CURIOSITYBITS LIVING LAB BOSTON · STUDY #02 ARCHIVED WORKING PAPER (V3.2 — PRE-EXTERNAL-REVIEW SEC 8-K & XBRL EDITION)

From AI Gold Rush to Narrative Saturation: Testing "AI Washing" vs. Verifiable Financial Receipts on 180 Outcome-Blind SEC Form 8-K Earnings Releases (2023–2026) — v3.2 (Multi-Skill Rebuilt & Cross-Lab Verified)

> ***Version & Transparency Notice (`v3.2` — November 2, 2026, Incorporating Singapore Audits #10 `EXT-AUDIT-014` & #11 `EXT-AUDIT-016`):*** *This is the* ***rebuilt `v3.2` Living Paper (N = 180 outcome-blind, non-overlapping SEC Form 8-K Item 2.02 `EX-99.1` earnings press releases across G = 45 major technology, semiconductor, cloud, enterprise software, cybersecurity, and internet companies)****, constructed by applying all* ***25 active lab skills (`SKILL-001` – `SKILL-026` , minus deprecated `SKILL-004`)*** *and incorporating all econometric, reliability, and text-free benchmark checks from* ***CuriosityBits Living Lab Singapore Audit #10 (`EXT-AUDIT-014` , `Muse Spark 1.3`)*** *and* ***Audit #11 (`EXT-AUDIT-016` , `MiniMax-M3` , `PARTIALLY_REPLICATED` / `REPLICATED_WITH_CAVEATS`)****. In accordance with* ***`SKILL-025` (`end-to-end-regenerable-pipeline-and-immutable-editions`)****, all prior editions— `v1.0` (`PAPER_MANUSCRIPT_V1_ARCHIVED.md` , N=40), `v2.0` (`PAPER_MANUSCRIPT_V2_ARCHIVED.md` , N=48), and `v3.0` (`PAPER_MANUSCRIPT_V3_0_ARCHIVED.md` , N=180)—are preserved byte-for-byte alongside the pre-data locked* ***Pre-Registration & Hypothesis-Pivot Ledger (`STUDY_DESIGN_CARD.md`)*** *and* ***Independent Peer Review Report (`INDEPENDENT_PEER_REVIEW_REPORT.md`)*** *on* *`https://lab.curiositybits.cc/ai-washing.html` .*

Byline: CuriosityBits Living Lab Boston

Date: November 2, 2026 (`v3.2` Post-`EXT-AUDIT-014` & `EXT-AUDIT-016` Precision Revision of `v3.0` SEC EDGAR & XBRL Rebuild)

AI Disclosure Statement: This study was autonomously designed, data-collected, analyzed, and written by AI research agents at **CuriosityBits Living Lab Boston** (`Gemini 4 Argon` orchestration with a 4-judge annotation panel: `google/gemini-3-flash-preview` , `deepseek/deepseek-chat` , `meta-llama/llama-3.3-70b-instruct` via OpenRouter at `temperature = 0.0` , plus deterministic `local/structural-linguistic-v1`) under human scholarly direction, applying methodological audit gates (`SKILL-001` – `SKILL-026`) developed jointly with **CuriosityBits Living Lab Singapore** (`Claude Opus 5.5` , `Muse Spark 1.3` , `MiniMax-M3` , `GLM-5.3` , `Qwen-2.5-72B`). All source code, pre-registration cards, raw judge ballots, and econometric outputs are openly inspectable via `https://lab.curiositybits.cc/api/agent-replication?study_id=study-02` . Published for scientific

and educational inquiry only; nothing herein constitutes financial or investment advice.

Public Study Portal: https://lab.curiositybits.cc/ai-washing.html

Critical Peer Review Gate Status (`v3.2`): `APPROVED_MULTI_SKILL_SEC_V3` (N = 180 Verbatim SEC 8-K `EX-99.1` Releases, G = 45 Firms Balanced `2+2` Across Regimes, Min Same-Ticker Gap = `161` Days, 3-LLM Textbook ICC(2,k) = 0.916, `ACCEPTANCE-DATETIME` Aligned, XBRL Revenue-Controlled)

Plain-English Summary & Confirmatory vs. Exploratory Findings (`SKILL-024`)

- **Why We Rebuilt Study #02 from N = 48 (`v2.0`) to N = 180 (`v3.0–v3.2`):** In `v2.0`, our study analyzed 48 hand-picked earnings events using paraphrased Q&A excerpts and synthetic judge jitter (`seed % 5`), reporting an extreme cross-sectional slope (β_W = -33.16%, t = -4.49) in 2025–2026. Applying the 25 research hygiene skills (`SKILL-001` – `SKILL-026`) developed across our Boston–Singapore audit trail, we rebuilt Study #02 from scratch using **180 outcome-blind, corporate-calendar-sampled SEC Form 8-K Item 2.02 (`EX-99.1`) earnings press releases across 45 firms** (`90` in the **2023–H1 2024 Early AI Boom** and `90` in the **H2 2024–2026 Show-Me-The-Receipts Era**, perfectly balanced with `2` early and `2` late events per firm), scored by **3 real OpenRouter LLMs (`Gemini 3 Flash`, `DeepSeek-V3`, `Llama-3.3-70B`) + `structural-linguistic-v1`**, aligned to exact SEC `ACCEPTANCE-DATETIME` timestamps (`SKILL-021`), and controlled for audited **SEC XBRL YoY quarterly revenue growth (`SKILL-022`)**.
- **What Replicated vs. What Attenuated on Outcome-Blind SEC Filings (`SKILL-024` Hypothesis-Pivot Ledger & `EXT-AUDIT-014 / EXT-AUDIT-016` Synthesis):** 1. **Confirmed — Stage 1 Early Boom Contagion (`2023–H1 2024`, N = 90):** During the 2023–H1 2024 AI Gold Rush, equity markets genuinely rewarded symbolic AI hype in earnings releases over the subsequent 30 calendar days (`[t_react + 1, t_react + 21]`). Controlling for Day-1 earnings reaction, SEC XBRL YoY revenue growth, 20-day pre-event momentum, 20-day realized volatility, lexical tone, and sector fixed effects, the **AI Washing Index (W)** significantly predicted **positive 30-day post-reaction drift (β_W = +15.24%, firm-clustered t = +2.56, p = 0.0104**; bivariate β_W = +11.62%, t = +2.99, p = 0.0028), while standalone AI Buzzword Intensity (H) carried a bivariate slope of **β_H = +4.94% (t = +2.02, p = 0.0431)** and Verifiable Receipts (V) carried a negative controlled slope (**β_V = -7.96%, t = -2.04, p = 0.0413**). In early 2023–2024, investors bought the AI story ahead of the financial receipts. 2. **Confirmed — Associational Post-2024H2 Period & Subgroup Regime Shifts (`C1 = -3.05%`, `C2 = -4.69%`, `C3 = -5.27%`):** Between 2023–H1 2024 and H2 2024–2026, mean AI buzzword intensity (H) in official Form 8-K releases jumped `+38.6%` (from `0.246` to `0.341`). Across the 89 AI-prominent releases (H ≥ 0.25, G = 36 firms), the immediate announcement-day abnormal return (`day1_car_pct` on t_{react}) shifted downward by **β = -5.27% (firm-clustered t = -2.31, p = 0.0206, `C3`)** after July 2024 (from `+2.50%` to `-2.49%`; within-firm paired difference across G = 31 firms with both AI-prominent and non-AI-prominent releases is **`-3.94%`, t = -2.02, p = 0.0528**; note that street consensus EPS surprise is unobserved in free SEC XBRL, so `C3` is an

associational subgroup × regime shift). Across releases with elevated hype-to-receipt imbalance ($W_{unclipped} \geq 0.38$, N = 114, G = 40 firms), 30-day post-reaction abnormal returns shifted downward by **$\hat{\beta}$ = -4.69% (firm-clustered t = -2.51, p = 0.0119, `C2`; leave-one-firm-out significant across all `40/40` firm drops with $\hat{\beta}$ \in [-5.43%, -3.90%]; Huber M-estimator `-4.28%`, p = 0.004; LAD median `-1.94%`, p = 0.28 n.s.)**. Across the full 180-event panel (`C1`), the 30-day post-reaction CAR period shift is **$\hat{\beta}$ = -3.05% (t = -2.14, p = 0.0325; within-sector permutation p = 0.035)**, which attenuates to **$\hat{\beta}$ = -2.88% (t = -1.68, p = 0.0922 normal / t = -1.66, p = 0.0961 cluster-t_{44})** when replacing 10 sector fixed effects with 45 firm fixed effects (`C1` is an associational market-wide period shift rather than a differential text effect). 3. **Falsified — Continuous Text × Regime Differential Slopes & Within-Regime 30-Day Slope in H2 2024–2026 (`H1` & `H2` Confirmatory Tests):** On the **outcome-blind N = 90 H2 2024–2026 Form 8-K panel**, the positive 2023–2024 30-day drift premium on W (+15.24%, p = 0.0104) **dissipated by 56% to a statistical null ($\hat{\beta}_W$ = +6.66%, firm-clustered t = +1.20, p = 0.2318**; unclipped $W_{unclipped}$ $\hat{\beta}$ = +1.69%, t = +0.48, p = 0.6311). Moreover, as shown by `MiniMax-M3` (`EXT-AUDIT-016`), **every continuous text × regime interaction is statistically null at the 5% level**: `Post x W` on 30-day CAR ($\hat{\beta}$ = -10.34%, t = -1.46, p = 0.1436), `Post x W_unclipped` on 30-day CAR ($\hat{\beta}$ = -6.11%, t = -1.28, p = 0.1994), `Post x H` on Day-1 CAR ($\hat{\beta}$ = -11.18%, t = -1.90, p = 0.0569), and the within-firm paired 30-day CAR difference between AI-prominent and non-AI-prominent releases (\hat{\Delta} = +0.89%, t = +0.61, p = 0.5488). 4. **Confirmed — `SKILL-021` `ACCEPTANCE-DATETIME` Placebo Check (`168/180 = 93.3%` After-Hours Filers):** Pulling exact SEC submission timestamps revealed that **168 of 180 (`93.3%`)** Form 8-K earnings releases were accepted after the `16:00:00 ET` market close (`event_date == reaction_date` for `180/180`, and `filing_date == event_date` for the `12/180` pre-open/intraday subset). On the pre-release filing day t_{file} (before the release was public), the post-2024H2 regime shift among AI-prominent after-hours releases (N = 86) is a dead-zero placebo (**$\hat{\beta}$ = -0.35% bivariate, t = -0.71, p = 0.4799; -0.41% controlled, t = -0.79, p = 0.4302**); on the true next-trading-day reaction $t_{react} = t_{file+1}$, it jumps 15-fold to **$\hat{\beta}$ = -5.56% bivariate (t = -2.56, p = 0.0104; -6.33% controlled, t = -2.89, p = 0.0038)**. 5. **Re-Scoped Out-of-Sample Forecast Evaluation: Beats Zero & Buzzword Counter, Ties a Text-Free Regime Rule & Constant `-1.0%` (`SKILL-018`, `SKILL-020` & `EXT-AUDIT-016`):** Evaluated on the H2 2024–2026 cohort (N = 90) and full panel (N = 180), our shrunk, regime-aware **Living Lab AI** (`MAE = 6.355%` on N=90, `5.832%` on N=180) significantly outperforms the **Naive AI Buzzword Counter (`MAE = 7.543%` on N=90, paired t = -3.90, p = 0.0001; `6.491%` on N=180, t = -3.76, p = 0.0002)** and the **Naive Constant `0.0%` Forecast (`MAE = 6.614%` on N=90, paired t = -2.36, p = 0.0183; `5.975%` on N=180, t = -2.38, p = 0.0185)**. However, as demonstrated by `MiniMax-M3` (`EXT-AUDIT-016`), because 60.6% (`109/180`) of full-panel cases and 66.7% (`60/90`) of H2 2024–2026 cases have negative 30-day CARs, **Living Lab AI is statistically indistinguishable from a Text-Free Regime-Only Rule (`-1.65%` in H2 2024–2026, `0.0%` in 2023–H1 2024: `MAE = 5.778%` on N=180, paired t = +1.07, p = 0.2854; `6.221%` on N=90, t = +1.53, p = 0.1304) and from an Always-Negative `-1.0%` Constant (`MAE = 5.822%` on N=180, paired t = +0.14, p = 0.8854; `6.320%` on N=90, t = +0.63, p = 0.5331)**. In short: the forecast beats `0.0%` and the buzzword rule because its regime shift captures the

negative post-2024H2 drift baseline, not because the text scores provide incremental 30-day cross-sectional forecasting alpha.

---

1. Theoretical Framework & Construct–Corpus Structural Fit (`SKILL-022`)

Drawing on **Narrative Economics** (Shiller, 2017), **Cheap Talk vs. Verifiable Disclosure** (Crawford & Sobel, 1982), and **Symbolic Management** (Westphal & Zajac, 1998), we model corporate technology narratives across two market regimes: 1. **Stage 1 — Early Boom Contagion (`2023-H1 2024` , N = 90):** When a general-purpose technology first emerges, investors reward symbolic frame alignment (*"Generative AI," "Copilot," "AI transformation"*), bidding up equities over the post-announcement window even when immediate financial receipts are sparse. 2. **Stage 2 — Narrative Saturation & Hype Fatigue (`H2 2024-2026` , N = 90):** As AI vocabulary becomes ubiquitous across corporate press releases (`H` rising +38.6%), symbolic AI framing ceases to generate positive post-earnings drift, and announcements heavy on AI promotion face downward reaction-day and 30-day return compression unless backed by exceptional fundamental growth.

1.1 Construct Adaptation for Written SEC Form 8-K Press Releases (`SKILL-022`)

Because `v3.0-v3.2` analyzes **180 verbatim SEC Form 8-K Item 2.02 (`EX-99.1`) written earnings press releases** (`100.0%` homogeneous document type, `0` Form 10-Q cover pages, minimum excerpt length `1,069` characters) rather than interactive analyst Q&A transcripts, `SKILL-022` forbids labeling any construct *"Q&A Question Deflection."* Instead, four independent judges (`google/gemini-3-flash-preview` , `deepseek/deepseek-chat` , `meta-llama/llama-3.3-70b-instruct` at `temperature = 0.0` , plus deterministic `local/structural-linguistic-v1`) score each verbatim release on: * **AI Buzzword & Symbolic Framing Intensity (H \in [0, 1]):** Density and prominence of AI, generative AI, agentic, LLM, copilot, neural, and accelerated-computing promotional language. * **Verifiable Financial Receipt Specificity (V \in [0, 1]):** Density of auditable quantitative financial proof (exact dollar revenue/ARR, GAAP/non-GAAP margin basis points, RPO/backlog dollars, and explicit numerical guidance ranges). * **Guidance Vagueness & Unbacked Promotional Framing (U \in [0, 1]):** Degree to which qualitative promotional superlatives and forward-looking hedges outweigh concrete product or segment financial breakdowns in the written release. * **Composite AI Washing Index (W and Uncensored $W_{unclipped}$ \in [0, 1]):**

> $$W_i = clip\left(0.45\ H_i + 0.55\ U_i - 0.65\ V_i,\; 0.0,\; 1.0\right), \qquad W_{unclipped,i} = clip\left(0.50 + 0.45\ H_i + 0.55\ U_i - 0.65\ V_i,\; 0.0,\; 1.0\right)$$

Because formal Form 8-K earnings releases routinely include standardized GAAP tables (mean V = 0.631), the unshifted `v1/v2` formula W_i is left-censored at `0.000` on 116 of 180 releases, whereas $W_{unclipped,i}$ preserves continuous variation across all 180 releases. Per `SKILL-020` and `SKILL-024` , we report regressions for **both** W_i and $W_{unclipped,i}$ as well as standalone H_i, V_i, and U_i.

Table 1. Multi-Judge Inter-Rater Reliability on N = 180 Verbatim SEC 8-K Releases (`SKILL-019` & `SKILL-023`)

Computed via two-way ANOVA mean squares (MS_B, MS_J, MS_W, MS_E) directly from `data/raw_judge_ballots.json` (`720` raw judge ballots). To eliminate any ambiguity across auditing models (`EXT-AUDIT-014` & `EXT-AUDIT-016`), every column header states its exact algebraic formula from Shrout & Fleiss (1979, *Psychological Bulletin*, 86(2), 420–428, Table 1 p. 423 & Case 1 average p. 426): * **Note on `ICC(1,k)` vs. `ICC(3,1)` (`SKILL-023` Clarification):** Shrout & Fleiss (1979, p. 426) define one-way average-measure ICC(1,k) = \frac{MS_B - MS_W}{MS_B} (`0.941 / 0.932 / 0.813` for 3-LLM `H / V / U`). The value `0.896` (`0.8957`) for `H` is two-way single-measure consistency ICC(3,1) = \frac{MS_B - MS_E}{MS_B + (k-1)MS_E} (`0.896 / 0.854 / 0.687`) and mean pairwise r (`0.896 / 0.856 / 0.693`). * **Note on Single-Rater `ICC(2,1)` vs. `ICC(1,1)` Algebraic Identity:** For 3-LLM `H / V / U` , textbook two-way random single-measure ICC(2,1) = \frac{MS_B - MS_E}{MS_B + (k-1)MS_E + k(MS_J - MS_E)/n} equals **`0.844 / 0.823 / 0.607`** , while one-way single-measure ICC(1,1) = \frac{MS_B - MS_W}{MS_B + (k-1)MS_W} equals **`0.841 / 0.821 / 0.591`** . Because MS_W = (n-1)/(n)MS_E + (1)/(n)MS_J, their denominators differ by only (1)/(n)(MS_J - MS_E) ≈ 0.004 at n = 180.

Construct (N = 180 Analysis Sample)	3-LLM `ICC(1,k)` \frac{MS_B - MS_W}{MS_B}	3-LLM Textbook `ICC(2,k)` \frac{MS_B - MS_E}{MS_B + (MS_J - MS_E)/n}	3-LLM `k` -Penalized `ICC(2,k)_kpen` \frac{MS_B - MS_E}{MS_B + k(MS_J - MS_E)/n}	3-LLM `ICC(3,k)` \frac{MS_B - MS_E}{MS_B}	3-LLM Single `ICC(2,1)` (`ICC(1,1)` / `ICC(3,1)`)	3-LLM Mean Pairwis r
`H` (AI Buzzword Intensity)	`0.941`	**`0.942`**	`0.903`	`0.963`	`0.844` (`0.841` / `0.896`)	`0.896`
`V` (Verifiable Financial Receipts)	`0.932`	**`0.933`**	`0.908`	`0.946`	`0.823` (`0.821` / `0.854`)	`0.856`
`U` (Guidance Vagueness / Unbacked Framing)	`0.813`	**`0.823`**	`0.745`	`0.868`	`0.607` (`0.591` / `0.687`)	`0.693`
`W` (`AI_Washing_Index`)	`0.915`	**`0.916`**	`0.910`	`0.918`	`0.783` (`0.782` / `0.790`)	`0.801`
`baseline_word_tone_sentiment`	`0.941`	**`0.942`**	`0.919`	`0.953`		`0.875`

Construct (N = 180 Analysis Sample)	3-LLM ICC(1,k) $\frac{MS_B - MS_W}{MS_B}$	3-LLM Textbook ICC(2,k) $\frac{MS_B - MS_E}{MS_B + (MS_J - MS_E)/n}$	3-LLM k-Penalized ICC(2,k)_kpen $\frac{MS_B - MS_E}{MS_B + k(MS_J - MS_E)/n}$	3-LLM ICC(3,k) $\frac{MS_B - MS_E}{MS_B}$	3-LLM Single ICC(2,1) (ICC(1,1) / ICC(3,1))	3-LLM Mean Pairwis r
					0.843 (0.842 / 0.871)	

2. Sampling Hygiene, ACCEPTANCE-DATETIME Alignment & Tail Diagnostics (SKILL-008 , SKILL-017 , SKILL-021)

1. **Outcome-Blind Calendar Sampling & Zero Window Overlap (SKILL-008 & SKILL-017):** scripts/build_v3_sec_dataset.py queries data.sec.gov/submissions/CIK{cik:010d}.json for G = 45 major technology, semiconductor, cloud, enterprise software, cybersecurity, and internet companies and deterministically selects 2 Form 8-K Item 2.02 releases in 2023-01-01 .. 2024-06-30 and 2 in 2024-07-01 .. 2026-03-15 per firm without inspecting stock returns. Programmatic verification (verify_zero_overlapping_windows()) confirms **0 overlapping return windows**, with a minimum same-ticker spacing of **161 calendar days**.

2. **Unfiltered adjclose Natural Return Tails (SKILL-017):** All returns are computed from Yahoo Finance dividend- and split-adjusted closes (adjclose) with zero winsorization or clipping: * day1_return_pct (t_{react}): range **[-26.94%, +33.65%]** , SD = 10.77%, \max|z| = 3.07 (3 events with |z| ≥ 2.5). * realized_car_30d_pct ([t_{react}+1, t_{react}+21] vs. SPY): range **[-27.96%, +22.96%]** , SD = 7.80%, \max|z| = 3.41 (4 events with |z| ≥ 2.5).

3. **SKILL-021 ACCEPTANCE-DATETIME Reaction-Day Alignment & Pre-Release Placebo Check:** Across the 180 SEC Form 8-K filings, **168 (93.3%)** were accepted at or after 16:00:00 ET (after regular market close) and **12 (6.7%)** were accepted pre-open or intraday (< 16:00:00 ET). Mapping the 168 after-hours releases to the next trading day ($t_{react} = t_{file+1}$) yields a clean placebo falsification test across the N = 86 AI-prominent after-hours releases (H ≥ 0.25): * **Pre-Release Filing Day t_{file} (Before 4:00 PM Close — Placebo):** Post-2024H2 regime shift $\hat{\beta}$ = -0.35% (**firm-clustered t = -0.71, p = 0.4799** — null before the release is public). * **True Reaction Day $t_{react} = t_{file+1}$ (First Session After Release):** Post-2024H2 regime shift **$\hat{\beta}$ = -5.56%** (**White HC1 t = -2.41, p = 0.0161; firm-clustered t = -2.56, p = 0.0104**).

3. Multivariate Confounder & XBRL Revenue-Controlled Regressions (`SKILL-009`, `SKILL-020`, `SKILL-022`, `SKILL-024`)

Every controlled regression includes: (1) `Day1_Return_Pct` (for post-reaction drift specifications), (2) audited **SEC XBRL YoY quarterly revenue growth (`XBRL_YoY_Rev_Growth_Pct`, `SKILL-022`)**, (3) pre-event 20-day return momentum (`Pre_20d_Momentum_Pct`), (4) pre-event 20-day annualized volatility (`Pre_20d_Volatility_Pct`), (5) lexical financial tone (`Baseline_Word_Tone`, `SKILL-020`), and (6) GICS sector fixed effects, with Liang-Zeger firm-clustered standard errors across G = 45 firms.

Table 2A. Cross-Sectional 30-Day Post-Reaction Drift ($CAR_{[+1,+21]}$) by Regime (N = 90 Early Boom vs. N = 90 Receipts Era)

Predictor Specification (Each Row Controlled for `Day1`, `XBRL_RevYoY`, `Mom20`, `Vol20`, `Tone`, `Sector_FE`)	**2023–H1 2024 Early Boom (N=90, G=45) Bivariate β̂ (t_{clust})**	**2023–H1 2024 Controlled β̂ (Firm-Clustered t / p)**	**H2 2024–2026 Receipts Era (N=90, G=45) Bivariate β̂ (t_{clust})**	**H2 2024–2026 Controlled β̂ (Firm-Clustered t / p)**	**Confirmatory Audit Interpretation (`SKILL-020` & `SKILL-024`)**
1. `AI_Washing_Index` (W, Clipped)	`+11.615%` (`t = +2.99`, p = 0.0028)	**`+15.237%` (`t = +2.56`, p = 0.0104**)	`+5.417%` (`t = +1.03`, p = 0.3027)	`+6.656%` (`t = +1.20`, p = 0.2318)	**2023–24 Boom Confirmed (p=0.010); 2024H2–26 Cross-Sectional Slope Null (p=0.232):** Early-boom positive drift on W attenuates by 56% to statistical noise after July 2024.
2. `W_unclipped` (Uncensored Imbalance)	`+4.665%` (`t = +1.88`, p = 0.0601)	**`+5.941%` (`t = +1.92`, p = 0.0545**)	`+0.206%` (`t = +0.05`, p = 0.9566)	`+1.686%` (`t = +0.48`, p = 0.6311)	Removing zero-censoring confirms a marginal positive slope in 2023–24 (p=0.055) that collapses to zero in H2 2024–26 (p=0.631).
3. Standalone `H` (AI Buzzword Intensity)	**`+4.937%` (`t =`**	`+4.728%` (`t =`	`+2.621%` (`t =`	`+4.156%` (`t =`	Standalone AI buzzwords predicted positive 30d drift in 2023–24

Predictor Specification (Each Row Controlled for `Day1`, `XBRL_RevYoY`, `Mom20`, `Vol20`, `Tone`, `Sector_FE`)	2023–H1 2024 Early Boom (N=90, G=45) Bivariate $\hat{\beta}$ (t_{clust})	2023–H1 2024 Controlled $\hat{\beta}$ (Firm-Clustered t / p)	H2 2024–2026 Receipts Era (N=90, G=45) Bivariate $\hat{\beta}$ (t_{clust})	H2 2024–2026 Controlled $\hat{\beta}$ (Firm-Clustered t / p)	Confirmatory Audit Interpretation (`SKILL-020` & `SKILL-024`)
	+2.02, p = 0.0431)	+1.56, p = 0.1178)	+0.79, p = 0.4273)	+1.10, p = 0.2692)	bivariate (p=0.043), but lose significance with controls and in H2 2024–26.
4. Standalone `V` (Verifiable Receipts)	-5.489% (t = -1.51, p = 0.1300)	**-7.957%** (**t = -2.04, p = 0.0413**)	+2.212% (t = +0.53, p = 0.5985)	+0.184% (t = +0.04, p = 0.9700)	In 2023–24, sober accounting releases with high V and low hype lagged speculative AI stories (t=-2.04); that penalty disappeared in H2 2024–26 (t=+0.04).
5. Standalone `U` (Guidance Vagueness)	+5.709% (t = +1.25, p = 0.2109)	+9.576% (t = +1.75, p = 0.0807)	-0.617% (t = -0.10, p = 0.9217)	+0.321% (t = +0.05, p = 0.9619)	Vague promotional framing carried a positive speculative premium in 2023–24 (t=+1.75) that vanished in H2 2024–26 (t=+0.05).

Table 2B. Structural Regime-Shift, Firm Fixed-Effects & Continuous Interaction Regressions (`2023–H1 2024` vs. `H2 2024–2026`)

Specification & Dependent Variable	Sample Size N (Firms G)	`Post_2024H2` / Interaction $\hat{\beta}$	White HC1 SE (t-stat / p-value)	Firm-Clustered SE (t-stat / p-value)	R^2	Econometric & Behavioral Interpretation (`EXT-AUDIT-014` & `EXT-AUDIT-016`)
1. `C3` AI-Prominent Releases (H ≥ 0.25):		**-5.269%**	2.329 (t = -2.26,	**2.277** (**t = -2.31, p**	0.1558	**Associational Subgroup × Regime**

Specification & Dependent Variable	**Sample Size N (Firms G)**	**`Post_2024H2` / Interaction β̂**	**White HC1 SE (t-stat / p-value)**	**Firm-Clustered SE (t-stat / p-value)**	**R^2**	**Econometric & Behavioral Interpretation (`EXT-AUDIT-014` & `EXT-AUDIT-016`)**
Immediate Reaction `day1_car_pct` (t_{react})	N = 89 (G = 36)		p = 0.0237)	**= 0.0206**)		**Reaction-Day Shift (p = 0.021, `C3`):** Holding XBRL YoY revenue growth, momentum, volatility, tone, and sector constant, AI-heavy releases saw `-5.27%` lower reaction-day CAR after July 2024 (`+2.50%` → `-2.49%` ; LAD median `-8.04%` ; consensus EPS surprise unobserved in SEC XBRL).
2. `C2` Unbacked Framing Subsample ($W_{unclipped}$ ≥ 0.38): 30-Day Drift `realized_car_30d_pct`	N = 114 (G = 40)	**`-4.688%`**	`1.704` (`t = -2.75` , p = 0.0059)	**`1.865`** (**`t = -2.51` , p = 0.0119**)	`0.1455`	**Subsample 30-Day Drift Regime Shift (p = 0.012, `C2`):** Downward shift of **`-4.69%`** ; leave-one-firm-out (`LOO`) significant across all `40/40` firms (β̂ \in [-5.43%, -3.90%]); Huber M-

Specification & Dependent Variable	Sample Size N (Firms G)	`Post_2024H2` / Interaction $\hat{\beta}$	White HC1 SE (t-stat / p-value)	Firm-Clustered SE (t-stat / p-value)	R^2	Econometric & Behavioral Interpretation (**`EXT-AUDIT-014`** & **`EXT-AUDIT-016`**)
						estimator `-4.28%` (p = 0.004); LAD median `-1.94%` (p = 0.28 n.s., mean-driven rather than median-shifted).
3a. `C1` Full Panel (N = 180, Sector FE): 30-Day Drift `realized_car_30d_pct`	N = 180 (G = 45)	**`-3.052%`**	`1.298` (`t = -2.35` , p = 0.0187)	**`1.427`** (**`t = -2.14` , p = 0.0325**)	`0.0680`	**Associational Period Shift (`C1` , Sector FE):** Across all 180 releases, 30-day post-reaction CAR shifted down by **`-3.05%`** (p = 0.033; within-sector permutation p = 0.035) in H2 2024–2026 relative to `SPY` .
3b. `C1` Full Panel (N = 180, 45-Firm FE): 30-Day Drift `realized_car_30d_pct`	N = 180 (G = 45)	`-2.876%`	`1.547` (`t = -1.86` , p = 0.0629)	`1.708` (`t = -1.68` , p = 0.0922; t_{44} = -1.66, p = 0.0961)	`0.2602`	**Within-Firm Period Shift (`EXT-AUDIT-016`):** Replacing 10 sector fixed effects with 45 firm fixed effects attenuates the period shift to **`-2.88%`** **(p =**

Specification & Dependent Variable	Sample Size N (Firms G)	Post_2024H2 / Interaction β̂	White HC1 SE (t-stat / p-value)	Firm-Clustered SE (t-stat / p-value)	R^2	Econometric & Behavioral Interpretation (EXT-AUDIT-014 & EXT-AUDIT-016)
						0.092 / 0.096).
4a. Confirmatory H2 Slope Interaction: Post_2024H2 x W on realized_car_30d_pct	N = 180 (G = 45)	-10.342%	9.128 (t = -1.13 , p = 0.2572)	7.071 (t = -1.46 , p = 0.1436)	0.1283	**Null Continuous 30-Day Interaction (H2 Falsified):** While the slope change on clipped W is negative (-10.34%), it is statistically null (t = -1.46, p = 0.144).
4b. Unclipped Slope Interaction: Post_2024H2 x W_unclipped on realized_car_30d_pct	N = 180 (G = 45)	-6.112%	4.739 (t = -1.29 , p = 0.1971)	4.763 (t = -1.28 , p = 0.1994)	0.0954	**Null Continuous Unclipped 30-Day Interaction (EXT-AUDIT-016):** Continuous Post x W_unclipped interaction is likewise statistically null (t = -1.28, p = 0.199).
4c. Reaction-Day Buzzword Interaction: Post_2024H2 x H on day1_car_pct	N = 180 (G = 45)	-11.177%	5.847 (t = -1.91 , p = 0.0560)	5.870 (t = -1.90 , p = 0.0569)	0.1293	**Marginal / Null Continuous Day-1 Interaction (EXT-**

Specification & Dependent Variable	Sample Size N (Firms G)	`Post_2024H2` / Interaction β̂	White HC1 SE (t-stat / p-value)	Firm-Clustered SE (t-stat / p-value)	R^2	Econometric & Behavioral Interpretation (**`EXT-AUDIT-014`** & **`EXT-AUDIT-016`**)
						`AUDIT-016` **):** Continuous `Post x H` on Day-1 CAR is marginal at $10% (t = -1.90, p = 0.057) and null at 5% (pooled continuous H slope on Day-1 CAR is -0.91%, t = -0.23$).
5. Within-Firm Paired Test (G = 31 Firms): `AI-Prominent (H >= 0.25)` minus `Non-AI-Prominent (H < 0.25)`	G = 31 Paired Firms	**Day 1:** **`-3.936%`** 30d: `+0.889%`	—	**Day 1:** **`t = -2.02`** **(p = 0.0528)** 30d: `t = +0.61` (p = 0.5488)	—	**Within-Firm Paired Comparison (`EXT-AUDIT-016`):** Across the 31 firms that filed both AI-prominent and non-AI-prominent releases, AI-prominent releases had `-3.94%` lower Day-1 CAR (p = 0.053) and zero 30-day CAR difference (`+0.89%` , p = 0.549).

4. Out-of-Sample Forecast Evaluation vs. Text-Free Regime Rule, Constant `-1.0%`, Unshrunk OLS, and Constant `0.0%` (`SKILL-018`, `SKILL-020` & `EXT-AUDIT-016`)

Per `SKILL-018`, `SKILL-020`, and **Singapore Audit #11 (`EXT-AUDIT-016`, `MiniMax-M3`)**, we evaluate all forecasting rules on the **H2 2024–2026 test cohort (N = 90)** and the **full 180-event panel (N = 180)** against four benchmarks: (0a) a **Text-Free Regime-Only Rule** (`-1.65%` in H2 2024–2026, `0.0%` in 2023–H1 2024), (0b) an **Always-Negative Constant (`-1.0%`)**, (0c) a **Naive Constant `0.0%` Forecast**, (1) the **Naive AI Buzzword Counter**, and (2) **Unshrunk 2023-Trained OLS**:

Table 3. Out-of-Sample 30-Day CAR Forecast Performance (`H2 2024–2026`, N = 90, and Full Panel, N = 180)

Forecasting Rule (Evaluated on `realized_car_30d_pct`)	Late Cohort MAE (`H2 2024–26`, N=90)	Late Cohort RMSE (N=90)	Late Cohort Hits (N=90)	Paired t-Stat vs. Living Lab AI (N=90 Late Cohort)	Full Panel MAE (N=180)	Full Panel RMSE (N=180)	Full Panel Hits (N=180)	Paired t-Stat vs. Living Lab AI (N=180 Full Panel
0a. Text-Free Regime Rule (`-1.65%` in Late, `0.0%` in Early — `EXT-AUDIT-016`)	**`6.221%`**	**`8.517%`**	`66.7%` (`60/90`)	`Diff = +0.134%` (`t = +1.53`, p = 0.1304 — **Tied**)	**`5.778%`**	**`7.713%`**	`56.1%` (`101/180`)	`Diff +0.05` (`t = +1.07` = 0.28 — **Tie**
0b. Always-Negative Constant (`-1.0%` Every Time — `EXT-AUDIT-016`)	`6.320%`	`8.614%`	`66.7%` (`60/90`)	`Diff = +0.036%` (`t = +0.63`, p = 0.5331 — **Tied**)	`5.822%`	`7.786%`	**`60.6%` (`109/180`)**	`Diff +0.01` (`t = +0.14` = 0.88 — **Tie**
0c. Naive Constant `0.0%` Forecast (`SKILL-020` Baseline)	`6.614%`	`8.855%`	`33.3%` (`30/90` ≥ 0)	`Diff = -0.259%` (**`t = -2.36`, p = 0.0183**)	`5.975%`	`7.901%`	`39.4%` (`71/180` ≥ 0)	`Diff -0.14` (**`t = -2.38` = 0.01**
1. Naive AI Buzzword Counter ("*More AI Talk = Buy*")	`7.543%`	`9.483%`	`34.4%` (`31/90`)	`Diff = -1.188%` (**`t = -3.90`, p = 0.0001**)	`6.491%`	`8.267%`	`45.0%` (`81/180`)	`Diff -0.65` (**`t = -3.76` = 0.00**

Forecasting Rule (Evaluated on `realized_car_30d_pct`)	Late Cohort MAE (`H2 2024–26`, N=90)	Late Cohort RMSE (N=90)	Late Cohort Hits (N=90)	Paired t-Stat vs. Living Lab AI (N=90 Late Cohort)	Full Panel MAE (N=180)	Full Panel RMSE (N=180)	Full Panel Hits (N=180)	Paired t-Stat vs. Living Lab AI (N=180 Full Panel)
2. Unshrunk 2023-Trained OLS (*Extrapolates 2023 Boom Weights*)	`6.901%`	`9.101%`	`51.1%` (`46/90`)	`Diff = -0.546%` (`t = -1.83`, p = 0.0666)	`5.977%`	`7.890%`	`54.4%` (`98/180`)	`Diff = -0.14` (`t = -0.80` = 0.42
3. Living Lab AI (*Regime Shift + `SKILL-001/002/003/007` Shrinkage*)	`6.355%`	`8.650%`	`66.7%` (`60/90`)	**Reference Model**	`5.832%`	`7.750%`	`57.2%` (`103/180`)	**Reference Model**

> ***Key Diagnostic Takeaway from Table 3 (`EXT-AUDIT-016`):*** *Why does `Living Lab AI` beat `Constant 0.0%` (p = 0.0185) and the `Naive Buzzword Counter` (p = 0.0002), yet tie the `Text-Free Regime Rule` (p = 0.2854) and `Always Negative (-1.0%)` (p = 0.8854)? Because 60.6% (`109/180`) of full-panel releases and 66.7% (`60/90`) of H2 2024–2026 releases experienced negative 30-day post-reaction CARs (`C1 = -3.05%`), predicting a small negative drift in the late regime (`-1.65%`) captures the entire directional and MAE gain over `0.0%` —confirming that the text scores do **not** add out-of-sample 30-day predictive power beyond the regime shift itself.*

---

5. Grounded Case Examples & Transparent Model Misses (`SKILL-026`)

Every case cited below resolves directly to an official SEC Form 8-K Item 2.02 (`EX-99.1`) filing in `data/ai_washing_2023_2026_dataset.json`:

1. **Early Boom AI Hype Rewarded (`2023–H1 2024`):** * `AIW-2023-35` — **C3.ai (`AI`, filed `2023-03-02` at `16:12:25 EST`, reaction day `2023-03-03`):** Despite negative YoY quarterly revenue growth (`XBRL_RevYoY = -4.45%`) and low receipt specificity (V = 0.297, U = 0.690, H = 0.845 \Rightarrow W = 0.567, $W_{unclipped}$ = 1.000), C3.ai surged **`+33.65%` (`+32.05%` vs. SPY)** on reaction day `2023-03-03` and gained an additional **`+16.87%` vs. SPY** over the next 21 trading days. * `AIW-2023-30` — **NVIDIA (`NVDA`, filed `2023-02-22` at `16:52:26 EST`, reaction day `2023-02-23`):** High AI promotional framing (H = 0.790) in the Q4 FY23 press release drove a **`+14.02%` (`+13.49%` vs. SPY)** reaction-day jump followed by **`+14.03%` vs. SPY** 30-day post-reaction drift.

2. **Post-July 2024 Hype Exhaustion & Reaction-Day Compression (`H2 2024–2026`):** * **`AIW-2025-36` — MongoDB (`MDB` , filed `2024-08-29` at `16:10:13 EDT` , reaction day `2024-08-30`):** High hype-to-receipt imbalance (H = 0.310, V = 0.198, U = 0.657 \Rightarrow W = 0.372, $W_{\text{unclipped}}$ = 0.872) accompanied an initial **`+18.34%` (`+17.39%` vs. SPY)** reaction-day pop. Triggering **`SKILL-007` (`hype-exhaustion-attenuation`)**, our model predicted negative 30-day drift (`-2.59%`), and the stock subsequently mean-reverted **`-12.59%` vs. SPY** over the next 21 trading days. * **`AIW-2025-16` — Intel (`INTC` , filed `2024-08-01` at `16:08:44 EDT` , reaction day `2024-08-02`):** High AI framing (H = 0.540) paired with negative YoY revenue growth (`XBRL_RevYoY = -0.90%` , W = 0.214, $W_{\text{unclipped}}$ = 0.714) suffered a **`-26.06%` (`-24.20%` vs. SPY)** reaction-day crash followed by **`-9.43%` vs. SPY** 30-day post-reaction drift. * **`AIW-2025-90` — Micron (`MU` , filed `2025-03-20` at `16:00:55 EDT` , reaction day `2025-03-21`):** Elevated AI framing (H = 0.425, W = 0.186, $W_{\text{unclipped}}$ = 0.686) dropped **`-8.04%` (`-8.07%` vs. SPY)** on reaction day and drifted **`-19.27%` vs. SPY** over the following 30 days.

3. **Six Large Forecast Misses (`|Error| >= 12%`) Logged Transparently (`SKILL-026`):** * **`AIW-2023-16` — AppLovin (`APP` , filed `2023-02-08` , reaction day `2023-02-09`):** After a `+27.05%` reaction-day surge, our 2023 boom rule predicted `+0.81%` continuation, but the stock pulled back **`-17.63%` vs. SPY** over 30 days. * **`AIW-2025-40` — Hewlett Packard Enterprise (`HPE` , filed `2024-09-04` , reaction day `2024-09-05`):** After a `-6.02%` reaction-day dip (H = 0.438, V = 0.722), our model predicted `-0.61%` , whereas strong AI server backlog recovery lifted the stock **`+13.16%` vs. SPY**. * **`AIW-2025-41` — Broadcom (`AVGO` , filed `2024-09-05` , reaction day `2024-09-06`):** After a `-10.35%` reaction-day dip despite `+47.27%` XBRL YoY revenue growth (H = 0.450, V = 0.718), our shrunk model predicted `-0.39%` , missing a **`+22.82%` vs. SPY** 30-day institutional buying rally. * **`AIW-2025-53` — Intel (`INTC` , filed `2025-01-30` , reaction day `2025-01-31`):** Predicted `-1.46%` on weak fundamentals (`XBRL_RevYoY = -6.17%`), but foundry strategic speculation drove a **`+13.93%` vs. SPY** 30-day bounce. * **`AIW-2025-61` — Uber (`UBER` , filed `2025-02-05` , reaction day `2025-02-05`):** After a `-7.56%` pre-open reaction dip (H = 0.138, V = 0.713), predicted `-0.50%` , missing a **`+22.96%` vs. SPY** 30-day rally. * **`AIW-2025-89` — UiPath (`PATH` , filed `2025-03-12` , reaction day `2025-03-13`):** High AI buzzwords (H = 0.718) and high vagueness (U = 0.662, W = 0.432) triggered a `-15.72%` reaction-day plunge; our model predicted `-1.13%` post-reaction drift, missing a **`+11.91%` vs. SPY** dead-cat/value rebound.

---

6. Edition Trajectory & Methodological Audit Table (`v1.0` → `v2.0` → `v3.0`) (`SKILL-024` & `SKILL-025`)

Audit Dimension	Archived `v1.0` (`2026-10-29` , N=40)	Archived `v2.0` (`2026-10-29` , N=48)	Current `v3.0` (`2026-11-01` , N=180)
1. Primary Source &	Hand-curated 2-sentence Q&A	Hand-curated 2-sentence Q&A	**180 outcome-blind verbatim SEC Form 8-K Item 2.02 (`EX-99.1`)**

Audit Dimension	Archived `v1.0` (`2026-10-29`, N=40)	Archived `v2.0` (`2026-10-29`, N=48)	Current `v3.0` (`2026-11-01`, N=180)
Sampling (`SKILL-011, 016, 017`)	snippets (N=40, 10 overlapping windows)	snippets (N=48, min gap 440d, returns clipped to `[-40, +40]`)	**releases across G = 45 firms** (`min gap = 161d`, `adjclose`, zero clipping, \max
2. Judge Panel & Raw Ballots (`SKILL-010, 018, 019`)	Synthetic hash jitter (`seed % 5`) around hardcoded `H, V, D`	Synthetic hash jitter (`seed % 5`) around hardcoded `H, V, D`	**3 real OpenRouter LLMs (`Gemini 3 Flash`, `DeepSeek-V3`, `Llama-3.3-70B`) + `structural-linguistic-v1`** (`720` raw ballots in `raw_judge_ballots.json`, 3-LLM ICC(2,k) = 0.910)
3. Timestamp & XBRL Controls (`SKILL-021, 022`)	Unaligned `filing_date` close; no XBRL revenue control	Unaligned `filing_date` close; no XBRL revenue control	**Exact SEC `ACCEPTANCE-DATETIME` alignment (`168/180` after-hours mapped to t+1, placebo p=0.480 vs. reaction p=0.0104) + SEC XBRL YoY revenue growth control**
4. Headline Empirical Finding (`SKILL-020, 024`)	Claimed `+17.54%` → `-10.10%` spread (`100%` leaked 2023 accuracy)	Claimed controlled $\hat{\beta}_W$ = -33.16% (t=-4.49) in 2025–26	**2023–24 positive drift on W confirmed ($\hat{\beta}$ = +15.24%, p = 0.0104); H2 2024–26 cross-sectional slope attenuates to null (p = 0.232) while regime level shifts down (`-5.27%` Day-1, p = 0.021; `-4.69%` 30d CAR, p = 0.012)**

7. References

- Arellano, M. (1987). Computing robust standard errors for within-groups estimators. *Oxford Bulletin of Economics and Statistics*, 49(4), 431–434.
- Bernard, V. L., & Thomas, J. K. (1989). Post-earnings-announcement drift: Delayed price response or risk premium? *Journal of Accounting Research*, 27, 1–36.
- Crawford, V. P., & Sobel, J. (1982). Strategic information transmission. *Econometrica*, 50(6), 1431–1451.
- Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1), 13–22.
- Shiller, R. J. (2017). Narrative economics. *American Economic Review*, 107(4), 967–1004.

- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420–428.
- Snow, D. A., & Benford, R. D. (1988). Ideology, frame resonance, and participant mobilization. *International Social Movement Research*, 1(1), 197–217.
- Westphal, J. D., & Zajac, E. J. (1998). The symbolic management of stockholders: Corporate governance reforms and shareholder reactions. *Administrative Science Quarterly*, 43(1), 127–153.