CURIOSITYBITS LIVING LAB BOSTON · STUDY #02 PRE-REGISTRATION & HYPOTHESIS-PIVOT LEDGER (V4.0.1)

Pre-Registration & Hypothesis-Pivot Ledger — Study #02: AI Washing vs. Verifiable Financial Receipts in Corporate Earnings Releases (2023–2026) (`SKILL-010`, `SKILL-021`, `SKILL-022`, `SKILL-024`, `SKILL-025`)

> ***Pre-Registration Lock & Provenance (`SKILL-024` & `SKILL-025`):*** *- **Original `v1/v2` Pre-Registration Archive:** `STUDY_DESIGN_CARD_V2_ARCHIVED.md` (`SHA-256: bb967a2c97c96bfa9fd46ce428c974cd3df9d1385e7d335b66b81c11015aca1a`) - **`v3.0` Pre-Data Amendment Locked At:** `2026-11-01T23:28:30Z` (`SHA-256: 49e0a43da24df0af052b34936a2242c8258e4313bb27bbedb5764d61aad6a8e2`, locked **prior** to running `scripts/build_v3_sec_dataset.py` and `scripts/run_ai_washing_study.py`). - **`v3.1` Post-Audit Documentation Clarification (`EXT-AUDIT-014`):** Added explicit formula documentation for `W_unclipped = clip(0.50 + 0.45 H_i + 0.55 U_i - 0.65 V_i, 0.0, 1.0)` alongside `W = clip(0.45 H_i + 0.55 U_i - 0.65 V_i, 0.0, 1.0)` per Singapore Audit #10 (`EXT-AUDIT-014`). - **`v3.2` Post-Audit Benchmark & Differential-Shift Update (`EXT-AUDIT-016`, `MiniMax-M3`):** Added a **Text-Free Regime-Only Benchmark (`-1.65%` in H2 2024–2026, `0.0%` in 2023–H1 2024)** and an **Always-Negative Constant (`-1.0%`)** to the out-of-sample forecast table, added 45-firm fixed effects for `C1`, continuous `Post x Text` interactions, and the G = 31 within-firm paired test to Table 2B, and printed exact algebraic Shrout & Fleiss (1979) formulas in every ICC column header. - **`v4.0` Post-External-Review & SG Audit Update (`DIRECTOR-FEEDBACK-S02-2026-11-02`, `SG-AUDIT-EXT-AUDIT-019-2026-11-02` & `SG-AUDIT-V4.0-EXT-AUDIT-020-2026-11-02`, authored by `Gemini 4 Argon (gemini-4-argon)`):** Preserved `STUDY_DESIGN_CARD_V3_2_ARCHIVED.md` (`SHA-256: 36c9e4e26dc7a7296d410506b92228305356cf1b2806d4dbd10a99467f96598e`) intact; documented that the claimed `v3.0` pre-data file (`SHA-256: 49e0a43da24df0af052b34936a2242c8258e4313bb27bbedb5764d61aad6a8e2`) was modified in place during `v3.1/v3.2` rather than archived separately in git (`62d9a7d`); retired `C1` (`-3.05%, t = -2.14, p = 0.0325` vs `SPY`; `-0.07%, t = -0.05, p = 0.9588` vs `XLK`; `-0.39%, t = -0.29, p = 0.7718` vs `QQQ`; `-1.21%, t = -0.82, p = 0.4135` when dropping `2025_Q1`) and `C2` (`-4.69%, t = -2.51, p = 0.0119` on `W_unclipped >= 0.38` vs `SPY`; `-1.83%, t = -1.08, p = 0.2807` vs `XLK`; `-2.17%, t = -1.22, p = 0.2227` vs `QQQ`; `-1.77%, t = -0.92, p = 0.3553` when dropping `2025_Q1`); reclassified `C3` (`-5.27%, t = -2.31, p = 0.0206` on `H >= 0.25`; `-7.65%, t = -2.33, p = 0.0197` with `Sector x Season` FE; `-5.08%, t = -1.36, p = 0.1744` with `45`-firm FE; `-2.57%, t = -0.95, p = 0.3423` when dropping `2025_Q1`) as an exploratory between-firm association concentrated in one season (`2025_Q1`); reported*

> *the `2025_Q1` within-season slope on `W` (`+8.05%, HC1 t = +1.10, p = 0.2715` without controls; `+15.44%, HC1 t = +2.38, p = 0.0172` vs `SPY` and `+14.69%, t = +2.34, p = 0.0193` vs `XLK` with controls; `+12.68%, t = +2.40, p = 0.0163` dropping the top-3 Cook's D observations); and documented the `v2.0` (`N=48` across `G=19` firms, commits `d8800a7` / `08fd4c7`) `score_four_judges()` jitter and `passed_all_reviewer_gates: True` provenance audit.* - ***Reason for `v3.0` Rebuild:*** *Applying all 25 active lab skills (`SKILL-001` – `SKILL-026`, learned from internal peer review and CuriosityBits Living Lab Singapore Red-Team Audits `EXT-AUDIT-001` – `EXT-AUDIT-011`) to replace the small (N = 48), hand-curated/paraphrased `v2.0` pilot with a statistically powered (N = 180 events across G = 45 firms), outcome-blind, primary-source **SEC EDGAR Form 8-K Item 2.02 (`EX-99.1`)** panel with exact `ACCEPTANCE-DATETIME` reaction-day alignment (`SKILL-021`), SEC XBRL fundamental revenue/earnings growth controls (`SKILL-022`), and real 4-judge LLM + structural linguistic ballots (`SKILL-018`, `SKILL-019`).*

---

1. Research Question & Theoretical Anchor

- **Primary Research Question:** Did US equity markets shift from rewarding generic AI promotional framing (H) during the **2023–H1 2024 "AI Gold Rush" regime** to penalizing unbacked AI hype (**"AI Washing"**, W) and rewarding verifiable financial receipts (V) during the **H2 2024–2026 "Show-Me-The-Receipts" regime**—and does that textual signal survive after controlling for SEC XBRL quarterly revenue growth, announcement-day return shocks, pre-event momentum/volatility, lexical negativity, and sector fixed effects?
- **Theoretical Anchors:** Shiller (2017, 2019) *Narrative Economics*; Crawford & Sobel (1982) *Cheap Talk vs. Verifiable Disclosure*; Snow & Benford (1988) *Frame Over-Extension*; Westphal & Zajac (1998) *Symbolic Management*; Bernard & Thomas (1989) *Post-Earnings Announcement Drift*.

---

2. Q1 — Exact Operational Measurement & Outcome-Blind Sampling (`SKILL-011`, `SKILL-016`, `SKILL-017`, `SKILL-021`, `SKILL-022`)

1. **Primary-Source Corpus (`SKILL-016` & `SKILL-022`):** - **100% Homogeneous Document Type:** Official **SEC EDGAR Form 8-K Item 2.02 ("Results of Operations and Financial Condition")** filings with attached **`EX-99.1` / `EX-99` quarterly earnings press releases** (`0` Form 10-Q cover pages). - **Construct–Corpus Structural Fit (`SKILL-022`):** Because SEC Form 8-K Item 2.02 (`EX-99.1`) exhibits are **written corporate earnings releases** rather than interactive analyst Q&A transcripts, we explicitly replace the `v1/v2` Q&A construct (`D = Core-Business Question Deflection`) with **`U = Guidance Vagueness & Unbacked Promotional Framing` (\in [0, 1])**—measuring the degree to which

management emphasizes qualitative promotional superlatives without concrete quantitative forward guidance or segment revenue breakdowns in the written press release.

2. **Outcome-Blind Calendar Sampling (`SKILL-011` & `SKILL-017`):** - **Universe:** G = 45 major US-listed Technology, Semiconductor, Cloud, Enterprise Software, Cybersecurity, FinTech, and Digital Platform companies with SEC CIKs and continuous 2022–2026 Yahoo Finance price history. - **Deterministic Selection Rule (Zero Return Peeking):** For each of the 45 firms, `scripts/build_v3_sec_dataset.py` queries `https://data.sec.gov/submissions/CIK{cik:010d}.json` and selects **4 non-overlapping Form 8-K Item 2.02 earnings releases** (2 in the **Early AI Boom cohort: Jan 2023 – Jun 2024** and 2 in the **Show-Me-The-Receipts cohort: Jul 2024 – Mar 2026**) separated by ≥ **75 calendar days** per ticker, yielding **N = 180 strictly non-overlapping events (N_1 = 90 in 2023–H1 2024; N_2 = 90 in H2 2024–2026)** across G = 45 firm clusters.

3. **Exact `ACCEPTANCE-DATETIME` Reaction-Day & Post-Reaction Drift Windows (`SKILL-021`):** - Every event records the exact SEC EDGAR `acceptance_datetime_utc` and Eastern Time `acceptance_datetime_et` (`America/New_York`). - **Market-Hours Release (`< 16:00:00 ET` on a trading day):** Reaction day $t_{react} = t_{file}$. - **After-Hours Release (`>= 16:00:00 ET` or non-trading day):** Reaction day $t_{react} = t_{file + 1}$ (next trading day). - **Immediate Reaction (`day1_return_pct`):** \frac{adjclose(t_{react}) - adjclose(t_{react} - 1)}{adjclose(t_{react} - 1)} × 100, plus market-adjusted `day1_car_pct` (minus `SPY` return on t_{react}). - **Pre-Release Filing-Day Placebo (`pre_release_day0_return_pct`):** For after-hours filers, \frac{adjclose(t_{file}) - adjclose(t_{file} - 1)}{adjclose(t_{file} - 1)} × 100 (`SKILL-021` falsification check). - **Post-Reaction Drift Windows (`SKILL-020` & `SKILL-021`):** Strictly forward from adjclose(t_{react}):
   - 5-trading-day post-reaction abnormal return: $CAR_{[+1,+5],i} = R_{i,[t_\{react}+1, t_{react}+5]\} - R_{SPY,[t_\{react}+1, t_{react}+5]\}$.
   - 21-trading-day (30-calendar-day) post-reaction abnormal return: $CAR_{[+1,+21],i} = R_{i,[t_\{react}+1, t_{react}+21]\} - R_{SPY,[t_\{react}+1, t_{react}+21]\}$.
   - **Zero Tail Capping (`SKILL-017`):** All returns are computed from Yahoo Finance dividend- and split-adjusted closes (`adjclose`) with **zero winsorization or clipping**.

4. **Primary Text Constructs (\in [0, 1]):** - `H` (**AI Buzzword & Symbolic Framing Intensity** \in [0, 1]): Density and prominence of AI / generative AI / agentic / LLM / neural / copilot / transformation terminology in the earnings release. - `V` (**Verifiable Financial & Operational Receipt Specificity** \in [0, 1]): Density of auditable quantitative financial commitments (explicit dollar revenue/ARR, GAAP/non-GAAP margin bps, RPO/backlog dollars, and concrete numerical guidance ranges). - `U` (**Guidance Vagueness & Unbacked Promotional Framing** \in [0, 1]): Degree to which qualitative promotional claims and forward-looking hedges outweigh concrete quantitative guidance in the release (`SKILL-022` replacement for Q&A `D`). - `W` (**Composite AI Washing Index** \in [0, 1]):

> $$W_i = clip\left(0.45\, H_i + 0.55\, U_i - 0.65\, V_i,\; 0.0,\; 1.0\right)$$

- `W_unclipped` (**Centered Continuous AI Washing Index** \in [0, 1], adding a +0.50 centering intercept so receipt-heavy releases with $0.45 H_i + 0.55 U_i - 0.65 V_i < 0$ retain continuous variation above zero rather than censoring at 0.0$):

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

- Plus standalone regressions testing H_i, V_i, and U_i individually (`SKILL-020`).

---

3. Q2 — Judge-Panel Amendment & Reliability Protocol (`SKILL-010` , `SKILL-018` , `SKILL-019` , `SKILL-024`)

- **Explicit Judge-Panel Amendment Log (`SKILL-024`):**

- In `v1/v2` (N=48), the script listed `deepseek/deepseek-v4.1-flash` , `typesafe/jev-router` , and `qwen/qwen3.8-flash` but actually generated synthetic hash-jittered scores (`seed % 5`) around hardcoded values (`SKILL-006` / `SKILL-018` violation).

- In `v3.0` (N=180), we replace synthetic hash jitter with **4 genuine, independent judge evaluations** stored in `data/raw_judge_ballots.json` (`SKILL-018`):
  1. `google/gemini-3-flash-preview` (via OpenRouter API, `temperature = 0.0`)
  2. `deepseek/deepseek-chat` (Open-Weight DeepSeek-V3 via OpenRouter API, `temperature = 0.0`)
  3. `meta-llama/llama-3.3-70b-instruct` (Open-Weight Llama-3.3-70B via OpenRouter API, `temperature = 0.0`)
  4. `local/structural-linguistic-v1` (Deterministic lexical & quantitative-receipt density analyzer)

- **Reliability Metrics (`SKILL-019`):**

- Compute standard **Shrout & Fleiss (1979) `ICC(1,k)` , `ICC(2,k)` , and `ICC(3,k)`** (using two-way ANOVA residual `MS_E` , never `MS_W`), plus mean pairwise Pearson r across the 3 LLM judges and all 4 judges on the exact N = 180 analysis sample in `data/annotation_validation.json` .

---

4. Q3 — Mandatory Confounders, Fundamental XBRL Controls & Falsification Rules (`SKILL-009`, `SKILL-020`, `SKILL-022`, `SKILL-024`)

1. **Confounders Controlled in Multivariate Regressions (`SKILL-009`, `SKILL-020`, `SKILL-022`):** - `xbrl_yoy_rev_growth_pct`: Audited YoY quarterly revenue growth rate extracted from SEC EDGAR XBRL (`https://data.sec.gov/api/xbrl/companyfacts/CIK{cik:010d}.json`) for the reported quarter (`SKILL-022` fundamental-news control). - `day1_return_pct`: Immediate announcement reaction return on t_{react} (when predicting post-reaction drift $CAR_{[+1,+21]}$). - `pre_20d_return_pct`: Pre-announcement 20-trading-day return ($[t_{react}-21, t_{react}-1]$). - `pre_20d_volatility_pct`: Pre-announcement 20-trading-day annualized realized volatility. - `baseline_word_tone_sentiment`: Lexical financial sentiment/negativity control (`SKILL-020`). - `Sector_FE`: GICS sub-industry / sector fixed effects + Liang-Zeger firm-clustered standard errors across G = 45 firms (`SKILL-008`).

2. **Confirmatory vs. Exploratory Separation & Falsification Ledger (`SKILL-024`): - Pre-Registered Primary Confirmatory Hypotheses (Inherited from `v1/v2`):**
   - **H1 (Post-2024 30-Day Drift Penalty on W):** In the H2 2024–2026 cohort (N = 90), higher `AI_Washing_Index` (W) predicts negative 21-trading-day (30-calendar-day) post-reaction drift $CAR_{[+1,+21]}$ ($\beta_W < 0$, $|t_{clustered}| \geq 1.96$) after controlling for `day1_return_pct`, `xbrl_yoy_rev_growth_pct`, `pre_20d_return_pct`, `pre_20d_volatility_pct`, `baseline_word_tone_sentiment`, and `Sector_FE`.
   - **H2 (Regime Interaction W × $Post_{2024H2}$ on 30-Day Drift):** Across the full N = 180 panel, the interaction slope $\beta_{W \times Post}$ is significantly negative ($|t_{clustered}| \geq 1.96$).
   - **Mandatory Falsification & Anti-Pivot Rule (`SKILL-024`):**
   - If H1 or H2 yields $|t_{clustered}| < 1.96$ on the outcome-blind N = 180 real SEC 8-K panel, the `v3.0` manuscript and public web feature **MUST state in the abstract and headline summary that the pre-registered 30-day post-reaction drift hypothesis from `v1/v2` was FALSIFIED / NULL (or attenuated) on outcome-blind SEC filings**.
   - Secondary tests—including (a) **Immediate Reaction-Day (`day1_car_pct` on t_{react}) effects** of W, V, H, and U controlling for `xbrl_yoy_rev_growth_pct`, and (b) **Out-of-Sample Forecast Evaluation vs. Constant-Zero (`0.0%`) Benchmark** (`SKILL-020`)—must be reported alongside H1/H2 with their exact t-statistics and p-values, clearly labeled as **Secondary / Decomposition Analyses**.

5. Hypothesis-Pivot & Edition Evolution Ledger (`v1.0` → `v2.0` → `v3.0`) (`SKILL-024`)

Edition & Date	Sample Size & Corpus	Primary Construct & Judge Setup	Primary Hypothesis Tested	Outcome & Methodological Audit Status
`v1.0` (Archived) `2026-10-29`	N = 40 hand-curated Q&A snippets (10 overlapping ticker windows)	`H, V, D, W` with synthetic hash jitter (`seed % 5`)	Unbacked AI hype flips from positive 30d drift in 2023–24 to negative in 2025–26	**REJECTED (`SKILL-005, 006, 008`):** 10 overlapping Aug–Sept 2026 return windows, ex-post `c["category"]` leakage, and no confounder controls.
`v2.0` (Archived) `2026-10-29`	N = 48 non-overlapping hand-picked/ paraphrased snippets (G = 19 firms)	`H, V, D, W` with synthetic hash jitter (`seed % 5`)	Controlled 30d CAR slope on W in 2025–26 ($\hat{\beta}$ = -33.16, t = -4.49)	**SUPERSEDED BY `SKILL-016` – `026` AUDIT:** Passed `SKILL-005/006/008` , but failed `SKILL-016` (paraphrased non-SEC text), `SKILL-017` (small hand-picked N=48, clipped `[-40, +40]`), `SKILL-018` (synthetic judge jitter), and `SKILL-021` (unaligned reaction timestamps).
`v3.0` – `v3.2` (Archived) `2026-11-01` – `2026-11-02`	**N = 180 outcome-blind verbatim SEC 8-K Item 2.02 (`EX-99.1`) releases across G = 45 firms** (`90` in 2023; `90` in 2024 Q3 + 2025 Q1)	`H, V, U, W` scored by **3 real OpenRouter LLMs (`Gemini-3-Flash` , `DeepSeek-V3` , `Llama-3.3-70B`) + `structural-linguistic-v1`**	Confirmatory test of H1/ H2 (30d post-reaction drift $CAR_{[+1,+21]}$) + Secondary test of Reaction-Day $CAR_{t_\{react\}}$ & OOS vs. `0.0%` baseline	**Archived (`PAPER_MANUSCRIPT_V3_2_ARCHIVED.md` , `f50de310...`):** Pre-registered H1 (`+6.66%, t = +1.20`) and H2 (`-10.34%, t = -1.46`) are NULL, and `v3.0` – `v3.2` over-claimed post-hoc claims `C1` , `C2` , and `C3` without `XLK` / `QQQ` benchmarking or leave-one-season-out (`R1–R2`).
`v4.0` (Current) `2026-11-02`	**N = 180 outcome-blind verbatim SEC 8-K Item 2.02 (`EX-99.1`) releases**	`H, V, U, W` (`ICC(2,1) = 0.844 / 0.823 / 0.607`), evaluated against `SPY` , `XLK` ,	Confirmatory H1/H2 (NULL) + De-confounded multi-benchmark & leave-one-	**Methodological Reckoning & Full Null / Exploratory Reclassification (`DIRECTOR-FEEDBACK-S02-2026-11-02` & `SG-AUDIT-V4.0-EXT-AUDIT-020`):** Retires `C1` (`-0.07%, t = -0.05` vs `XLK`) and `C2` (`-1.83%, t = -1.08` vs `XLK`); reclassifies `C3` as between-firm association concentrated in `2025_Q1`

Edition & Date	Sample Size & Corpus	Primary Construct & Judge Setup	Primary Hypothesis Tested	Outcome & Methodological Audit Status
	across G = 45 firms (`45` per season: `2023_Q1`, `2023_Q3`, `2024_Q3`, `2025_Q1`)	`QQQ`, and pre-event $\hat{\beta}_i$ \cdot SPY	season-out tests of `C1`–`C3` & `Early_W` + within-season `2025_Q1` slope on `W`	(`-5.08%, t = -1.36` with 45-firm FE; `-2.57%, t = -0.95` without `2025_Q1`); reports `2025_Q1` slope on `W` (`+8.05%, t = +1.10` bivariate; `+15.44%, t = +2.38` with controls); documents full git & archive provenance (`GIT_COMMIT_HISTORY_AUDIT.txt`).