CURIOSITYBITS LIVING LAB BOSTON · STUDY #02 RESEARCH ARTICLE (V4.0.1 — CORPORATE AI WASHING, SECTOR BENCHMARKS & EXPLORATORY LIMITS)

Corporate "AI Washing," Verifiable Financial Receipts, and the Benchmark Mirage: Evidence from 180 SEC Form 8-K Earnings Releases (2023–2025)

CuriosityBits Living Lab Boston

Abstract

Did U.S. equity markets shift from rewarding generic artificial intelligence (AI) enthusiasm during the 2023–2024 technology boom to penalizing unbacked promotional claims—"AI washing"—in 2024–2025? We test this hypothesis across N = 180 outcome-blind SEC Form 8-K Item 2.02 quarterly earnings press releases filed by G = 45 major technology firms across four reporting seasons (Q1 2023, Q3 2023, Q3 2024, and Q1 2025), scored by three independent large language models and a deterministic linguistic analyzer. Both pre-registered confirmatory hypotheses are refuted: within the late cohort (N = 90), the controlled slope of 30-day post-announcement cumulative abnormal returns (CAR) on our continuous AI Washing Index is +6.66% (t = +1.20, p = 0.2318 vs. S&P 500) and +6.64% (t = +1.31, p = 0.1894 vs. XLK), while the full-panel regime interaction is null (-10.34%, t = -1.46, p = 0.1436 vs. S&P 500; -7.28%, t = -1.02, p = 0.3057 vs. XLK). Apparent 30-day return shifts of -3.05% overall and -4.69% among high-washing releases vs. the S&P 500 vanish against technology benchmarks (-0.07% and -1.83% vs. XLK; -0.39% and -2.17% vs. QQQ). An exploratory announcement-day drop of -5.35% (vs. XLK) on AI-intensive releases attenuates under 45 firm fixed effects (-5.08%, clustered t = -1.36, p = 0.1744) and disappears without Q1 2025 (-2.57%, t = -0.95). Finally, text-based forecasting ($5.832% MAE) underperforms text-free baselines (5.778%$ MAE).

1. Introduction

Every major technological wave invites symbolic corporate storytelling. Following ChatGPT's debut in late 2022, executives across enterprise software, semiconductors, cloud infrastructure, and digital platforms rapidly wove generative AI vocabulary into quarterly earnings releases. As AI mentions proliferated, asset managers, journalists, and securities regulators warned of "AI washing"—a disclosure strategy analogous to corporate greenwashing in which managers deploy technological buzzwords and vague promises to capture valuation premia without commensurate capital commitment or auditable revenue (Juma'h, 2026; Montgomery et al., 2023; Sun et al., 2026). By late 2024 and early 2025, market commentary converged on a popular regime-shift

thesis: equity investors had supposedly entered a "show-me-the-receipts" era in which unsubstantiated AI rhetoric triggered post-announcement stock penalties unless backed by hard financial numbers.

From the standpoint of disclosure theory and financial econometrics, this thesis is plausible yet empirically hazardous. Strategic communication models show that costless verbal claims ("cheap talk") lose separating power when managerial and investor incentives diverge, requiring credible issuers to rely on verifiable disclosures (Crawford & Sobel, 1982; Lipnowski & Doron, 2020; Skinner, 1994). Theories of symbolic decoupling similarly suggest that external scrutiny eventually penalizes gaps between ceremonial talk and operational execution (Bothello et al., 2023; Durand et al., 2017; Sánchez et al., 2020). Yet event-study scholarship warns that multi-week abnormal return tests on industry-concentrated panels are acutely vulnerable to benchmark misspecification and factor rotations (Barber & Lyon, 1997; Goldsmith-Pinkham & Lyu, 2025; Lyon et al., 1999). When every sample firm belongs to high-beta technology and earnings cluster in calendar time, benchmarking drift against the S&P 500 risks mistaking a sector drawdown for a backlash against corporate prose.

We test the "AI washing backlash" hypothesis across N = 180 official SEC Form 8-K Item 2.02 quarterly earnings releases filed by G = 45 major U.S. technology companies between January 24, 2023, and March 20, 2025. Our protocol programmatically selects the earliest two eligible releases per firm on or after January 1, 2023, and the earliest two on or after July 1, 2024, enforcing a 75-day minimum spacing rule (realized minimum gap of 161 days) and anchoring reaction days to exact SEC EDGAR acceptance timestamps. Each narrative excerpt is scored independently by three large language models and a deterministic linguistic analyzer across *AI Promotional Intensity* (H), *Verifiable Financial Receipt Specificity* (V), and *Guidance Vagueness* (U), forming a continuous *AI Washing Index* (W).

Four findings demonstrate both the absence of a post-earnings AI washing penalty and how naive benchmarks manufacture illusory narrative effects.

First, both pre-registered confirmatory hypotheses are refuted. Conditional on audited SEC XBRL year-over-year revenue growth, announcement-day returns, pre-event momentum, pre-event volatility, lexical sentiment, and sub-industry fixed effects, the late-cohort (N = 90, G = 45) slope of 30-day post-announcement CAR on W is +6.66% (t = +1.20, p = 0.2318) vs. the S&P 500 and +6.64% (t = +1.31, p = 0.1894) vs. the Technology Select Sector Index (XLK)—statistically zero and positive in sign (Hypothesis 1). Across all N = 180 filings, the continuous regime interaction (W × Post) is -10.34% (t = -1.46, p = 0.1436) vs. the S&P 500 and -7.28% (t = -1.02, p = 0.3057) vs. XLK (Hypothesis 2).

Second, benchmarking technology stocks against the S&P 500 creates a textbook benchmark mirage. Because calendar sampling selects the earliest two filings per firm in each era, the 180 events fall into four eight-week seasons of N = 45 filings each: Q1 2023, Q3 2023, Q3 2024, and Q1 2025. Against the S&P 500, period-shift regressions suggest a -3.05% full-panel 30-day CAR decline (t = -2.14, p = 0.0325) and a -4.69% decline among releases with elevated Unclipped Washing Scores (t = -2.51, p = 0.0119), neither of which vanishes under static CAPM beta adjustment (-2.63%, t = -2.00 and -4.37%, t = -2.47, mean beta = 1.650). Yet

benchmarking against technology indices eliminates both shifts completely: against XLK, estimates collapse to -0.07% (t = -0.05, p = 0.9588) and -1.83% (t = -1.08, p = 0.2807); against the Nasdaq-100 (QQQ), they fall to -0.39% (t = -0.29, p = 0.7718) and -2.17% (t = -1.22, p = 0.2227). Technology beat the S&P 500 by +3.48 percentage points in Q1 2023 (+2.44% vs. -1.05%) but fell -2.67 percentage points harder in Q1 2025 (-8.60% vs. -5.93%) amid the DeepSeek-R1 shock and tariff selloff. Moreover, within Q1 2025 itself (N = 45), the controlled slope of 30-day CAR on W is *positive* (+15.44%, t = +2.38, p = 0.0172 vs. S&P 500; +14.69%, t = +2.34, p = 0.0193 vs. XLK), as steep reversals hit low-washing February earnings winners.

Third, on the immediate announcement day, releases with high AI Promotional Intensity (H ≥ 0.25, N = 89, G = 36) experienced an exploratory shift of -5.35% (t = -2.41, p = 0.0158) vs. XLK and -5.23% (t = -2.34, p = 0.0195) vs. QQQ. Although this reaction-day shift survives Sub-Industry × Season fixed effects (-7.65%, clustered t = -2.33, p = 0.0197), adding 45 firm fixed effects attenuates the pooled differential to -5.08% (clustered standard error = 3.74%, clustered t = -1.36, p = 0.1744), and dropping Q1 2025 reduces it to noise (-2.57%, t = -0.95, p = 0.3423). The pattern is thus a single-season between-firm association rather than a generalizable within-firm penalty.

Fourth, in return prediction, a 22-constant heuristic AI narrative rule ($5.832% full-panel MAE; 6.355% late-cohort MAE) fails to beat a two-number text-free regime baseline (5.778% and 6.221% MAE) or a constant -1.0% rule (5.822% and 6.320%$ MAE).

By tracing how apparent penalties in early hand-curated pilots (N = 40 and N = 48) vanished in an outcome-blind SEC panel (N = 180), our study advances empirical standards for textual finance and AI disclosure research (Ante & Saggu, 2025; Babina et al., 2020; Basnet et al., 2025). Verification scripts, archived editions, and version-control histories appear in the *Technical and Econometric Supplement*.

2. Theoretical Framework: Cheap Talk, Symbolic Washing, and Benchmark Attribution

Cheap Talk, Verifiable Disclosure, and Earnings Guidance Credibility

Economic signaling theory draws a foundational boundary between costless qualitative claims and verifiable quantitative commitments. In Crawford and Sobel's (1982) model of strategic information transmission, an informed sender communicates costless messages to a receiver whose action shapes both agents' payoffs. When sender preferences are upward-biased—as when executives benefit from higher short-term equity valuations—equilibrium cheap talk can sustain at best coarse, noisy partitions and always admits an uninformative babbling equilibrium. When sender motives are transparent and state-independent, Lipnowski and Doron (2020) show that purely self-serving qualitative claims cannot separate high-capability types from imitators unless the sender degrades promotional claims or commits to verifiable benchmarks.

Empirical accounting research operationalizes this boundary in corporate financial reporting. Skinner (1994) documents a stark asymmetry in voluntary disclosure: managers typically convey favorable earnings news via

quantitative point or range estimates that can be audited ex post against GAAP realizations, while communicating bad news through qualitative framing to temper litigation exposure. Confirming quantitative earnings forecasts measurably reduce analyst dispersion and elicit positive stock reactions because markets price reduced fundamental uncertainty (Clement et al., 2003). Moreover, Hutton (2005) shows that active earnings guidance concentrates among complex, high-market-to-book, institutionally owned firms—precisely the profile of major technology corporations—while Skinner (2023) demonstrates that managers match disclosure channel richness to subject-matter complexity to lower investor processing costs.

Crucially, investors condition the weight placed on forward-looking narrative commentary on accompanying financial verifiability. Athanasakou and Hussainey (2014) find that narrative performance disclosures earn market credibility primarily when supported by high-quality audited earnings, whereas opportunistic tone in qualitative non-earnings forward-looking statements is associated with managerial overinvestment (Chen et al., 2024). Similarly, when peer disclosure lawsuits heighten litigation risk, managers shift away from verifiable historical metrics toward less verifiable forward-looking talk (Robinson et al., 2025). Together, disclosure theory implies that AI buzzwords (H) paired with vague guidance (U) represent low-cost cheap talk, whereas explicit AI revenue, backlog, or margin commitments (V) provide verifiable anchors.

Corporate Washing, Symbolic Decoupling, and Technological Narratives

Organizational scholars analyze gaps between external rhetoric and internal execution through symbolic decoupling. Modeling organizational responses to external pressures, Durand et al. (2017) show that when issue salience surges faster than firms can mobilize operational capabilities, managers frequently adopt symbolic conformity—embracing fashionable language while leaving core operations unchanged. In corporate sustainability, Montgomery et al. (2023) identify the rise of "Greenwash 3.0": *narratives about the future*, where firms rely on sweeping, hard-to-falsify transformation promises rather than falsifiable historical statements. High-status group affiliations can partially insulate prominent firms from perceived washing penalties (Bothello et al., 2023). Nevertheless, where analyst scrutiny exposes persistent decoupling, empirical studies find wider analyst forecast errors and higher implied cost of capital (Sánchez et al., 2020), alongside negative stock reactions to documented greenwashing news (Xu et al., 2025).

Recent scholarship on AI adoption and corporate filings mirrors this tension between substantive capability and symbolic washing. Using resume and job-posting microdata, Babina et al. (2020) demonstrate in the *Journal of Financial Economics* that genuine firm-level AI talent investments boost sales growth, product innovation, and market value, concentrating gains among ex ante large firms. Although generative AI tools deliver measurable productivity gains in operational settings—such as the $15% increase in customer-support resolutions documented by Brynjolfsson et al. (2024)—equity markets do not blindly reward every AI announcement: Lui et al. (2021) find an average -1.77%$ announcement-day return for AI investment announcements, concentrated among firms with weak IT capabilities. In corporate Form 10-K filings, Ante and Saggu (2025) show that deeper pre-existing AI engagement predicted positive abnormal returns around ChatGPT's launch, while Basnet et al. (2025) find that *actionable* AI implementation disclosures enhance firm value and R&D

whereas *speculative* AI mentions have negligible valuation effects. Juma'h (2026) documents wide dispersion between standards-aligned Form 10-K disclosures and promotional "AI-washing" rhetoric, and Sun et al. (2026) link AI narrative-capability gaps to R&D investment and investor sentiment.

Shiller's (2017) narrative economics connects these cross-sectional signaling dynamics to aggregate market cycles: viral technological stories initially propel investor enthusiasm before fatiguing as audiences demand tangible proof. Integrating Shiller (2017) with Crawford and Sobel (1982) motivated our pre-registered regime hypothesis: markets might tolerate or reward unbacked AI rhetoric during the 2023 boom, but penalize high AI Washing Index scores over the 30-day post-announcement window in 2024–2025.

Why Naive Event Studies Create Illusory Narrative Penalties

Financial econometrics establishes that multi-week abnormal return tests are joint tests of the economic hypothesis and the benchmark return model (Barber & Lyon, 1997; Lyon et al., 1999). Barber and Lyon (1997) prove that calculating abnormal returns against broad market indices induces severe test misspecification, demonstrating the necessity of size- and style-matched reference portfolios. Lyon et al. (1999) further warn that in nonrandom, industry-clustered samples, multi-week abnormal return inference is "treacherous" because shared industry factors violate cross-sectional independence. Formalizing causal inference in financial event studies, Goldsmith-Pinkham and Lyu (2025) prove that misspecified factor models yield inconsistent treatment estimates whose bias peaks during volatile periods and when event timing correlates with sector conditions.

In a panel of technology earnings releases spanning 2023 to early 2025, market-adjusting returns against the S&P 500—or even subtracting a constant pre-event CAPM beta times the S&P 500—creates a potent omitted-variable trap. First, if technology stocks outperform the S&P 500 during early-regime earnings seasons (Q1 2023) and sell off more steeply than the S&P 500 during late-regime seasons (Q1 2025), market-adjusted CARs mechanically turn negative even if every stock matches its sector index. Second, because routine AI product commentary diffused broadly across technology releases after mid-2023, conditioning on elevated AI washing or buzzword thresholds isolates technology issuers exposed to the Q1 2025 sector drawdown, misattributing an aggregate sector rotation to a narrative penalty.

---

3. Data, Multi-Judge Measurement, and Research Design

Primary-Source SEC Corpus and Four-Season Panel Structure

Our dataset consists exclusively of primary-source **SEC EDGAR Form 8-K Item 2.02** ("Results of Operations and Financial Condition") filings with attached quarterly earnings press releases (Exhibit 99), excluding Form 10-Q cover pages and third-party transcripts.

The sample spans G = 45 U.S.-listed technology corporations across four sub-industries: Enterprise Software and SaaS (13 firms, including Microsoft, Oracle, Salesforce, Adobe, ServiceNow, Palantir, UiPath, and C3.ai), Cloud Data, Cybersecurity, and Networking (12 firms, including Snowflake, MongoDB, Datadog, Cloudflare,

CrowdStrike, Arista, and Dell), Semiconductors and AI Hardware (11 firms, including NVIDIA, AMD, Intel, Broadcom, Micron, Applied Materials, and Marvell), and Internet and Cloud Platforms (9 firms, including Alphabet, Meta, Amazon, AppLovin, Pinterest, Twilio, and Uber). This 45-firm roster was specified with hindsight of prominent participants in the 2023–2025 AI cycle; outcome-blindness governs event selection *within* each firm.

For each firm, an automated script queries EDGAR submission histories and selects four non-overlapping Item 2.02 filings without inspecting stock returns: the earliest two eligible releases on or after January 1, 2023 (**Early AI Boom cohort**, N = 90), and the earliest two on or after July 1, 2024 (**Show-Me-The-Receipts cohort**, N = 90), separated by at least 75 days (minimum realized same-firm gap = 161 days). Because corporate calendars align closely, all N = 180 events fall into four ~8-week seasons of N = 45 filings each (one per firm), ending March 20, 2025: **Q1 2023** (`2023-01-24` to `2023-03-28`), **Q3 2023** (`2023-07-19` to `2023-09-27`), **Q3 2024** (`2024-07-22` to `2024-09-25`, spanning the `2024-09-18` Federal Reserve rate cut), and **Q1 2025** (`2025-01-23` to `2025-03-20`, coinciding with the late-January 2025 DeepSeek-R1 shock and tariff selloff).

Reaction dates follow exact SEC acceptance timestamps in Eastern Time: market-hours filings before `16:00` ET ($12/180, or 6.7%) map to the filing date, whereas after-hours filings at or after `16:00` ET (168/180, or 93.3%; 0/180$ at or after `17:30` ET) roll to the next trading day (`2023-01-25` to `2025-03-21`). Post-announcement 30-calendar-day (21-trading-day) CARs are computed strictly forward from the reaction-day adjusted close without winsorization.

Excerpt Scope and Multi-Judge Construct Scoring

Our parser extracts narrative paragraphs (≥ 50% alphabetic characters) prioritized by AI and guidance keywords. Across all N = 180 filings, excerpts average $298.4 words (median 288.5, range 159 to 807) and 1987.3 characters (range 1069 to 4878). Auditing 30 complete Form 8-K releases (15 per cohort; mean full length 4414.9 words vs. 299.5 excerpt words) shows that excerpts represent 7.3% of full press-release character length (7.1% early, 7.4% late) and enrich AI-term density by 5.42 times overall (3.57 times early, 2.00 vs. 0.56 terms per 1,000 words; 6.26 times late, 7.76 vs. 1.24 terms per 1,000$ words). Thus, our constructs measure management's framing of AI-relevant narrative passages rather than full-document word counts.

Tailored to written press releases, three primary constructs on [0, 1] are scored: (1) **AI Promotional Intensity (H)**, capturing generative AI, agentic, LLM, copilot, and transformation prominence; (2) **Verifiable Financial Receipt Specificity (V)**, capturing auditable dollar revenue, recurring revenue, margin basis points, backlog dollars, and quantitative guidance ranges; and (3) **Guidance Vagueness (U)**, capturing qualitative superlatives and non-committal hedges relative to concrete numbers. These form the pre-registered **AI Washing Index** (W) and centered **Unclipped Washing Score** (W^*):

> *$W = clip\left(0.45\ H + 0.55\ U - 0.65\ V,\; 0.0,\; 1.0\right), \qquad W^{*} = clip\left(0.50 + 0.45\ H + 0.55\ U - 0.65\ V,\; 0.0,\; 1.0\right)$*

Because $109/180 (60.6%) of filings contain sufficient quantitative disclosures that 0.45 H + 0.55 U - 0.65 V ≤ 0, W equals zero for 65.6% (59/90) of early filings (71.1%, or 32/45, in Q1 2023), making W^{*} an essential uncensored check.

Every excerpt is scored independently at zero temperature by Google Gemini 3 Flash Preview, DeepSeek-V3, and Llama 3.3 70B Instruct, plus a deterministic lexical-receipt analyzer. Tables 1 and 2 report seasonal descriptive statistics and Shrout-Fleiss intraclass correlation coefficients (`ICC(1,k)`, `ICC(2,k)`, `ICC(3,k)`, `ICC(2,1)`) across all N = 180 filings. Ensemble `ICC(2,k)` across the three LLMs reaches $0.942 for H, 0.933 for V, 0.823 for U, and 0.916 for W$, ruling out measurement attenuation as an explanation for null return slopes.

Table 1. Panel Descriptive Statistics Across the Four SEC Form 8-K Earnings Seasons (N = 180, G = 45)

Season / Panel	Events (N)	Filing Window	Day-1 CAR vs. SPY	30-Day CAR vs. SPY (SD)	30-Day CAR vs. XLK	Mean H	Mean V	Mean U	Mean W	Mean W^{*}	W = 0 Share
Q1 2023	45	Jan 24 – Mar 28, 2023	+3.01%	-0.37% ($7.96%)	-3.85%	0.170	0.599	0.397	0.072	0.401	71.1% (32/45$)
Q3 2023	45	Jul 19 – Sep 27, 2023	+0.53%	-0.04% ($5.60%)	-0.26%	0.322	0.589	0.406	0.095	0.487	60.0% (27/45$)
Q3 2024	45	Jul 22 – Sep 25, 2024	+0.69%	-0.91% ($7.31%)	-0.77%	0.326	0.616	0.384	0.075	0.457	60.0% (27/45$)
Q1 2025	45	Jan 23 – Mar 20, 2025	-1.61%	-4.30% ($9.34%)	-1.62%	0.357	0.585	0.415	0.084	0.508	51.1% (23/45$)
Full Panel	**180**	**Jan 24, 2023 – Mar 20, 2025**	**+0.65%**	**-1.40% ($7.80%)**	**-1.63%**	**0.294**	**0.597**	**0.401**	**0.082**	**0.463**	**60.6% (109/180$)**

Notes: Full-panel standard deviations (N = 180): Day-1 CAR vs. SPY = 10.52% (min -25.16%, max +32.05%); 30-day CAR vs. SPY = 7.80% (min -27.96%, max +22.96%); 30-day CAR vs. XLK = 7.32% (min -25.02%, max +24.96%); H = 0.225 (early mean $0.246, late mean 0.341); V = 0.207; U = 0.144; W = 0.148; W^{*} = 0.241$.

Table 2. Multi-Judge Inter-Rater Reliability Across the Analysis Sample (N = 180)

Textual Construct	Judge Panel	ICC(1,k)	ICC(2,k)	ICC(3,k)	ICC(2,1)	Mean Pairwise r
AI Promotional Intensity (H)	Three LLM Judges	$0.941	0.942	0.963	0.844	0.896$
Verifiable Financial Receipts (V)	Three LLM Judges	$0.932	0.933	0.946	0.823	0.856$
Guidance Vagueness (U)	Three LLM Judges	$0.813	0.823	0.868	0.607	0.693$
Lexical Financial Sentiment	Three LLM Judges	$0.941	0.942	0.953	0.843	0.875$
AI Washing Index (W)	Three LLM Judges	$0.915	0.916	0.918	0.783	0.801$
AI Promotional Intensity (H)	All Four Judges	$0.939	0.940	0.953	0.797	0.846$
Verifiable Financial Receipts (V)	All Four Judges	$0.824	0.839	0.915	0.566	0.726$
Guidance Vagueness (U)	All Four Judges	$0.762	0.781	0.847	0.471	0.578$
AI Washing Index (W)	All Four Judges	$0.857	0.861	0.884	0.608	0.670$

Notes: Across all four judges, Lexical Financial Sentiment achieves `ICC(1,k) = 0.880`, `ICC(2,k) = 0.880`, `ICC(3,k) = 0.885`, `ICC(2,1) = 0.648`, and mean pairwise r = 0.731.

Econometric Specifications and Benchmark Design

Every multivariate specification controls for **Audited SEC XBRL Year-over-Year Quarterly Revenue Growth**, **20-Day Pre-Event Stock Momentum**, **20-Day Pre-Event Annualized Realized Volatility**, **Lexical Financial Sentiment**, and **Sub-Industry Fixed Effects**, with **Immediate Announcement-Day Return** included in all 30-day post-announcement drift models. Multi-season models cluster standard errors by firm (G = 45); single-season cross-sections (N = 45) report `HC1` robust standard errors.

Two confirmatory hypotheses were pre-registered: **Hypothesis 1** tests whether the controlled slope of 30-day CAR on continuous W is negative within the post-July 2024 cohort (N = 90, G = 45), and **Hypothesis 2** tests whether the continuous regime interaction W × Post is negative across the full panel (N = 180, G = 45). All specifications are evaluated across four benchmarks: **S&P 500 (SPY)**, **Technology Select Sector (XLK)**, **Nasdaq-100 (QQQ)**, and **CAPM Beta-Adjusted SPY** using pre-event [t-280, t-30] market betas (mean = 1.650, median = 1.593, range $0.50 to 3.07$).

4. Empirical Results

Refutation of Pre-Registered Confirmatory Hypotheses

Table 3 Panel A reports the confirmatory tests. Both pre-registered hypotheses fail cleanly. Within the post-July 2024 cohort (N = 90, G = 45), Hypothesis 1 predicted a negative 30-day CAR slope on the AI Washing Index (W). The actual estimate is **+6.66%** (t = +1.20, p = 0.2318) vs. SPY and **+6.64%** (t = +1.31, p = 0.1894) vs. XLK—statistically indistinguishable from zero and opposite in sign to a penalty. Across the full panel (N = 180, G = 45), Hypothesis 2's continuous interaction (W × Post) is **-10.34%** (t = -1.46, p = 0.1436) vs. SPY and **-7.28%** (t = -1.02, p = 0.3057) vs. XLK, likewise null.

Table 3. Confirmatory Hypothesis Tests and Multi-Benchmark Exploratory Regime Comparisons

Specification & Return Window	Sample (N, G)	vs. S&P 500 (SPY)	vs. Tech Sector (XLK)	vs. Nasdaq-100 (QQQ)	CAPM Beta SPY
Panel A: Confirmatory Hypotheses (Null)					
Hypothesis 1: Late-Cohort 30-Day Slope on W	N = 90, G = 45	+6.66% (t = +1.20, p = 0.2318)	+6.64% (t = +1.31, p = 0.1894)	—	—
Hypothesis 2: Full-Panel Interaction (W × Post)	N = 180, G = 45	-10.34% (t = -1.46, p = 0.1436)	-7.28% (t = -1.02, p = 0.3057)	—	—
Panel B: Retired 30-Day Period Shifts					
Full-Panel 30-Day Shift (Post)	N = 180, G = 45	-3.05% (t = -2.14, p = 0.0325)	**-0.07% (t = -0.05, p = 0.9588)**	**-0.39% (t = -0.29, p = 0.7718)**	-2.63% (t = -2.00, p = 0.0459)
High-Washing 30-Day Shift ($W^* \geq 0.38$)	N = 114, G = 40				

Specification & Return Window	Sample (N, G)	vs. S&P 500 (SPY)	vs. Tech Sector (XLK)	vs. Nasdaq-100 (QQQ)	CAPM Beta SPY
		-4.69% (t = -2.51, p = 0.0119)	**-1.83% (t = -1.08, p = 0.2807)**	**-2.17% (t = -1.22, p = 0.2227)**	-4.37% (t = -2.47, p = 0.0134)
Pooled 30-Day Differential (Post × \mathbb{I}[W* ≥ 0.38])	N = 180, G = 45	-2.78% (t = -1.02, p = 0.3066)	-2.72% (t = -1.12, p = 0.2624)	-2.79% (t = -1.10, p = 0.2721)	—
Panel C: Exploratory Patterns & Limits					
High-AI Day-1 CAR Shift (H ≥ 0.25)	N = 89, G = 36	-5.27% (t = -2.31, p = 0.0206)	-5.35% (t = -2.41, p = 0.0158)	-5.23% (t = -2.34, p = 0.0195)	-5.25% (t = -2.32, p = 0.0203)
Pooled Day-1 Differential (Post × \mathbb{I}[H ≥ 0.25])	N = 180, G = 45	-6.39% (t = -2.05, p = 0.0407)	-6.90% (t = -2.24, p = 0.0254)	-6.65% (t = -2.15, p = 0.0312)	—
Early-Cohort 30-Day Slope on W (2023 Boom)	N = 90, G = 45	+15.24% (t = +2.56, p = 0.0104)	+13.29% (t = +2.29, p = 0.0220)	+13.90% (t = +2.48, p = 0.0130)	+16.70% (t = +2.74, p = 0.0062)

The Benchmark Mirage: Four-Season Decomposition and Sector Attribution

Why did SPY-benchmarked period shifts in Table 3 Panel B initially suggest a -3.05% full-panel 30-day drop (t = -2.14, p = 0.0325) and a -4.69% high-washing subgroup drop (t = -2.51, p = 0.0119)? Table 4 decomposes returns across the four seasons and leaves out one season at a time.

Table 4. Four-Season Benchmark Decomposition, Within-Season Slopes, and Leave-One-Season-Out Tests

Diagnostic Panel / Season	N	SPY 30d	XLK 30d (XLK - SPY)	QQQ 30d (QQQ - SPY)	CAR vs. SPY	CAR vs. XLK	CAPM Beta CAR	Within-Season W Slope (vs. SPY)
Panel A: Season Returns								
Q1 2023	45	-1.05%	+2.44% (+3.48 pp)	+1.86% (+2.91 pp)	-0.37%	-3.85%	+0.34%	+26.604% (t = +2.72, p = 0.0066)

Diagnostic Panel / Season	N	SPY 30d	XLK 30d (XLK - SPY)	QQQ 30d (QQQ - SPY)	CAR vs. SPY	CAR vs. XLK	CAPM Beta CAR	Within-Season W Slope (vs. SPY)
Q3 2023	45	-1.99%	-1.77% (+0.22 pp)	-1.55% (+0.44 pp)	-0.04%	-0.26%	+1.01%	+5.408% (t = +0.66, p = 0.5115)
Q3 2024	45	+3.12%	+2.99% (-0.13 pp)	+2.89% (-0.23 pp)	-0.91%	-0.77%	-3.23%	+2.274% (t = +0.23, p = 0.8146)
Q1 2025	45	-5.93%	-8.60% (-2.67 pp)	-7.78% (-1.85 pp)	-4.30%	-1.62%	+0.27%	+15.445% (t = +2.38, p = 0.0172)
Panel B: Leave-One-Season-Out	**N**	**Full-Panel 30d Shift**	**High-Washing 30d Shift**	**High-AI Day-1 Shift**				
Drop None (Full Panel)	180	-3.05% (t = -2.14)	-4.69% (t = -2.51, N = 114)	-5.27% (t = -2.31, N = 89)				
Drop Q1 2023	135	-2.21% (t = -1.55)	-3.34% (t = -1.90, N = 93)	-4.07% (t = -1.72, N = 82)				
Drop Q3 2023	135	-3.66% (t = -1.92)	-6.70% (t = -2.19, N = 84)	-6.53% (t = -1.47, N = 59)				
Drop Q3 2024	135	-4.69% (t = -2.63)	-6.73% (t = -3.03, N = 84)	-7.86% (t = -3.00, N = 65)				
Drop Q1 2025	**135**	**-1.21% (t = -0.82)**	**-1.77% (t = -0.92, N = 81)**	**-2.57% (t = -0.95, N = 61)**				

Notes: Sample 30-day CARs vs. QQQ are -3.27% (Q1 2023), -0.48% (Q3 2023), -0.68% (Q3 2024), and -2.44% (Q1 2025). Unclipped W^* within-season slopes: +10.190% (t = +1.51) in Q1 2023, +2.962% (t = +0.73) in Q3 2023, -0.244% (t = -0.05) in Q3 2024, and +5.655% (t = +1.05) in Q1 2025.

First, mean 30-day CARs vs. SPY were flat across Q1 2023 (-0.37%), Q3 2023 (-0.04%), and Q3 2024 (-0.91%) before falling to -4.30% in Q1 2025. Dropping Q1 2025 (N = 135) collapses all three shifts to noise: -1.21% (t = -0.82, p = 0.4135), -1.77% (t = -0.92, p = 0.3553), and -2.57% (t = -0.95, p = 0.3423).

Second, CAPM beta adjustment against SPY fails (-2.63%, t = -2.00, p = 0.0459 and -4.37%, t = -2.47, p = 0.0134) because SPY rose +3.12% in Q3 2024; subtracting $1.650 × 3.12% pushes Q3 2024's beta-adjusted mean down to -3.23% (vs. +0.27% in Q1 2025). Benchmarking directly against technology ETFs absorbs the true industry rotation—XLK beat SPY by +3.48 percentage points in Q1 2023 (+2.44% vs. -1.05%) but trailed SPY by -2.67 percentage points in Q1 2025 (-8.60% vs. -5.93%). Against XLK, both 30-day shifts vanish (-0.07%, t = -0.05, p = 0.9588 and -1.83%, t = -1.08, p = 0.2807), as they do against QQQ (-0.39%, t = -0.29, p = 0.7718 and -2.17%, t = -1.22, p = 0.2227). Across every W^{} *cut-off from 0.25 (N = 143, G = 43: -0.11%, t = -0.08) to 0.50 (N = 71, G = 34: -3.11%, t = -1.37), XLK-adjusted shifts remain statistically zero, and the pooled differential (Post × \mathbb{I}[W^{} ≥ 0.38]) is null across all benchmarks (-2.78% vs. SPY, -2.72% vs. XLK, -2.79% vs. QQQ). Although Q1 2025 still averages -1.62% CAR vs. XLK (-2.44%$ vs. QQQ), SPY benchmarking supplied the sector-rotation wedge that created artificial significance.

Third, inside Q1 2025 itself (N = 45), there is zero penalty on W: the bivariate slope is +8.05% (`HC1` SE = 7.32%, t = +1.10, p = 0.2715; W > 0 filings [n = 22/45, 48.9%] averaged -4.52% vs. -4.08% for W = 0 [n = 23/45, 51.1%]), and with controls the slope is *positive* across every benchmark: **+15.44%** (t = +2.38, p = 0.0172) vs. SPY (+17.44%, t = +2.71 without the Day-1 control), **+14.69%** (t = +2.34, p = 0.0193) vs. XLK, **+15.28%** (t = +2.46, p = 0.0137) vs. QQQ, and **+12.68%** (`HC1` SE = 5.28%, t = +2.40, p = 0.0163, N = 42) dropping the top three Cook's distance observations (Marvell Technology, Uber Technologies, and Twilio). Case inspection explains why: Q1 2025's worst 30-day reversals hit low-washing software and internet winners that had soared on Day 1 in early February before reversing in the March tariff selloff—AppLovin (W = 0.097, Day-1 +22.96%, 30-day -27.96%), Cloudflare (W = 0.010, Day-1 +18.68%, 30-day -20.49%), Palantir Technologies (W = 0.000, Day-1 +23.32%, 30-day -17.68%), and Pinterest (W = 0.000, Day-1 +20.00%, 30-day -14.06%)—whereas high-W filings that fell on Day 1, including NVIDIA (W = 0.555, Day-1 -6.88%, 30-day -3.97%), C3.ai (W = 0.502, Day-1 -8.08%, 30-day -3.93%), UiPath (W = 0.432, Day-1 -14.39%, 30-day +11.91%), and Intel (W = 0.191, Day-1 -2.37%, 30-day +13.93%), stabilized or rebounded afterward.

Announcement-Day Shift and Firm Fixed-Effects Attenuation

In Table 3 Panel C, the announcement-day CAR shift on AI-intensive filings (H ≥ 0.25, N = 89, G = 36) survives XLK (-5.35%, t = -2.41, p = 0.0158), QQQ (-5.23%, t = -2.34, p = 0.0195), pooled interaction (-6.90%, t = -2.24, p = 0.0254 vs. XLK), and 16-cell Sub-Industry × Season fixed effects (-7.65%, clustered SE = 3.28%, t = -2.33, p = 0.0197). Yet four diagnostics bound this exploratory result: 1. **Firm Fixed Effects:** Adding **45 firm fixed effects** across all N = 180 filings attenuates the differential to **-5.08%** (`HC1` SE = 3.46%, t = -1.47, p = 0.1417; clustered SE = 3.74%, **clustered t = -1.36, p = 0.1744**). On the N = 89 (G = 36) subgroup, 36 firm fixed effects yield -5.91% (clustered t = -1.91, p = 0.0566), matching the within-firm paired difference across G = 31 firms with both high-H and low-H filings (-3.94%, t = -2.02, p = 0.0528; G = 27

paired regime difference = -4.96%, t = -2.22, p = 0.0355). Thus, the shift largely reflects *which firms* write AI-heavy releases. 2. **Single-Season Concentration:** The $24 high-H filings in Q3 2024 averaged +0.27% Day-1 CAR (median +0.995%); the drop comes entirely from Q1 2025 (28 high-H filings averaging -4.85%, median -6.70%, vs. +3.72% mean and -0.03% median across the 17 low-H filings), and vanishes without Q1 2025 (-2.57%, t = -0.95$). 3. **Small Early Cell and Cut-Off Fluctuation:** Only $7/45 (15.6%) Q1 2023 releases had H ≥ 0.25 (median +3.000%, led by C3.ai +32.05% and NVIDIA +13.49%), and shifting the cut-off to H ≥ 0.30 (N = 73, G = 32) drops the estimate to -3.23% (t = -1.19, p = 0.2346 vs. SPY) and -3.21% (t = -1.21, p = 0.2260$ vs. QQQ). 4. **Unobserved Consensus EPS Surprise:** SEC XBRL revenue growth cannot capture street consensus EPS surprises.

Likewise, the exploratory early-cohort 30-day slope on W (+15.24%, t = +2.56 vs. SPY; +13.29%, t = +2.29 vs. XLK) stems solely from Q1 2023 (+26.604%, t = +2.72 vs. +5.408%, t = +0.66 in Q3 2023), where $32/45 (71.1%) of filings equal zero, attenuating to +5.94% (t = +1.92, p = 0.0545) on W^{*}$.

Forecasting Evaluation Against Text-Free Baselines

Table 5 evaluates whether our 22-constant *Heuristic AI Narrative Rule*—despite in-sample parameter calibration—improves 30-day return prediction relative to text-free benchmarks.

Table 5. Post-Announcement 30-Day Return Forecast Comparison: Text-Based vs. Text-Free Rules

Forecasting Rule	Uses Text?	Full Panel MAE (N = 180)	Full Hits	Late Cohort MAE (N = 90)	Late Hits	Paired t vs. AI Rule (Full / Late)
Text-Free Regime Baseline ($0.0% / -1.65%)	No	**5.778%**	101/180	**6.221%**	60/90	Beats AI Rule (t = +1.07, p = 0.2840 / t = +1.53, p = 0.1269$)
Constant Negative One Percent Rule (-1.0%)	No	**$5.822%**	109/180	**6.320%**	60/90	Beats AI Rule (t = +0.14, p = 0.8852 / t = +0.63, p = 0.5315$)
Heuristic AI Narrative Rule (22 Constants)	Yes	$5.832%	103/180	6.355%	60/90$	Reference Rule
Constant Zero Forecast ($0.0%)	No	5.975%	—	6.614%	—	Worse (t = -2.38, p = 0.0174 / t = -2.36, p = 0.0183$)
	Yes	$5.977%	98/180	6.901%	46/90$	Worse than Text-Free Baselines

Forecasting Rule	Uses Text?	Full Panel MAE (N = 180)	Full Hits	Late Cohort MAE (N = 90)	Late Hits	Paired t vs. AI Rule (Full / Late)
Unshrunk Linear Regression Forecast						
Naive Buzzword Momentum Forecast	Yes	$6.491%	81/180	7.543%	31/90$	Worst Specification

Even with 22 in-sample constants, the Heuristic AI Narrative Rule ($5.832% full-panel MAE; 6.355% late-cohort MAE) performs *worse* than both the two-number **Text-Free Regime Baseline** (5.778% and 6.221% MAE) and the **Constant Negative One Percent Rule** (5.822% and 6.320% MAE). Its apparent victory over Constant Zero (5.975%, t = -2.38, p = 0.0174) came entirely from embedding a -1.65%$ late-regime intercept rather than genuine textual alpha.

5. Discussion: Why Narrative Penalties Vanish Under Sector and Firm Controls

Reconciling Disclosure Theory with Null Post-Earnings Drift

Why do regulated Form 8-K earnings releases exhibit zero post-announcement AI washing penalty despite clear cheap-talk incentives (Crawford & Sobel, 1982; Lipnowski & Doron, 2020)? Three mechanisms reconcile theory with our null evidence. First, rational discounting of cheap talk occurs *immediately*, not via slow 30-day drift: once AI commentary became routine across technology issuers in 2023–2024, institutional analysts discounted unverified superlatives upon release rather than gradually discovering missing receipts over 21 trading days, echoing Basnet et al.'s (2025) finding that speculative AI mentions carry zero valuation weight. Second, unlike voluntary marketing or sustainability pledges (Montgomery et al., 2023), Item 2.02 earnings releases are formal financial disclosures where mean Verifiable Receipt Specificity remained high throughout our sample (V = 0.597), censoring $60.6% of filings at W = 0 (Hutton, 2005; Skinner, 1994). Third, the exploratory Q1 2025 announcement-day drop on high-H filings attenuates under firm fixed effects (-5.08%, t = -1.36$) because AI hardware and cloud incumbents concentrated both AI narrative intensity and elevated valuation multiples (Babina et al., 2020), exposing them jointly to the Q1 2025 DeepSeek-R1 and macro repricing.

Methodological Implications and Limitations

Methodologically, standard market-adjusted event studies and zero-benchmark forecast tests can construct an illusory "narrative penalty" (-4.69% vs. SPY, p = 0.0119; CAPM beta -4.37%, p = 0.0134). Four safeguards

proved decisive: (1) benchmarking industry-concentrated panels against sector indices (XLK, QQQ) rather than broad market proxies (Barber & Lyon, 1997; Goldsmith-Pinkham & Lyu, 2025; Lyon et al., 1999); (2) leave-one-season-out and within-season slope checks, showing that Q1 2025 drove the period drop while exhibiting a *positive* within-season W slope (+15.44% vs. SPY; +14.69% vs. XLK); (3) firm fixed effects to absorb persistent disclosure styles; and (4) benchmarking textual forecasts against text-free regime means.

Three boundaries merit future study: our 45-firm universe reflects hindsight selection of major technology incumbents rather than micro-cap rebranders (Lui et al., 2021); keyword-enriched excerpts ($7.3% of release length; 5.42× AI density enrichment) exclude conference-call Q&A (Skinner, 2023); and SEC XBRL revenue growth omits consensus EPS surprises. Replication scripts, pilot archives (N = 40, N = 48), and audit logs appear in the *Technical and Econometric Supplement*. Across 180$ Form 8-K filings, once sector rotations and firm identities are controlled, the post-earnings "AI washing backlash" dissolves into a benchmark mirage.

References

Ante, L., & Saggu, A. (2025). Quantifying a firm's AI engagement: Constructing objective, data-driven, AI stock indices using 10-K filings. *Technological Forecasting and Social Change*. https://doi.org/10.1016/j.techfore.2024.123965

Athanasakou, V. E., & Hussainey, K. (2014). The perceived credibility of forward-looking performance disclosures. *Accounting and Business Research*. https://doi.org/10.1080/00014788.2013.867403

Babina, T., Fedyk, A., He, A. X., & Hodson, J. (2020). Artificial intelligence, firm growth, and industry concentration. *Journal of Financial Economics*. https://doi.org/10.1016/j.jfineco.2023.103745

Barber, B. M., & Lyon, J. D. (1997). Detecting long-run abnormal stock returns: The empirical power and specification of test statistics. *Journal of Financial Economics*. https://doi.org/10.1016/s0304-405x(96)00890-2

Basnet, A., Elias, M., Salganik-Shoshan, G., Walker, T. J., & Zhao, Y. (2025). Analyzing the market's reaction to AI narratives in corporate filings. *International Review of Financial Analysis*. https://doi.org/10.1016/j.irfa.2025.104378

Bothello, J., Ioannou, I., Porumb, V.-A., & Karaibrahimoglu, Y. (2023). CSR decoupling within business groups and the risk of perceived greenwashing. *Strategic Management Journal*. https://doi.org/10.1002/smj.3532

Brynjolfsson, E., Li, D., & Raymond, L. (2024). Generative AI at work. *The Quarterly Journal of Economics*. https://doi.org/10.1093/qje/qjae044

Chen, J. J., Song, P., & Loi, F. L. (2024). Strategic forward-looking nonearnings disclosure and overinvestment. *The British Accounting Review*. https://doi.org/10.1016/j.bar.2024.101431

Clement, M. B., Frankel, R. M., & Miller, J. (2003). Confirming management earnings forecasts, earnings uncertainty, and stock returns. *Journal of Accounting Research*. https://doi.org/10.1111/1475-679x.00119

Crawford, V. P., & Sobel, J. (1982). Strategic information transmission. *Econometrica*. https://doi.org/10.2307/1913390

Durand, R., Hawn, O. V., & Ioannou, I. (2017). Willing and able: A general model of organizational responses to normative pressures. *Academy of Management Review*. https://doi.org/10.5465/amr.2016.0107

Goldsmith-Pinkham, P. S., & Lyu, T. (2025). Causal inference in financial event studies. *arXiv Working Paper*. https://doi.org/10.48550/arxiv.2511.15123

Hutton, A. P. (2005). Determinants of managerial earnings guidance prior to Regulation Fair Disclosure and bias in analysts' earnings forecasts. *Contemporary Accounting Research*. https://doi.org/10.1506/6qur-cr5m-aqqx-kx1a

Juma'h, A. H. (2026). Artificial intelligence disclosure practices in US 10-K filings: Accounting standards, signalling incentives and reporting quality. *Journal of Applied Accounting Research*. https://doi.org/10.1108/jaar-07-2025-0355

Lipnowski, E., & Doron, R. (2020). Cheap talk with transparent motives. *Econometrica*. https://doi.org/10.3982/ecta15674

Lui, A. K. H., Lee, M. C. M., & Ngai, E. W. T. (2021). Impact of artificial intelligence investment on firm value. *Annals of Operations Research*. https://doi.org/10.1007/s10479-020-03862-8

Lyon, J. D., Barber, B. M., & Tsai, C.-L. (1999). Improved methods for tests of long-run abnormal stock returns. *The Journal of Finance*. https://doi.org/10.1111/0022-1082.00101

Montgomery, A. W., Lyon, T. P., & Barg, J. (2023). No end in sight? A greenwash review and research agenda. *Organization & Environment*. https://doi.org/10.1177/10860266231168905

Robinson, S., Skinner, A. N., & Wang, J. (2025). Does litigation risk shape environmental disclosure decisions? Evidence from peers' environmental disclosure lawsuits. *The Accounting Review*. https://doi.org/10.2308/tar-2023-0352

Sánchez, I., Hussain, N., Khan, S. A., & Martínez-Ferrero, J. (2020). Do markets punish or reward corporate social responsibility decoupling? *Business & Society*. https://doi.org/10.1177/0007650319898839

Shiller, R. J. (2017). Narrative economics. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2896857

Skinner, A. N. (2023). Subject matter complexity and disclosure channel richness. *The Accounting Review*. https://doi.org/10.2308/tar-2021-0337

Skinner, D. J. (1994). Why firms voluntarily disclose bad news. *Journal of Accounting Research*. https://doi.org/10.2307/2491386

Sun, Z., Wen, Y., Zhao, L., Almugren, I., & Galgotia, A. (2026). Unveiling AI washing: Bridging corporate technological gaps through a cognitive dissonance lens. *Technological Forecasting and Social Change*. https://doi.org/10.1016/j.techfore.2025.124511

Xu, M., Tse, Y. K., Geng, R., Liu, Z., & Potter, A. T. (2025). Greenwashing and market value of firms: An empirical study. *International Journal of Production Economics*. https://doi.org/10.1016/j.ijpe.2025.109606