio-fill comparison, where positive-class coverage is the binding constraint on the priority-fill construction.

Table 7: Pre-2024 gpt-3.5-turbo-0613 versus current-vintage 3-vendor majority (unmasked), $n = 7{,}656$ sub-panel (61% of the 12,527-event labeled pool).

Sub-cohort	n	Agreement	Cohen's κ
Full sub-panel	7,656	97.35%	0.7828
pre-cutoff (2015–2021)	4,778	97.63%	0.8061
post-cutoff (2022+)	2,878	96.87%	0.7444

Table 8: Confusion matrix, pre-2024 gpt-3.5-turbo-0613 versus current-vintage 3-vendor majority (unmasked), $n = 7{,}656$.

	Current = POS	Current = NEUT
pre-2024 = POS	398	41
pre-2024 = NEUT	162	7,055

The pre-2024 model calls 5.73% (439) of the sub-panel positive against the current majority's 7.31% (560). Both shares are elevated relative to the panel-wide positive rate of 4.5% because the sub-panel oversamples positive events by construction (see above); the like-for-like comparison is that pre-2024 is more conservative on the positive class within the sub-panel. Of the 560 current-majority positive events, pre-2024 confirms 398 (71.1%) and demotes 162 (28.9%) to neutral; the 71.1% preservation figure is an exact population statistic rather than a sample estimate, because the carry-over guarantees that every current-majority positive event in the labeled pool is covered by a pre-2024 label. Of the 7,096 current-majority neutral events in the sub-panel (a random $\sim 59\%$ draw from the panel-wide neutral pool), pre-2024 flips only 41 (0.6%) to positive. Negative-class

consensus is 99.4%. Year-stratified: the pre-cutoff sub-cohort (2015–2021) yields $\kappa = 0.8061$ and the post-cutoff sub-cohort (2022+) yields $\kappa = 0.7444$, a drop of 0.06 that is directionally consistent with a modest vintage-leakage effect on the recent portfolio period.

Table 9: Portfolio priority-fill net Sharpe under current-vintage versus pre-2024 labels, 2015–2025 panel, five-day holding period, 25 basis-point round-trip cost, 35,161-event portfolio-eligible pool. Both constructions use identical no-disclosure events. All 540 current-vintage positive events (post-CRSP-filter) are present in the pre-2024 sub-panel, so the positive-priority slot is populated from the same set in both constructions; current-vintage-labeled events without a pre-2024 label (approximately 4,871, all current-vintage neutral) fall to the last-priority no-disclosure bucket in the pre-2024 construction, which lowers the neutral-priority pool relative to the current-vintage construction. Sharpe figures are displayed to three decimals so that the reported lifts and their difference reproduce arithmetically from the table itself.

Construction	N_{POS}	Gross Sharpe	Net Sharpe	Lift vs baseline
Baseline (equal-weight, no tone tilt)	—	+2.188	+1.775	—
Current 3-vendor majority priority-fill	540	+2.672	+2.284	+0.509
Pre-2024 gpt-3.5-turbo-0613 priority-fill	418	+2.517	+2.116	+0.341

The pre-2024 priority-fill construction, which strictly cannot have used forward-outcome information for the 2022–2025 sub-period, delivers a net Sharpe of +2.116 against the current-vintage construction's +2.284. The lift attenuation, computed on the displayed table values, is $(0.509 - 0.341)/0.509 = 33.0\%$: about one-third of the current-vintage lift is potentially attributable to vintage-embedded text-to-outcome signal, and the remainder (+0.341 net lift) survives a strictly pre-outcome-data model. This is the leakage-corrected lower bound on the Finding 1 lift and the appropriate quantity to cite where a vintage-robust figure is required. The upper bound remains the current-vintage +0.509, which also captures the pre-2024 model's weaker fine-grained trigger recall (the gpt-3.5-turbo generation is materially smaller and older than the current ensemble members, so some of the 29% positive-demotion is genuine capability differential rather than vintage-driven label inflation). Either way, the priority-fill lift remains economically material.

The pre-2024 priority-fill construction is not perfectly comparable to the current-vintage construction on the neutral-priority slot. Across the 2,766 trading days on which either construction fills at least one position, the pre-2024 construction draws on the no-disclosure bucket on 2,583 days (93.4%), with 42,536 of 75,574 total slot-days (56.3%) filled from that bucket; the corresponding figures for the current-vintage construction are 2,152 days (77.8%) and 27,893 slot-days (36.9%). The difference — 19.4 percentage points more no-disclosure slot-days in the pre-2024 fill — comes primarily from the 4,871 current-vintage-labeled events without a pre-2024 label falling into the no-disclosure bucket. The direction of the bias this asymmetry introduces rests on an empirical premise about the relative mean CAR of the buckets: on the sample, current-vintage-labeled neutral events carry a mean five-day market-adjusted CAR of +1.81%, the labeled-neutral events that fell into the pre-2024 no-disclosure bucket carry +1.69%, and the true no-disclosure events carry +1.32%. Moving events from the neutral-priority slot (+1.81%) into the no-disclosure slot (where their +1.69% mixes with the true no-disclosure +1.32%) lowers per-slot expected CAR by approximately 12 basis points on the marginal event. The bias therefore runs toward over-attenuation of the pre-2024 lift: giving the pre-2024 construction those missing neutral-labeled events would raise its neutral-priority saturation and lift its Sharpe. The +0.341 pre-2024 lift and 33.0% attenuation should be read as, respectively, a conservative lower bound on the vintage-robust Sharpe lift and a conservative upper bound on the vintage-inflated component of Finding 1. The magnitude of the bias is bounded by the 12-basis-point per-event CAR gap; a sharper point estimate would require

completing the remaining pre-2024 classification calls.

5.5 Finding 2: Disclosure-Timing Gap

Table 10 partitions the labeled events by the calendar gap between the concurrent disclosure's filed date and the cluster terminal transaction date.

Table 10: Priority-fill net Sharpe by disclosure-timing gap tier, 2015–2025 panel, five-day holding period, 25 basis-point round-trip cost. POS = positive-outlook, NEUT = neutral (as in Table 5).

Gap tier	N_{events}	N_{POS}	N_{NEUT}	Gross Sharpe	Net Sharpe
Same-day (0 days)	1,285	48	1,237	+1.04	+0.64
Close (1–2 days)	4,112	215	3,897	+1.87	+1.63
Wide (3–5 days)	6,274	277	5,997	+2.08	+1.86

The same-day cohort — clusters whose concurrent 8-K is filed on the same calendar day as the cluster terminal transaction — yields a priority-fill net Sharpe of +0.64, materially below the close-tier +1.63 and the wide-tier +1.86. The direction is counterintuitive: the standard event-study convention treats immediate co-disclosure as the cleanest signal date.

The natural reading is that the same-day 8-K is typically a reactive filing — an earnings release, a material-contract announcement, or a management-transition disclosure that has already been absorbed into the closing price on the disclosure day itself. By the time the market participant observes the cluster at $t+1$ open under the point-in-time entry convention, the news content has already been priced. The three-to-five-day gap filings, by contrast, are typically post-hoc explanatory filings in which the cluster of insider purchases predates the disclosure and the disclosure confirms a latent operational or strategic conviction. In this case the market has not yet fully absorbed the joint signal at the time the participant observes the cluster on EDGAR, and the forward window captures the price adjustment.

The deployment implication is that the strategy should overweight three-to-five-day gap disclosures relative to same-day disclosures. The research implication is that future joint insider-plus-disclosure event studies should not pool the two gap sub-cohorts.

6 Robustness

6.1 Size-Tier Partition with Tier-Specific Transaction Costs

Table 11 reports the priority-fill net Sharpe by dollar-volume tier at $t-1$ (twenty-trading-day trailing average of price times volume). The dollar-volume proxy is available for 24,560 of the 35,161 portfolio-eligible events; the residual 10,601 events without a computable trailing dollar-volume proxy are dropped from the tier partition. Round-trip cost by tier is calibrated to the empirical size-decile bid-ask spread plus proportional market-impact estimates of Novy-Marx and Velikov (2016) (Table 3) and Frazzini et al. (2018) (Table 5): the large tier's 25 basis-points corresponds to a mid-decile institutional execution cost, the mid tier's 50 to a below-median cost, the small tier's 80 to the small-decile average, and the micro tier's 150 to the smallest-decile tail cost.

Table 11: Priority-fill net Sharpe by dollar-volume tier at tier-specific round-trip transaction cost, 24,560-event dollar-volume-tier-partitioned subpool.

Size tier	Floor (dollar-volume, per day)	N_{events}	RT cost (bps)	Gross Sharpe	Net Sharpe
Large	$\geq$ \$50M	1,977	25	+0.79	+0.36
Mid	\$5M to \$50M	5,209	50	+1.35	+1.06
Small	\$1M to \$5M	5,466	80	+2.05	+1.86
Micro	$<$ \$1M	11,908	150	+0.91	+0.51

The small-capitalization tier delivers the strongest performance at +1.86 net Sharpe even after the 80 basis-point round-trip cost. The large-capitalization tier delivers a modest +0.36 net Sharpe, consistent with rapid information absorption in the efficient large-cap universe. The micro-tier gross Sharpe is comparable to large-cap but the 150 basis-point round-trip cost erodes it substantially. The pattern is consistent with the small-cap concentration of anomaly returns documented by Stambaugh et al. (2012) and the limits-of-arbitrage frame of Shleifer and Vishny (1997).

The tier distribution has an immediate capacity implication that the paper does not sidestep. Micro-cap events ($<$ \$1 million average daily dollar volume) account for 48% of the tier-partitioned pool, small-cap events (\$1–\$5 million) another 22%; the two combined — the tiers in which the priority-fill net Sharpe is highest after tier-specific cost — comprise 70% of the event count but a small share of the aggregate deployable notional. On a single-day participation limit of 50 basis points of the twenty-day trailing average dollar volume (a conservative single-trade share of a name's own liquidity), the small tier tolerates roughly \$5,000–\$25,000 per position on the median event and the micro tier below \$5,000; at thirty concurrent positions this puts the strategy's deployable notional at approximately \$150,000–\$750,000 on the small tier and one order of magnitude smaller on the micro tier. The strategy is tractable at retail scale; an institutional deployment would need to concentrate on the mid- and large-cap tiers, where the reported net Sharpe is materially lower.

6.2 Sector Partition

Table 12 reports the priority-fill net Sharpe by Fama-French forty-eight industry classification. The industry mapping follows the Ken French Standard Industrial Classification-range table applied to each firm's primary Standard Industrial Classification code from the CRSP `dsenames` name-history file.

Table 12: Per-industry priority-fill net Sharpe (25 basis-point round-trip cost), Fama-French 48 industries with sample size ≥ 500.

Industry	N_{events}	Gross Sharpe	Net Sharpe
Medical equipment	533	+2.21	+2.04
Other	6,660	+2.30	+2.11
Household products	658	+1.97	+1.75
Telecommunications	511	+1.91	+1.69
Oil	740	+1.56	+1.40
Drugs	1,992	+1.41	+1.24
Business services	1,913	+1.39	+1.18
Retail	509	+1.20	+0.99
Chips	528	+1.15	+0.98
Financials	1,386	+1.10	+0.83
Insurance	929	+1.31	+0.97
Utilities	734	+0.56	+0.13
Banks	5,032	+0.47	−0.07

The sector Sharpe dispersion is wide, from Banks at −0.07 to Medical equipment at +2.04. Banks and Utilities carry weak or negative net Sharpe, consistent with the interpretation that their 8-K and DEF 14A body content is dominated by regulatory, capital, and rate-adjustment items that carry limited forward-return content. The distribution has direct deployment implications: a sector-neutrality overlay would substantially reduce the aggregate signal by forcing capital into low-signal sectors; the natural deployment weights the concentration toward high-signal sectors.

6.3 Winsorization Sensitivity

Table 13 reports the sensitivity of the priority-fill net Sharpe to the choice of winsorization percentile on the pooled forward cumulative abnormal return. The winsorization percentile is computed on a past-events-only basis to prevent foresight in the winsorization cut.

Table 13: Priority-fill net Sharpe sensitivity to the winsorization percentile.

Rule	Priority-fill net Sharpe
No winsorization	+2.02
Winsor 0.5%/99.5%	+2.24
Winsor 1%/99% (base)	+2.28
Winsor 2%/98%	+2.26
Hard truncation at ±50%	+1.87

The Sharpe is stable across the 0.5–2% winsorization band. Removing the winsorization entirely drops the net Sharpe by 0.26 (from +2.28 to +2.02); hard truncation at plus-or-minus fifty percent drops it by 0.41 (from +2.28 to +1.87), which implies that observations with absolute five-day return above 50% account for roughly one-fifth of the realized Sharpe. The strategy is not exclusively tail-driven, but the tail contribution is material and a deployment estimate should not assume that the pooled Sharpe survives a policy that excludes extreme observations.

6.4 Cluster-Window Sensitivity

Table 14 reports the sensitivity of the aggregate mean five-day cumulative abnormal return and the corresponding priority-fill net Sharpe to the choice of cluster window (thirty, forty-five, and sixty calendar days).

Table 14: Cluster-window sensitivity of the aggregate cumulative abnormal return.

Cluster window	N	Mean CAR	Priority-fill net Sharpe
30 days (base)	32,098	+1.68%	+2.28
45 days	34,770	+1.94%	+2.34
60 days	37,013	+2.07%	+2.31

The Sharpe is stable within a narrow band across the three cluster window definitions. The choice of thirty days is retained for the primary specification because it is broadly consistent with the multi-week rolling-window convention in Alldredge and Blank (2019).

7 Discussion

The cluster mechanism itself is a well-documented informational anchor (Cohen et al. 2012; Alldredge and Blank 2019): at the point-in-time signal date, the market observes multiple distinct insiders and at least one officer having transacted an open-market purchase within a rolling thirty-day window, which concentrates the informational content of the disclosure feed. On this anchor, the concurrent 8-K and DEF 14A content contributes an additional stratification axis: the language-model reading identifies the events at which the disclosure text is concordant with a specific positive-outlook trigger. The priority-fill portfolio then implements the joint signal as a two-stage screen: at each entry date, positive-labeled events are inserted first into the thirty-position portfolio, then the remaining slots are filled with neutral-labeled and no-disclosure events. The construction preserves the baseline cluster-mechanism informational content while opportunistically overweighting the additional textual information carried by the concurrent disclosure.

The size-tier partition (Section 6.1) and the sector partition (Section 6.2) are consistent with two well-established structural readings. The small-capitalization concentration of the tone lift is consistent with the limits-of-arbitrage frame of Shleifer and Vishny (1997) and the small-cap anomaly concentration documented by Stambaugh et al. (2012): the small-capitalization universe is where sell-side coverage is thin, institutional attention to the concurrent disclosure is bounded, and a mildly-informative textual signal is most likely to survive as a mispricing. The sector partition reveals that the positive-outlook label is not uniformly distributed across industries: sectors whose 8-K and DEF 14A content is dominated by concrete disclosures (medical equipment, telecommunications, household products) carry both a high baseline Sharpe and a disproportionately high positive-label share, while sectors dominated by regulatory and rate-adjustment disclosures (banks, utilities) carry weak or negative baseline Sharpe. The two-stage priority-fill construction inherits this joint distribution: aggregate deployment tilts toward the sectors whose disclosure feed is most legible under the trigger taxonomy.

The disclosure-timing gap finding (Finding 2) is the empirically most novel result. The natural expectation is that same-day co-disclosure carries the cleanest and most timely signal. The empirical result is the opposite: same-day co-disclosure yields +0.64 priority-fill net Sharpe, whereas three-

to-five-day gap disclosure yields +1.86. The interpretation is that the same-day 8-K is typically a reactive filing whose news content has already been absorbed into the closing price on the disclosure day itself; by the time the market participant observes the cluster at the point-in-time entry date, the news has been priced. The three-to-five-day gap disclosure is a post-hoc explanatory filing whose alignment with the pre-existing insider cluster confirms a latent operational or strategic conviction that has not yet been priced. The finding has direct implications for event-study designs on insider-plus-disclosure joint events: pooling same-day and gap filings understates the informational content of the gap sub-cohort.

The aggregate tone lift of +0.51 Sharpe is economically material at the panel level, but the reported figures are position-level and not deployment-adjusted for the overlapping-position diversification that a monthly-entry-with-five-day-holding portfolio would realize. A deployment-realistic per-capital construction would maintain approximately six overlapping positions per dollar of capital at any point in time, and the corresponding per-capital annualized Sharpe would be lower than the position-level figure. Reporting the per-capital construction alongside the position-level figure is a natural follow-up.

8 Limitations

I catalog limitations directly.

Language-model vintage look-ahead — primary channel closed, secondary channel bounded and quantified. All three vendors in the ensemble have training-corpus coverage that extends into 2024 or later. Two channels by which vintage leakage could inflate Finding 1 are addressed. The primary channel is filer-date recall: the model recognizes a specific filer plus filing date and recalls the subsequent forward return. Finding 1b (Section 5.4) closes this channel: a re-extraction with the filer's central index key, ticker, company name, filed date, and all explicit date strings redacted from the text preserves the three-vendor majority-vote label at 99.33% agreement (Cohen's kappa = 0.9240) on a 7,652-event shared sample, and the positive share drifts by only three events (4.57% $\rightarrow$ 4.61%). If filer-date recall were the operative mechanism, removing those identifiers should collapse the positive share back toward neutral; it does not. The secondary channel is a text-to-outcome mapping embedded in the pre-training corpus: the model has learned during pre-training that disclosures with certain phrasings predict certain forward returns even without identifier metadata. Finding 1c (Section 5.4) bounds and quantifies this channel with a pre-2024-training-cutoff cross-check on a 7,656-event sub-panel (61% of the labeled pool). The pre-2024 model confirms 71% of the current-majority positive labels and 99.4% of the current-majority neutral labels; year-stratified kappa is 0.81 on the pre-cutoff (2015–2021) sub-cohort and 0.74 on the post-cutoff (2022+) sub-cohort. Substituting the pre-2024 labels into the priority-fill portfolio delivers a net Sharpe of +2.116 versus the current-vintage +2.284, an attenuation of 33.0%: the vintage-robust lower bound on the Finding 1 lift is +0.341 Sharpe, and the current-vintage +0.509 should be read as an upper bound that also captures the pre-2024 model's smaller-and-older capability differential. The comparability asymmetry documented in Section 5.4 (the pre-2024 fill draws 56.3% of slot-days from the no-disclosure bucket versus 36.9% for the current-vintage fill) biases the attenuation figure upward, so 33.0% is a conservative upper bound on the vintage-inflated component.

The concurrent-disclosure fetch is bounded to ± 5 trading days of the cluster terminal date. An extension to ± 10 days is possible but requires additional body-fetch and language-model classification budget. The sensitivity on the $\pm 2/\pm 3/\pm 5$ cutoff shows a monotonic improvement

in the priority-fill net Sharpe as the window widens (from +1.96 at ±2 days to +2.28 at ±5 days), suggesting that a wider window may deliver a further lift, but the lift may also plateau as increasingly-loose disclosure matches accumulate more noise.

Delisting-return coverage. The CRSP daily stock file with delisting-return flag covers the majority of the panel. The 23.4% delisted-issuer share is recovered via the CRSP `dsenames` and SEC EDGAR full-text-search bridge documented in Section 3.2, but the bridge may under-recover a small fraction of the tail. The direction of the residual bias is toward under-including the most-delisted-prone firms, which is a conservative direction for the long-only cluster-plus-body-tone strategy documented here.

Portfolio construction is stylized. The thirty-position long-only portfolio at equal weight is a first-cut construction. A realistic deployment would layer per-position position-sizing (weighted by cluster strength; a per-body confidence field could be added to the classification schema to support this refinement), sector-neutrality overlay, and full-cost bid-ask spread plus linear impact per name. The tier-specific round-trip cost of 25/50/80/150 basis points calibrated to Novy-Marx and Velikov (2016) and Frazzini et al. (2018) is an approximation to the tier-specific cost scaling. The reported net Sharpe is therefore a stylized baseline that a realistic deployment would refine both upward (finer signal weighting) and downward (heavier realistic costs, sector-neutrality dilution).

Capacity is small-cap-bounded. As discussed in Section 6.1, 48% of the size-tier partitioned pool falls in the micro-cap tier (below one million dollars per day average dollar volume) and another 22% in the small tier. These two tiers carry the highest priority-fill net Sharpe after tier-specific cost, but the aggregate deployable notional is limited: an institutional deployment at large-cap participation would need to accept the substantially lower large-cap net Sharpe of +0.36.

Tail-observation contribution is material. As reported in Section 6.3, hard truncation of forward returns at ±50% drops the priority-fill net Sharpe from +2.28 to +1.87, so extreme observations account for roughly one-fifth of the realized figure. The pooled Sharpe is not exclusively tail-driven, but a realistic deployment estimate should not assume the pooled figure survives a policy that excludes extreme observations.

No live paper-trading pilot is reported here. The reported Sharpe figures are historical backtest results with the point-in-time entry convention. A live thirty-day paper-trading pilot on a SEC EDGAR real-time feed against the same priority-fill construction is a natural follow-up.

The tone lift is estimated on the 11,671 **labeled events within the** 35,161**-event portfolio-eligible pool.** The residual 23,490 events without a concurrent 8-K or DEF 14A filing are treated as last-priority slots in the priority-fill construction. If the tone signal is highly concentrated in the events with a concurrent disclosure, a hypothetical wider-net fetch that shifted no-disclosure events into the labeled cohort may or may not preserve the lift magnitude.

9 Conclusion

This paper documents that the tone of the disclosures a firm files around a multi-insider buy cluster carries portfolio-relevant information beyond the cluster itself. A zero-shot language-model classification of the concurrent 8-K and DEF 14A text, used to order the fill of a thirty-position long-only portfolio, raises the five-day net Sharpe ratio from 1.77 to 2.28 on a survivorship-free 2015–2025 panel. The lift is invariant to the choice of language model, survives redaction of all filer identifiers from the classified text, and retains two-thirds of its magnitude under a model

whose training data predate the outcome period — a vintage-robust lower bound of +0.34 Sharpe. Separately, the timing of the concurrent disclosure matters in a direction the standard event-study convention does not anticipate: disclosures arriving three to five days after the cluster's final trade carry nearly three times the net Sharpe of same-day co-disclosures, consistent with same-day filings being reactive news that is already priced and delayed filings exposing insider conviction not yet reflected in prices. The underlying cluster mechanism replicates the prior literature's magnitudes on a decade of fresh data and is sign-consistent across seven macro regimes.

Three directions for follow-up are immediate. First, the pre-2024 vintage cross-check halted at 61% panel coverage; completing the remaining events would tighten the 33.0% lift-attenuation estimate. Second, the concurrent-disclosure fetch window is currently ± 5 trading days, and the cutoff sensitivity suggests a further lift may be available at a wider window. Third, no live paper-trading results are reported here; a live pilot on the SEC EDGAR real-time feed against the same priority-fill construction is the natural next step.

The full pipeline, including the classification prompt, the three-vendor ensemble output, the cluster detection rule, the priority-fill portfolio construction, and the historical CRSP-with-delisting-return bridge, is publicly released for replication and extension.

10 References

Alldredge, Dallin M., and Brian Blank. 2019. "Do Insiders Cluster Trades with Colleagues? Evidence from Daily Insider Trading." *Journal of Financial Research* 42 (2): 331–60.

Carhart, Mark M. 1997. "On Persistence in Mutual Fund Performance." *Journal of Finance* 52 (1): 57–82.

Cassell, Cory A., Lauren M. Dreher, and Linda A. Myers. 2013. "Reviewing the SEC's Review Process: 10-K Comment Letters and the Cost of Remediation." *The Accounting Review* 88 (6): 1875–908.

Chen, Yifei, Bryan T. Kelly, and Dacheng Xiu. 2024. "Expected Returns and Large Language Models." *SSRN Working Paper*.

Cohen, Lauren, Karl B. Diether, and Christopher J. Malloy. 2013. "Misvaluing Innovation." *Review of Financial Studies* 26 (3): 635–66.

Cohen, Lauren, Christopher Malloy, and Lukasz Pomorski. 2012. "Decoding Inside Information." *Journal of Finance* 67 (3): 1009–43.

Cziraki, Peter, and Jasmin Gider. 2021. "The Dollar Profits to Insider Trading." *Review of Finance* 25 (5): 1547–80.

Fama, Eugene F., and Kenneth R. French. 2015. "A Five-Factor Asset Pricing Model." *Journal of Financial Economics* 116 (1): 1–22.

Frazzini, Andrea, Ronen Israel, and Tobias J. Moskowitz. 2018. "Trading Costs." *SSRN Working Paper*.

Jeng, Leslie A., Andrew Metrick, and Richard Zeckhauser. 2003. "Estimating the Returns to Insider Trading: A Performance-Evaluation Perspective." *Review of Economics and Statistics* 85 (2): 453–71.

Johnston, Rick, and Reining Petacchi. 2017. "Regulatory Oversight of Financial Reporting: Securities and Exchange Commission Comment Letters." *Contemporary Accounting Research* 34 (2): 1128–55.

Lakonishok, Josef, and Inmoo Lee. 2001. "Are Insider Trades Informative?" *Review of Financial Studies* 14 (1): 79–111.

Landis, J. Richard, and Gary G. Koch. 1977. "The Measurement of Observer Agreement for Categorical Data." *Biometrics* 33 (1): 159–74.

Lopez-Lira, Alejandro, and Yuehua Tang. 2023. "Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models." *SSRN Working Paper*.

Loughran, Tim, and Bill McDonald. 2011. "When Is a Liability Not a Liability? Textual Analysis,

Dictionaries, and 10-Ks." *Journal of Finance* 66 (1): 35–65.

Novy-Marx, Robert, and Mihail Velikov. 2016. "A Taxonomy of Anomalies and Their Trading Costs." *Review of Financial Studies* 29 (1): 104–47.

Ryans, James P. 2021. "Textual Classification of SEC Comment Letters." *Review of Accounting Studies* 26 (1): 37–80.

Shleifer, Andrei, and Robert W. Vishny. 1997. "The Limits of Arbitrage." *Journal of Finance* 52 (1): 35–55.

Shumway, Tyler. 1997. "The Delisting Bias in CRSP Data." *Journal of Finance* 52 (1): 327–40.

Stambaugh, Robert F., Jianfeng Yu, and Yu Yuan. 2012. "The Short of It: Investor Sentiment and Anomalies." *Journal of Financial Economics* 104 (2): 288–302.

11 Appendix A: Classification Prompt and Masking Function

11.1 A.1 Prompt template

The prompt template is a single user-role message reproduced verbatim from the replication package source (`src/signals/llm_intent.py`). The template contains no filer identifier fields (no ticker, no central index key, no filing date) in either a system-role preamble or the user-message header: the only variable interpolated into the template is the disclosure text under the `{text}` placeholder. Consequently the identifier-mask robustness check of Finding 1b is applied entirely to the disclosure text itself, not to prompt scaffolding. The label strings `POSITIVE-outlook` and `neutral-diversification` below are the literal schema values returned by the models; the paper refers to them as the positive-outlook and neutral classes.

```
You are classifying a corporate disclosure (8-K or DEF 14A body)
filed within +/-5 days of a cluster of open-market insider purchases
by multiple company officers.

Classify into exactly ONE label.

Assign "POSITIVE-outlook" ONLY if the body contains at least ONE of
these SPECIFIC triggers:
  (T1) RAISED forward guidance (revenue, EPS, or unit volume
       increase over prior guidance)
  (T2) NEW share repurchase authorization or upsized existing buyback
  (T3) NEW material commercial contract win, customer announcement,
       or regulatory approval
  (T4) NEW product launch or clinical/regulatory milestone hit
  (T5) POSITIVE preliminary/actual earnings surprise vs consensus
  (T6) Explicit management confidence statement citing near-term
       operational improvement
  (T7) M&A announcement where the disclosing company is BUYER at
       accretive terms

Otherwise assign "neutral-diversification". Neutral includes
(non-exhaustive):
  - Item 5.02/5.03/5.07/8.01 routine administrative 8-Ks
    (director changes, bylaw changes, proxy results)
  - Standard proxy (DEF 14A) content without new forward-looking
    positive content
  - Earnings reports that MEET or MISS consensus (no positive surprise)
  - Guidance MAINTAINED at prior range (no raise)
  - Debt/credit facility amendments with neutral terms
  - Cost restructuring / layoff / write-down announcements
  - Regulatory investigations, litigation updates
  - Insider selling / gifting / 10b5-1 plan adoptions
  - Boilerplate risk-factor updates

BE CONSERVATIVE. If the body describes routine business without one
of triggers T1-T7, assign "neutral-diversification". Ambiguity
```

```
favors "neutral-diversification".

Respond with ONLY a JSON object:
{"body_intent": "POSITIVE-outlook"} or
{"body_intent": "neutral-diversification"}

Disclosure text (may be truncated):
---
{text}
---
```

Each vendor is queried independently at temperature zero with a token cap of one hundred. The three-vendor cross-rater agreement (minimum pairwise Cohen's κ) on the $n = 97$ pilot sample is $\kappa = 0.823$. The maximum body-text length passed under `{text}` is 40,000 characters; longer bodies are truncated at that limit.

11.2 A.2 Identifier-masking function (Finding 1b)

The identifier-masking function operates on the body text before it is inserted into the prompt template. Regex-based redaction covers seven categories: the filer's central index key (padded and unpadded forms), the filer's ticker symbol, the filer's company name and its common corporate-suffix variants (auto-detected from the standard 8-K header line "([Name]) (Exact name of registrant as specified in its charter)"), the concurrent-filing filed date and the cluster terminal date in ISO and human-readable formats, all four-digit years in the range 2010–2030, all `MonthName Day, Year` and `MonthName Year` strings, fiscal-year and quarter references, and structural identifiers (IRS Employer Identification Number `##-#######`, Securities Exchange Commission file number `###-####-#`, phone number `(###) ###-####`, ZIP code `#####[-####]`). The full implementation is in `scripts/classify_bodies_masked.py`.

12 Appendix B: Replication

The full pipeline runs from the project root. Each step is idempotent on its respective output file.

```
python scripts/fetch_form4_index.py \
    --start-year 2015 --end-year 2025 \
    --output data/form4_index.jsonl

python scripts/fetch_form4_accessions.py

python scripts/bridge_delisted_via_dsenames.py \
    --output data/edgar-form4-parsed-delisted/

python scripts/detect_clusters.py \
    --window-days 30 --min-insiders 2 --require-officer

python scripts/fetch_concurrent_bodies.py \
    --window-days 5

python scripts/classify_bodies_llm.py \
    --vendors openai/gpt-4o-mini,meta-llama/llama-3.3-70b-instruct,deepseek/deepseek-chat-v3.1

# Identifier-masked re-classification (Finding 1b)
python scripts/classify_bodies_masked.py

# Pre-2024 vintage cross-check (Finding 1c)
python scripts/classify_bodies_pre2024.py --full-panel

python scripts/backtest_cluster_baseline.py

python scripts/backtest_priority_fill.py

python scripts/robustness_size_tier.py
python scripts/robustness_sector_ff48.py
python scripts/robustness_alt_llm_stack.py
python scripts/robustness_body_window.py

# Robustness comparisons for Finding 1b, 1c
python scripts/robustness_masked_vs_unmasked.py
python scripts/robustness_pre2024_full_panel.py

python scripts/make_paper_figures.py
```

13 Appendix C: Data File Manifest

Table 15: Data files underlying the reported tables.

File	Used in
`data/edgar-form4-parsed/{Q}.jsonl`	Sample waterfall (Table 1), Section 3.2
`data/edgar-form4-parsed-delisted/{Q}.jsonl`	Delisted-issuer bridge, Section 3.2
`data/universe/dsenames_bridge.parquet`	CRSP-to-SIC map, Section 3.3 and Section 6.2
`data/backtest-cache/crsp_survivorship_daily.json`	Daily returns, Tables 2–14
`data/llm-labels/full_history.jsonl`	Three-vendor labels, Tables 5, 6, 9
`data/llm-labels/full_history_masked.jsonl`	Identifier-masked labels, Table 6 (Finding 1b)
`data/llm-labels/pre2024_vintage_full.jsonl`	Pre-2024 vintage labels (7,656 events), Tables 7–9 (Finding 1c)
`data/llm-cache-pre2024/*.json`	Pre-2024 raw response cache (per item x model)
`data/size_tier_priority_fill.log`	Size-tier partition (Table 11)
`data/sector_priority_fill.log`	Sector partition (Table 12)
`data/body_window_sensitivity.log`	Timing-gap decomposition (Table 10)