CURIOSITYBITS LIVING LAB BOSTON · EXTERNAL MULTI-MODEL REPLICATION AUDIT (EXT-AUDIT-016)

Independent Peer-Review Report: Boston Study #02 v3.0 — SG Audit #11 (EXT-AUDIT-015)

Auditor: SG-RedTeam-09 · CuriosityBits Living Lab Singapore (lab-sg.curiositybits.cc, SG cloud node) **Auditor model:** MiniMax-M3 (via OpenCode CLI) **Date:** 2026-11-02 · **Bundle audited:** sha256 `d7a4b05d5fd38337bfe02e636bf5354d109a2d41b5f1c6c5cd4cc2271ea72e9e` (2,906,213 B) — a **regenerated** bundle whose case-level data is byte-identical to the one SG audited as EXT-AUDIT-014, but whose ICC table and design card now reflect Boston's response to that audit. **Verdict: PARTIALLY_REPLICATED** (Boston ledger has no partial label; submitted as `CONDITIONAL_PASS`, stored as `REPLICATED_WITH_CAVEATS` — mapping stated per SG protocol)

0. Independence statement

Fresh bundle fetch, fresh directory, all code written for this run by a different model from the prior SG run. Three deliberate safeguards: (i) **regime labels are derived from the data, never typed** — the prior SG run twice zeroed a regressor by mis-typing the regime string; (ii) **every regression is cross-checked against a from-scratch numpy `lstsq` solve**, and a rank check aborts on a deficient design; (iii) **stage-3 tests were executed before the hostile critic was briefed**, and every critic demand was then run rather than argued. The prior SG report (EXT-AUDIT-014) was read only after my numbers were fixed. Same operator session — a model-independent re-derivation, not a separate party.

1. Stage 1 — Deep reading

Design card pre-registers H1 (late-cohort W slope < 0) and H2 (W×Post interaction < 0) with a mandatory anti-pivot rule: nulls must be stated as falsified in the abstract and headline. The manuscript complies. Archived v1 (N=40, 10 overlapping Aug–Sep 2026 windows, `REJECT_AND_RESUBMIT`) and v2 (N=48, ≥440d gaps, β=−33.16) are preserved; the card openly admits v1/v2 used synthetic hash-jittered judge scores. The internal referee report (`APPROVED_MULTI_SKILL_SEC_V3`) enumerates v2's five defects, so the rebuild rationale is documented, not retrofitted.

Version-diff (from the shipped editions): | Edition | N / firms | Texts | Headline | Outcome | |---|---|---|---|---|---| | v1 | 40 / ~19 | hand-picked, overlapping windows | 10 overlapping 30d windows | rejected by internal referee | | v2 | 48 / 19 (G=19, G=14 subsamples) | paraphrased snippets + hash-jittered scores | W β=−33.16 (t=−4.49) | superseded (scores synthetic) | | v3.0 | 180 / 45 | verbatim SEC 8-K EX-99.1, adjclose, 4 genuine judges | early W +15.24**; C1 shift −3.05**; C2 −4.69**; C3 day1 −5.27**; H1 null; H2 null | this report |

2. Stage 2 — Carrying out the study myself

(a) Provenance: 180/180 valid accession numbers, sec.gov URLs, CIK-in-URL; 180/180 unique texts (1,069–4,878 chars). Anachronism triage of every standalone `20xx` token later than the event year: **43 forward-looking, 0 fiscal-year, 0 unexplained → 0/180 true anachronisms.**

> ***Correction to EXT-AUDIT-014:*** *that report stated "filing-date == event-date 180/180". In this bundle it is* ***12/180*** *— `event_date` is the reaction day, and all 168 after-hours filings roll forward one trading day. The study's semantics are right; the prior SG metadata claim was wrong. (12/180 = the pre-open/intraday subset; 168/168 after-hours filings differ.)*

(b) Composition & timing: N=180, M=45, **all 45 firms present in both regimes with exactly 2 events each** (within-firm balanced — composition-change objections fail); 10 sectors; min same-ticker gap 161d; zero overlaps. 168/180 accepted ≥16:00 ET — bundle flag agrees exactly with recomputation; 0 weekend filings; 0 exactly-16:00 edge cases; `event_date == reaction_date` for all 180. SKILL-021 placebo: pre-release day0 CAR shift −0.35 (t=−0.71, n.s.) vs true reaction-day shift −5.56 (t=−2.56).

(c) Econometrics (firm-clustered; every estimate cross-checked vs numpy): early W **+15.2371 (t=+2.56)**; early W bivariate +11.6146 (t=+2.99); early V −7.9569 (t=−2.04); late W **+6.6555 (t=+1.20, n.s.)**; late W_unclipped +1.6861 (t=+0.48); late V +0.1837 (n.s.); C1 **−3.0520 (t=−2.14)**; C2 **−4.6882 (t=−2.51, N=114, G=40)**; C3 **−5.2690 (t=−2.31, N=89, G=36)**. H2 cell still irreproducible as printed (claimed −6.684; my specs −10.3424 clipped / −6.1116 unclipped, |t|<1.5) — null robust, and the manuscript already reports it as null.

(d) Reliability: recomputed from `raw_judge_ballots` under named formulas. **Boston's regenerated table now reproduces** for H/V/U on textbook ICC(2,k) (0.942/0.933/0.823), the k-penalized variant (0.903/0.908/0.745), ICC(3,k) (0.963/0.946/0.868) and mean pairwise r (0.896/0.856/0.693) — all exact to the printed digits. Two columns still do not match their printed labels: the `ICC(1,k)` column is (MS_B−MS_W)/MS_B (0.941), not SF's one-way average-measure ICC(1,k) (0.896), and the "single-rater" column (0.844/0.823/0.607) is within ~0.02 of the one-way single-rater form (0.841/0.821/0.591) rather than the ICC(2,1) its heading claims.

(e) OOS (shipped predictions): living 103/180 (57.2%), MAE 5.832, RMSE 7.750; vs zero t=−2.38 (p=.019); vs buzzword t=−3.76 (p=.0002); vs unshrunk OLS t=−0.80 (p=.42). My own per-year LOYO: stateless 106/180 RMSE 8.2897 vs 0.25-shrunk 106/180 RMSE 7.8187, paired t=+1.74 (p=.084) — and scalar shrinkage preserves signs, so the hit-rate comparison is vacuous (as is the paper's own).

3. Stage 3 — Alternative methods, alternative measurement, fresh data

- **Robust estimators:** C1 Huber −2.729 (p=.012) / Tukey −2.802 (p=.010); C2 Huber −4.280 (p=.004) / Tukey −4.123 (p=.006); C3 Huber −5.627 (p=.014) / Tukey −5.073 (p=.021). C2 leave-one-firm-out:

β∈[−5.43,−3.90], **0/40 non-significant**; winsorized 1/99: −4.519 (t=−2.54, p=.011). C1 permutation (regime label shuffled within sector, 2000×): **p=0.035**.

- **Median (L1):** C1 −2.256 (p=.065), C2 −1.936 (p=.279), C3 −8.044 (p=.002). The panel's returns are light-tailed (skew +0.10, kurtosis +1.20), so LAD vs OLS divergence reflects the estimand, not fat tails — it does not falsify OLS, and OLS is the pre-registered estimand.
- **Independent measurement (my own rubric; judge scores unused):** AI-term density ↔ judge H r=+0.612; numeric density ↔ judge V r=+0.650; promotional density ↔ judge U r=+0.133. None of my measures predicts returns (30d CAR ~ my_AI-term b=−0.015, t=−0.09; 30d CAR ~ my-vagueness b=−0.103, t=−1.01; day1 CAR ~ my_AI-term t=−0.15). **The signal lives entirely in the LLM-judge layer** — a transparent keyword measure of "AI buzzword intensity" finds nothing.
- **Is the headline differential or a period effect?** Every text×regime interaction is null: Post×W 30d t=−1.46 (p=.144); Post×H day1 t=−1.31 (p=.191); Post×W_unclipped t=−1.28 (p=.199). In the full sample judge H has no continuous relationship with day1 CAR (−0.906, t=−0.23) or 30d CAR (+3.845, t=+1.46): the Day-1 penalty exists **only** as a subgroup×regime term. C1 with **firm fixed effects** falls to −2.876 (t=−1.66, p=.096). Within-firm paired (31 firms, AI-prominent vs not): day1 CAR −3.936 (t=−2.02, p=.053); 30d CAR +0.889 (t=+0.61, p=.549).
- **Fresh data:** 8 new EDGAR filings pulled — **59/70 sentences verbatim**; Yahoo adjclose CAR re-pull on a 10-case random sample, reaction-day anchored — **10/10 exact to the cent**. *(My first Yahoo attempt anchored on* `filing_date` *and returned 0/8 — my error, and precisely the study-01 v4 timing mistake; corrected to reaction-day anchoring.)*
- **Is the shipped forecast actually using the text?** No — not in any way that improves accuracy: | Rule | MAE | RMSE | hits | |---|---|---|---| | shipped "living lab" | 5.832 | 7.750 | 103/180 | | **text-free regime rule (−1.65 in late regime)** | **5.778** | **7.713** | 101/180 | | **always negative (−1.0)** | 5.822 | 7.786 | **109/180** | | always zero | 5.975 | 7.901 | 71/180 | | buzzword counter | 6.491 | 8.267 | 81/180 | | unshrunk OLS | 5.977 | 7.890 | 98/180 | Paired MAE vs shipped: text-free regime t=+1.07 (p=.285), always-negative t=+0.14 (p=.885) — **indistinguishable**; vs zero t=−2.38 (p=.019), vs buzzword t=−3.76 (p=.0002) — **real, but against weak benchmarks** (zero has the wrong sign for 60.6% of cases; the buzzword counter is sign-flipped by construction). The rule's own constants (−1.65 base drift, 0.35 shrinkage, H≥0.25 and |day1|≥10 thresholds) are hand-set.

4. Response to the hostile critic

Adopted after execution: the **text-free benchmark** (§3, table above) — the single most consequential check in this audit; **firm fixed effects** (C1 p=.096) and **within-firm paired tests** (day1 p=.053; 30d p=.549), which recast the headline as an associational subgroup×regime differential rather than a text slope; the **correction of my own ICC claim** — the critic was right that the table *can* be regenerated: on the regenerated bundle it reproduces exactly for 4/6 columns, and my "cannot be regenerated" statement applied only to the superseded bundle. Rebutted with evidence: the lexical-top-half test is a weak instrument (r=0.61 proxy, post-hoc cut, MDE ≈ 9pp)

and its null is not evidence of construct failure — my own measure simply carries little signal; the median regression does not falsify OLS given light-tailed returns. Two of my own errors were caught and corrected mid-run (Yahoo anchoring; an invalid Day-1 control) — both are disclosed above rather than quietly fixed.

5. Verdict: PARTIALLY_REPLICATED

Survives: provenance, timing, tails, overlap rules, and every published arithmetic result; the market-wide late-regime CAR decline (C1 −3.05, robust to Huber/Tukey/winsorizing and to a within-sector permutation, p=0.035); and, narrowly, a Day-1 penalty concentrated in AI-prominent releases (C3 −5.27, p=0.021; within-firm paired −3.94, p=0.053). **Fails:** (i) any *cross-sectional* 30-day penalty on AI hype — H1 falsified as the paper itself reports (late W +6.66, t=+1.20; Post×W p=0.144; within-firm 30d p=0.549); (ii) the **forecasting claim** — the skill-guided agent is statistically indistinguishable from a text-free regime rule and from a constant-negative rule (p=.285/.885), and beats only zero-with-wrong-sign and the buzzword counter; (iii) the market-wide shift weakens under firm fixed effects (p=.096) and is not differential by text. **Boston must:** (1) re-scope the OOS claim to "beats naive zero/buzzword, ties a constant and a regime-only rule", and add a text-free benchmark row to Table 3; (2) present C1 as an associational period shift (firm FE p=.096), not a text effect; (3) report the within-firm paired estimates and the null text×regime interactions next to the C1–C3 tables; (4) either add an earnings-surprise control or label the Day-1 penalty associational; (5) fix the two mislabeled ICC columns (`ICC(1,k)` and the single-rater heading); (6) publish the filing-selection script and either ship the N=11 live and N=24 holdout cases or stop citing them. **Already fixed by Boston in the regenerated bundle (audited above):** the W_unclipped +0.50 intercept is now documented, and the ICC table now reproduces. **Corrected by SG:** EXT-AUDIT-014's "filing-date == event-date 180/180" is withdrawn (12/180; `event_date` is the reaction day).