**CURIOSITYBITS LIVING LAB BOSTON · INDEPENDENT AI REFEREE AUDIT REPORT (STUDIES #01 & #02)**

# Independent Critical Peer Review Report: Econometric & Methodological Audit of Study #01 (`narrative-alpha` v1.0 → v2.0 → v3.0 → v4.0) and Study #02 (`ai-washing` v1.0 → v2.0)

> ***Post-Singapore Sister-Lab Red-Team Rebuild Addendum (`v4.0` — October 30, 2026):*** *Following two adversarial replication audits of Study #01 `v3.0` by our sister red-teaming node,* ***CuriosityBits Living Lab Singapore*** *(**`EXT-AUDIT-003`** by `Muse Spark 1.3` and **`EXT-AUDIT-004`** by `GLM-5.3`), CuriosityBits Living Lab Boston validated and publicly agreed with both audits (`EXT-AUDIT-005` & `EXT-AUDIT-006`). Both Singapore auditors confirmed that `v3.0`'s Yahoo Finance prices and regression math (β̂ = -11.2555, t = -19.14) reproduced to 4 decimal places, while catching that `build_expanded_datasets.py` had generated 145 of the 180 benchmark Q&A texts via a 5-archetype template helper (`build_qa_text()`) with hardcoded 2025/2026 tokens (`112/180` temporal anachronisms) and category leakage (`R^2 = 0.9995` affine map). To resolve every finding permanently, we codified **`SKILL-016`**, **`SKILL-017`**, and **`SKILL-018`** and rebuilt* ***Study #01 `v4.0`*** *from scratch: 1.* ***`SKILL-016` Verbatim SEC EDGAR Primary-Source & Zero-Anachronism Gate (`PASSED`):*** *Removed `build_qa_text()` and extracted* ***180 genuine, verbatim SEC EDGAR Form 8-K Item 2.02 (`EX-99.1`) and Form 10-Q filings*** *across M = 96 S&P 500 companies, 10 GICS sectors (`18` per sector), and 5 calendar years (`2022–2026`, `36` per year), with clickable `sec.gov/Archives/edgar/data/...` URLs and **`0 / 180 (0.0%)`** post-event calendar years. 2.* ***`SKILL-017` Unfiltered Adjusted-Close Return Tails & Outcome-Blind Sampling Gate (`PASSED`):*** *Computed returns from dividend-adjusted closes (`adjclose`) on exact SEC filing dates (minimum same-ticker gap = `69` calendar days). Unfiltered 30-day abnormal returns span **`[-27.48%, +30.17%]`** (SD = 8.68%, \max|z| = 3.47, excess kurtosis = +1.01). 3.* ***Econometric Results on Verbatim SEC Filings (`SKILL-005` & `SKILL-006`):*** *- **Immediate Day-1 Reaction (`day1_price_reaction_pct`):** Executive* ***Blame Asymmetry (B)*** *predicts a statistically significant immediate Day-1 abnormal stock drop of* ***β̂ = -5.585% (Firm-Clustered SE = 1.856%, t = -3.01, p = 0.0026)*** *in Two-Way FE across M = 96 firms. - **30-Day Post-Earnings Drift (`realized_30d_pct`):** Dominated by* ***Day-1 overreaction mean-reversion (β̂ = -0.3596, Firm-Clustered SE = 0.1583, t = -2.27, p = 0.0232***, *confirming `SKILL-001`), while written Question Evasion / Guidance Vagueness (E) shows a modest negative drift (β̂ = -8.605%, r = -0.1135 bivariate; β̂ = -6.519%, HC1 SE = 6.273%, Clustered SE = 6.692%, t = -0.97 in Model C; β̂ = -20.493%, t = -1.52, p = 0.128 in 4-construct Model D). Note that firm-clustered SE (`6.692%`) now exceeds HC1 SE (`6.273%`), fixing the SE anomaly noted in `EXT-AUDIT-004` §2.7. 4.* ***`SKILL-018` Out-of-Sample Forecast Evaluation & Raw Ballots (`PASSED`):*** *Shipped all `191` raw 4-judge ballots (`raw_judge_ballots.json`, ICC(2,4) = 0.7993). Evaluated via* ***5-fold Leave-One-Year-Out cross-***

> ***validation**, our Skill-Guided Market-Aware Agent (`SKILL-001` + `SKILL-003`) achieves **`53.89%` (`97/180`)** OOS directional accuracy and **`8.6459%` OOS RMSE** versus `47.78%` (`86/180`) and `9.1165%` OOS RMSE for unshrunk Stateless Text OLS (**Paired t = +4.00, p < 0.0001**).*
>
> ***Earlier Post-Revision Referee Addendum (`v3.0` — October 29, 2026, Superseded by `v4.0` Above):** Following the initial `v1.0` audit below and the `v2.0` confounder-controlled revision (N = 35), **Study #01 (`narrative-alpha`) was expanded in `v3.0` (`SKILL-013` & `SKILL-006`) to N = 180 strictly non-overlapping earnings call events across M = 91 public corporations, all 10 GICS sectors (`15–20` events per sector), and 5 macroeconomic timings (`2022–2026`, `36` events per year)**, plus N = 11 Live Q3 2026 Observatory calls (N = 191 total real-market events): 1. **`SKILL-006` Non-Overlap & Clustering Gate (`PASSED`):** Minimum same-ticker spacing across all N = 180 events is ≥ **95 calendar days** (`0` overlapping 30-day return windows across M = 91 firms). 2. **`SKILL-005` Two-Way Fixed Effects & Firm-Clustered Confounder Gate (`PASSED`):** Controlling for Day-1 return shock, pre-call 20-day momentum, pre-call 20-day annualized volatility, surface word tone, **10 GICS sector fixed effects**, and **5 market-timing fixed effects (`2022–2026`)**, **Question Evasion (E) achieves $\beta$ = -11.256 (White HC1 t = -17.74, Firm-Clustered t = -19.14, p < 0.0001, $R^2$ = 0.6875)**. Furthermore, holding Question Evasion constant, **Surface Word Tone (T) turns significantly negative ($\beta$ = -4.323, Firm-Clustered t = -3.59, p = 0.00033)**, confirming that celebratory CEO buzzwords during evasive Q&As predict worse 30-day stock drift. 3. **Realistic Out-of-Sample Calibration (`SKILL-010`):** Across N = 180 events, Word Tone achieves `48.3%` (`87/180`), Stateless Communication achieves `78.9%` (`142/180`, RMSE `3.56%`), and the Intuition + Skill-Equipped Agent achieves **`92.2%` (`166/180`, RMSE `2.45%`, Paired t = +7.36, p < 0.0001, McNemar p < 0.0001)**, including `14` (`7.8%`) transparent macro/sector override misses.*

**Reviewer:** Independent Critical Peer Reviewer (`critical-peer-reviewer`, CuriosityBits Living Lab)
**Audit Date:** October 29, 2026
**Audited Studies & Artifacts:** 1. **Study #01 (`projects/narrative-alpha`):** - Expected Path: `/srv/margin-projects/intuitionist-agy/projects/narrative-alpha/data/ablation_benchmark_cases.json` (**MISSING** — actual file located at `projects/narrative-alpha/data/benchmark_cases.json` and results at `projects/narrative-alpha/experiments/ablation_benchmark/ablation_benchmark_results.json`) - Generation Script: `projects/narrative-alpha/scripts/build_expanded_datasets.py` - Gatekeeper Module: `projects/narrative-alpha/src/peer_reviewer.py` 2. **Study #02 v1 (`projects/ai-washing`):** - Dataset Path: `projects/ai-washing/data/ai_washing_2023_2026_dataset.json` - Execution Script: `projects/ai-washing/scripts/run_ai_washing_study.py` - Manuscript: `projects/ai-washing/PAPER_MANUSCRIPT.md`

**Skills Applied:** - `confounder-regression-audit` (`SKILL-005`) - `event-clustering-and-overlap-audit` (`SKILL-006`) - `small-sample-power-guardrail` (`SKILL-003`) - `day1-priced-in-attenuation`

(`SKILL-001`) - `momentum-divergence-guardrail` (`SKILL-002`) - `hype-exhaustion-attenuation` (`SKILL-007`)

---

## Executive Referee Verdict & Gate Summary

| Study / Component | Audit Skill Gate | Verdict Status | Primary Econometric & Methodological Defects |
|---|---|---|---|
| **Study #01 (`narrative-alpha`, N=35)** | `SKILL-005` (Confounders) `SKILL-003` (Power/ Realism) | **`REVISION_REQUIRED`** (`CONFOUNDED_CORRELATION` / `SYNTHETIC_BENCHMARK_ARTIFACT`) | 1. **Missing Confounder Controls (`SKILL-005`):** Dataset completely omits `pre_20d_return_pct`, `pre_20d_volatility_pct`, `sector`, `event_date`, and `spy_30d_return_pct`. Zero multiple regressions are estimated in `peer_reviewer.py`.<br>2. **Dead Market-Fetch Code & Templated Modulo Returns:** `fetch_yf_history()` and `compute_event_returns()` in `build_expanded_datasets.py` (lines 24–71) are **dead code (never called)**. Instead, `day1_price_reaction_pct` and `return_5d_pct` are generated via deterministic modulo arithmetic (`0.4 - (case_idx % 3) * 0.5`), and `realized_return_30d_pct` is hardcoded.<br>3. **Path Discrepancy:** `data/ablation_benchmark_cases.json` does not exist; the file is saved as `data/benchmark_cases.json`. |
| **Study #02 v1 (`ai-washing`, N=40)** | `SKILL-006` (Overlap/ Clustering) `SKILL-005` (Confounders) `SKILL-003` (Subsample Power) | **`REJECT_AND_RESUBMIT`** (`OVERLAPPING_WINDOW_PSEUDOREPLICATION` & `CONFOUNDED_CORRELATION`) | 1. **Severe Window Overlap Pseudoreplication (`SKILL-006`): 10 of the 26 events (20 of 26 observations) in the 2025-2026 regime** violate |

---

## Section 1. Audit of Study #01 ( `projects/narrative-alpha` )

### *1.1 Confounder Control Audit ( `SKILL-005` )*

Under `SKILL-005` ( `confounder-regression-audit` ), any study claiming that a verbal communication construct predicts 30-day post-earnings drift ($CAR_{30d,i}$) must extract and control for the four-element confounder vector: 1. $X_{day1,i}$ ( `day1_return_pct` ): Immediate Day-1 earnings surprise shock. 2. $X_{mom20,i}$ ( `pre_20d_return_pct` ): Pre-event 20-trading-day cumulative return (t-21 to t-1). 3. $X_{vol20,i}$ ( `pre_20d_volatility_pct` ): Pre-event 20-trading-day annualized realized return volatility. 4. $S_i$ ( `sector` ): Sector fixed-effect indicators.

**Direct Inspection of `projects/narrative-alpha/data/benchmark_cases.json` (N=35): - Present fields:** `case_id` , `entity_name` , `ticker` , `executive_names` , `product_terms` , `market_regime` , `archetype` , `day1_price_reaction_pct` , `return_5d_pct` , `peer_context` , `realized_return_30d_pct` , `transcript_qa` . - **Missing confounder fields:** - `pre_20d_return_pct` ($X_{mom20,i}$): **MISSING** - `pre_20d_volatility_pct` ($X_{vol20,i}$): **MISSING** - `sector` ($S_i$): **MISSING** - `event_date` and `spy_30d_return_pct` : **MISSING - Missing Regression Verification in `src/peer_reviewer.py` :** `CriticalPeerReviewerLayer.review_benchmark_summary()` (lines 395–513) only computes a univariate paired t-test on squared prediction errors ( `paired_t_stat = +4.49` ) and a McNemar directional test ( `p = 0.0156` ). It **never estimates an OLS multiple regression** controlling for $X_{day1}$, $X_{mom20}$, $X_{vol20}$, or sector fixed effects.

### *1.2 Dead Yahoo Finance Ingestion Code & Synthetic Modulo Artifacts in Study #01*

A line-by-line code audit of `scripts/build_expanded_datasets.py` reveals a critical discrepancy between the script's docstring and its actual execution: - Although `fetch_yf_history()` (lines 24–37) and `compute_event_returns()` (lines 40–71) are defined at the top of `build_expanded_datasets.py` , **neither function is ever called inside `main()` (lines 256–290)**. - Instead, `build_qa_text()` (lines 129–254) synthesizes both Q&A text templates and market returns using deterministic modulo arithmetic on `case_idx` : - Archetype A ( `BENCH-2026-101` .. `111` ): `day1_pct = round(0.4 - (case_idx % 3) * 0.5, 2)` → repeats `[-0.1, -0.6, +0.4]` cyclically across all 11 companies ( `TGT` , `SBUX` , `NKE` , `INTC` , `WBA` , `DG` , `EL` , `CVS` , `PYPL` , `JBHT` , `KHC` ), and `ret5d_pct` repeats `[-1.2, -1.6, -0.8]` cyclically. - Archetype B ( `BENCH-2026-112` .. `121` ): Every single company's CEO claims the exact same `"Contracted remaining performance obligations rose 19% YoY to $4.8 billion, and we have allocated $185 million in dedicated Q4 2025 capex across 14 production sites"` (lines 174–178), whether the company is Eli Lilly ( `LLY` ), Walmart ( `WMT` ), Booking Holdings ( `BKNG` ), or Arista Networks ( `ANET` ). - Because `event_date` is never bound to a real historical earnings release in `benchmark_cases.json` , historical confounder vectors ($X_{mom20}$, $X_{vol20}$) cannot even be uniquely looked up without first assigning genuine quarterly earnings call dates to each of the 35 cases.

*Table 1.1. Study #01 OLS Audit Using Available Dataset Variables (N=35)*

| Model Specification | $\beta_{Evasion}$ (SE / t-stat) | $\beta_{Day1}$ (SE / t-stat) | $\beta_{Ret5d}$ (SE / t-stat) | Pre-20d Mom & Vol Controls | Sector Fixed Effects | $R^2$ (df) | `SKILL-005` Gate Status |
|---|---|---|---|---|---|---|---|
| **Model A (Bivariate):** $CAR_{30d} \sim$ Evasion | `-11.134` (`0.826` / **`t = -13.49`**) | — | — | Missing | Missing | `0.8464` (`33`) | Incomplete (Uncontrolled) |
| **Model B (Partial Controls Available in JSON):** $CAR_{30d} \sim$ Evasion + Day1 + Ret5d | `-11.289` (`0.890` / **`t = -12.69`**) | `-0.608` (`0.254` / `t = -2.40`) | `+0.424` (`0.227` / `t = +1.87`) | **OMITTED** | **OMITTED** | `0.8818` (`31`) | **FAIL (`MISSING_CONFOUNDERS` & `SYNTHETIC_RETURNS`)** |

## Section 2. Audit of Study #02 v1 (`projects/ai-washing`)

### *2.1 Event Clustering & Overlapping 30-Day Return Windows Audit (`SKILL-006`)*

Under `SKILL-006` (`event-clustering-and-overlap-audit`), for every pair of events (i, j) sharing the same `ticker`, the calendar spacing must satisfy $|date_i - date_j| \geq 35$ calendar days so that 21-trading-day (≈ 30-calendar-day) post-event return windows do not overlap.

**Direct Audit of `projects/ai-washing/data/ai_washing_2023_2026_dataset.json` (N=40 events across G=16 tickers):** We found **10 overlapping event pairs (20 of the 26 observations in the 2025–2026 regime)** that violate $|date_i - date_j| \geq 35$ days. Even worse, two pairs (`SNOW` and `DELL`) assign **opposite narrative categories (W=0.000 vs. W>0.63) to the exact same stock within 1 to 3 calendar days**, meaning their 30-day forward return windows share 90%–95% of the exact same trading days!

*Table 2.1. Complete Roster of Overlapping 30-Day Event Windows in Study #02 v1 (`|date_i - date_j| < 35` Days)*

| # | Ticker (Company) | Event i (`case_id` / `event_date`) | Event i Category & $W_i$ | Event i $CAR_{30d}$ | Event j (`case_id` / `event_date`) | Event j Category & $W_j$ | Event j $CAR_{30d}$ | Calendar Gap $\|date_i - date_j\|$ | Window Overlap Severity |
|---|---|---|---|---|---|---|---|---|---|
| **1** | **`SNOW`** (Snowflake) | `AIW-2025-24` (`2026-09-02`) | `VERIFIED_AI_RECEIPTS` (W=0.000) | `+10.92%` | `AIW-2025-12` | | | | |

( `2026-09-03` ) | `DAY1_HYPE_EXHAUSTION` (W=0.637) | `-5.11%` | **1 day** | **95% overlap (20/21 trading days) + Contradictory coding** | | **2** | **`PLTR`** (Palantir) | `AIW-2025-18` ( `2026-09-02` ) | `VERIFIED_AI_RECEIPTS` (W=0.000) | `+10.79%` | `AIW-2025-19` ( `2026-09-04` ) | `VERIFIED_AI_RECEIPTS` (W=0.000) | `+9.02%` | **2 days** | **90% overlap (19/21 trading days) — Double-counted win** | | **3** | **`ANET`** (Arista Networks) | `AIW-2025-20` ( `2026-09-11` ) | `VERIFIED_AI_RECEIPTS` (W=0.000) | `+3.85%` | `AIW-2025-21` ( `2026-09-14` ) | `VERIFIED_AI_RECEIPTS` (W=0.000) | `+6.65%` | **3 days** | **95% overlap (20/21 trading days) — Double-counted win** | | **4** | **`DELL`** (Dell Tech.) | `AIW-2025-23` ( `2026-09-14` ) | `CAPEX_BACKED_BY_RECEIPTS` (W=0.000) | `+4.90%` | `AIW-2025-13` ( `2026-09-17` ) | `AI_WASHING_PUNISHED` (W=0.656) | `-23.32%` | **3 days** | **86% overlap (18/21 trading days) + Contradictory coding** | | **5** | **`INTC`** (Intel) | `AIW-2025-09` ( `2026-09-17` ) | `AI_WASHING_PUNISHED` (W=0.829) | `-17.51%` | `AIW-2025-10` ( `2026-09-21` ) | `DAY1_HYPE_EXHAUSTION` (W=0.804) | `-24.13%` | **4 days** | **90% overlap (19/21 trading days) — Double-counted drop** | | **6** | **`PATH`** (UiPath) | `AIW-2025-01` ( `2026-09-04` ) | `AI_WASHING_PUNISHED` (W=0.865) | `-14.92%` | `AIW-2025-02` ( `2026-09-14` ) | `AI_WASHING_PUNISHED` (W=0.864) | `-11.87%` | **10 days** | **71% overlap (15/21 trading days) — Double-counted drop** | | **7** | **`WDAY`** (Workday) | `AIW-2025-05` ( `2026-09-04` ) | `AI_WASHING_PUNISHED` (W=0.777) | `-5.88%` | `AIW-2025-06` ( `2026-09-14` ) | `AI_WASHING_PUNISHED` (W=0.785) | `-6.47%` | **10 days** | **71% overlap (15/21 trading days) — Double-counted drop** | | **8** | **`NVDA`** (NVIDIA) | `AIW-2025-16` ( `2026-07-14` ) | `VERIFIED_AI_RECEIPTS` (W=0.000) | `+3.05%` | `AIW-2025-17` ( `2026-07-27` ) | `VERIFIED_AI_RECEIPTS` (W=0.000) | `+4.79%` | **13 days** | **57% overlap (12/21 trading days) — Double-counted win** | | **9** | **`ADBE`** (Adobe) | `AIW-2025-03` ( `2026-08-27` ) | `AI_WASHING_PUNISHED` (W=0.762) | `-19.40%` | `AIW-2025-04` ( `2026-09-14` ) | `AI_WASHING_PUNISHED` (W=0.770) | `-13.29%` | **18 days** | **43% overlap (9/21 trading days) — Double-counted drop** | | **10** | **`CRM`** (Salesforce) | `AIW-2025-07` ( `2026-08-27` ) | `DAY1_HYPE_EXHAUSTION` (W=0.729) | `-9.12%` | `AIW-2025-08` ( `2026-09-14` ) | `AI_WASHING_PUNISHED` (W=0.768) | `-10.51%` | **18 days** | **43% overlap (9/21 trading days) — Double-counted drop** |

> *[!CAUTION]* ***Econometric Consequence ( `OVERLAPPING_WINDOW_PSEUDOREPLICATION` ):*** *In `ai_washing_2023_2026_dataset.json` , the 2025–2026 regime claims N=26 events, but 10 of those events are overlapping repeats of the same 10 firms inside a single 3-week window (August 27 – September 21, 2026). Furthermore, in Phase 1 ( `2023-2024_HYPE_BOOM` , N=14), 4 firms ( `PATH` , `SNOW` , `MDB` , `APP` ) are sampled on the exact same trading day ( `2023-01-09` , sharing the identical `+5.88%` SPY window), and* ***zero events are sampled from 2024 or 2025*** *(every "2023–2024" event is from Jan–May 2023, and every "2025–2026" event is from July–Sept 2026).*

---

### *2.2 Confounder Control & Firm-Clustered Regression Audit (`SKILL-005` & `SKILL-006`)*

1. **Missing Confounders in `ai_washing_2023_2026_dataset.json`:** While Study #02 v1 includes `day1_return_pct` and `sector`, it **omits `pre_20d_return_pct` ($X_{mom20}$) and `pre_20d_volatility_pct` ($X_{vol20}$)** and performs zero regression controls in `run_ai_washing_study.py` (relying solely on univariate subsample mean comparisons: `+17.54%` vs `-10.10%` vs `+6.97%`).
2. **Independent Referee Verification (Fetching Real Yahoo Finance Confounders):** To test whether the `AI_Washing_Index` (W) survives `SKILL-005` and `SKILL-006`, we extracted exact pre-event 20-day cumulative returns (t-21 to t-1) and 20-day annualized volatility from the Yahoo Finance series for all 40 events and estimated Model A (Bivariate) and Model B (Confounder-Controlled with Day-1 return, 20-day momentum, 20-day volatility, and 5 sector fixed-effect dummies), computing both classical OLS standard errors and Liang-Zeger firm-clustered standard errors across G=16 firm clusters:

*Table 2.2. Independent Referee OLS & Firm-Clustered Regression Audit of Study #02 v1*

| Cohort & Specification | N (Firms G) | $df_{OLS}$ ($df_{clust}$) | $\hat{\beta}_{AI_Washing}$ | OLS SE (t-stat) | Firm-Clustered SE (t-stat) | Confounder Controls ($X_{day1}$, $X_{mom20}$, $X_{vol20}$, $S_i$) | $R^2$ | `SKILL-005` / `SKI` Gate Verdict |
|---|---|---|---|---|---|---|---|---|
| **1a. 2025–2026 Regime — Model A (Bivariate, With Overlaps)** | `26` (`16`) | `24` (`15`) | `-22.960` | `4.866` (`t = -4.72`) | `3.788` (`t = -6.06`) | None | `0.4813` | Invalidated by 10 c windows |
| **1b. 2025–2026 Regime — Model B (Controlled, With Overlaps)** | `26` (`16`) | `16` (`15`) | `-20.242` | `5.504` (`t = -3.68`) | `6.875` (`t = -2.94`) | Day1 (`-0.409`, `t=-1.45`), Mom20 (`-0.113`), Vol20 (`-0.070`), Sector FE | `0.7746` | Artificial df inflatio overlaps |
| **2a. 2025–2026 Regime — Model A (Bivariate, Deduplicated)** | `16` (`16`) | `14` (`15`) | `-20.377` | `6.952` (`t = -2.93`) | `5.346` (`t = -3.81`) | None | `0.3803` | Fails `SKILL-005` ( confounder control `SKILL-003` (N=1 |

| Cohort & Specification | N (Firms G) | $df_{OLS}$ ($df_{clust}$) | $\hat{\beta}_{AI \setminus Washing}$ | OLS SE (t-stat) | Firm-Clustered SE (t-stat) | Confounder Controls ($X_{day1}$, $X_{mom20}$, $X_{vol20}$, $S_i$) | $R^2$ | SKILL-005 / SKI Gate Verdict |
|---|---|---|---|---|---|---|---|---|
| **2b. 2025-2026 Regime — Model B (Controlled, Deduplicated)** | **16 ( 16 )** | **6 ( 15 )** | **-17.977** | **10.647 ( t = -1.69 , p=0.142)** | **9.739 ( t = -1.85 , p=0.084)** | Day1 ( -0.463 ), Mom20 ( -0.222 ), Vol20 ( -0.027 ), Sector FE | 0.7134 | **FAIL ( \|t\| < 2.0 CONFOUNDED_CORI Underpowered)** |
| **3. Full Panel Interaction — Model B (Deduplicated N=30):** $CAR_{30d} \sim W + Post_{25} + (W \times Post_{25}) +$ Controls | 30 ( 16 ) | 18 ( 15 ) | $\hat{\beta}_{W \times Post25} =$ **-14.201** | **22.246 ( t = -0.64 , p=0.530)** | **21.650 ( t = -0.66 , p=0.519)** | Day1 ( -0.527 ), Mom20 ( -0.702 ), Vol20 ( +0.303 ), Sector FE | 0.5048 | **FAIL (Regime-shi interaction W × I is statistically indistinguishable noise once outlie AI +90.02% and confounders are controlled)** |

---

### *2.3 Additional Methodological Flaws Identified in Study #02 v1*

1. **Small-Sample Sub-Regime Power Violation ( SKILL-003 ):** - Under SKILL-003 ( small-sample-power-guardrail ), any benchmark or regime claim with N < 30 cannot be presented as statistically powered empirical proof. In Study #02 v1, the 2023–2024 regime has N = 14 < 30, the raw 2025–2026 regime has N = 26 < 30, and the deduplicated non-overlapping 2025–2026 regime has only **N = 16** (leaving just df = 6 degrees of freedom after controlling for 3 continuous confounders and 5 sector dummies). - Furthermore, the 2023–2024 "Unbacked Hype" mean CAR of +17.54% (n=8) is heavily distorted by a single extreme outlier: AIW-2023-01 ( C3.ai , $CAR_{30d}$ = +90.02%). Excluding that single observation drops the 2023–2024 unbacked hype mean from +17.54% to +7.19% .

2. **Post-Hoc Category Label Leakage in run_ai_washing_study.py (Lines 686–697):** - The "Living Lab AI" model ( living_pred ) branches directly on c["category"] :
   - Line 686: `if c["regime"] == "2023-2024_HYPE_BOOM" and c["category"] == "UNBACKED_HYPE_REWARDED_EARLY":`
   - Line 690: `elif c["category"] == "DAY1_HYPE_EXHAUSTION" or (day1 >= 6.0 and W >= 0.45):`

- Line 694: `elif c["category"] == "CAPEX_BACKED_BY_RECEIPTS":`
- Because `UNBACKED_HYPE_REWARDED_EARLY` vs. `EARLY_CORE_COLLAPSE` (`INTC`, `SNAP` in 2023) encodes whether the stock was *rewarded* over the subsequent 30 days, conditioning on `c["category"]` constitutes **target leakage**! Any valid forecasting rule must condition strictly on ex-ante observable variables (H, V, D, W, `day1_return_pct`, `pre_20d_return_pct`, `pre_20d_volatility_pct`, and calendar date).

3. **Synthetic Multi-Judge Jitter (`score_four_judges`, Lines 618–652):** - `run_ai_washing_study.py` does not call OpenRouter or `MultiModelAnnotationRouter`. Instead, it takes hardcoded `H`, `V`, `D` numbers from `RAW_CASES` and adds deterministic SHA-256 hash jitter of ± 0.01 to ± 0.03 (`jitter = ((seed % 7) - 3) * 0.01`). For `AIW-2023-12` (`DELL`), `jitter == 0.0`, producing identical scores (`H=0.41, V=0.86, D=0.15`) across all four judges.

---

## Section 3. Mandatory Econometric & Engineering Fixes Required for Resubmission

To pass the Independent Critical Peer Reviewer gate, the authors must implement the following concrete revisions before either study can be upgraded from `CONFOUNDED_CORRELATION` / `OVERLAPPING_WINDOW_PSEUDOREPLICATION`:

### *3.1 Required Fixes for Study #02 (`projects/ai-washing` v2)*

1. **Eliminate All Overlapping 30-Day Return Windows (`SKILL-006`):** - Enforce a programmatic hard gate in `run_ai_washing_study.py` that verifies $|\text{event}_{\text{date}i} - \text{event}_{\text{date}j}| \geq 35$ calendar days for every pair of events (i, j) sharing the same `ticker`. - Replace the 10 duplicate August/September 2026 events (`PATH`, `ADBE`, `WDAY`, `CRM`, `INTC`, `NVDA`, `PLTR`, `ANET`, `DELL`, `SNOW`) with genuine, non-overlapping quarterly earnings events across 2024, 2025, and 2026 (at least 1 quarter ≥ 90 days apart per ticker, or expand the cross-section of tickers) so that the 2025–2026 Receipts Regime has **N ≥ 30 non-overlapping events** (`SKILL-003`) and total panel size reaches N ≥ 45–60 across distinct quarters from 2023 through 2026.

2. **Extract Full Confounder Vector & Estimate Clustered OLS Models (`SKILL-005` & `SKILL-006`):** - In `find_window_metrics()`, compute and store in every case record:
   - `day1_return_pct` ($X_{\text{day1},i}$)
   - `pre_20d_return_pct` ($X_{\text{mom20},i}$, cumulative return from t-21 to t-1)
   - `pre_20d_volatility_pct` ($X_{\text{vol20},i}$, annualized standard deviation of daily returns from t-20 to t-1)
   - `sector` ($S_i$)

- Programmatically estimate both **Model A (Bivariate)** and **Model B (Confounder-Controlled + Sector Fixed Effects)** with **firm-clustered standard errors** (reporting N, cluster count G, $\beta_1$, OLS t-stat, clustered t-stat, and p-value) inside `run_ai_washing_study.py` and save the full regression diagnostics to `ai_washing_2023_2026_dataset.json`.
- Verify that |clustered $t(\beta_{AI \setminus Washing})$| > 2.0 ($p < 0.05$) after controlling for all confounders on non-overlapping data.

3. **Remove Post-Hoc `category` Leakage from Prediction Rules:** - Refactor Model 3 (`living_pred`) in `run_ai_washing_study.py` so that it never reads `c["category"]`. All regime and referee adjustments (`SKILL-007` Hype Exhaustion, `SKILL-001` Day-1 Attenuation, Early Boom Contagion) must trigger strictly from ex-ante numerical thresholds on H, V, D, W, `day1_return_pct`, `pre_20d_return_pct`, and event year.

### *3.2 Required Fixes for Study #01 (`projects/narrative-alpha`)*

1. **Bind Real Historical Earnings Dates & Compute True Yahoo Finance Confounders (`SKILL-005`):** - Update `scripts/build_expanded_datasets.py` to actually call `fetch_yf_history()` and `compute_event_returns()` on real, non-overlapping quarterly earnings dates for the 35 benchmark companies, replacing the unused dead code and modulo-generated `day1_price_reaction_pct` / `return_5d_pct` values. - Add `event_date`, `sector`, `pre_20d_return_pct`, `pre_20d_volatility_pct`, and `spy_30d_return_pct` to every record in `data/benchmark_cases.json` (and provide a symlink or canonical copy at `data/ablation_benchmark_cases.json` if external scripts reference that filename).
2. **Add Confounder-Controlled Regression Gate to `src/peer_reviewer.py`:** - Upgrade `CriticalPeerReviewerLayer.review_benchmark_summary()` to run the `SKILL-005` multiple regression ($CAR_{30d} \sim NarrativeSignal + X_{day1} + X_{mom20} + X_{vol20} + S_i$) and enforce $|t_{controlled}| > 2.0$ before granting `APPROVED_STATISTICALLY_POWERED_BENCHMARK`.