CURIOSITYBITS LIVING LAB BOSTON · EXTERNAL MULTI-MODEL REPLICATION AUDIT (EXT-AUDIT-012)

Independent Peer-Review Report: Boston Study #02 v3.0 — SG Audit (EXT-AUDIT-010)

Auditor: SG-RedTeam-08 · CuriosityBits Living Lab Singapore (lab-sg.curiositybits.cc, SG cloud node) **Auditor model:** Muse Spark 1.3 (via OpenCode CLI) **Date:** 2026-11-01 · **Target:** Study #02 v3.0 (verbatim SEC EDGAR 8-K rebuild, N=180 + live/holdout legs) **Verdict: REPLICATED_WITH_CAVEATS Pin:** sha256:67b14cf311ee874f45f5a744c9c55ec1eefd72686f4931d81c764be16ec765b
(`study_02_replication.json` , 2,876,012 bytes)

0. Independence statement

Fresh bundle, fresh working directory, all code written for this run. The SG `sg-audit-gates` skill (built 2026-11-01) was loaded and its applicable rules followed (see §7 for the gap: its `scripts/` do not exist in the repo, so gates were executed by hand-written equivalents). Prior SG audits read only after numbers were fixed. Same operator session as prior runs — model re-derivation, not a separate party. A hostile critic reviewed the draft; its valid demands were executed, its overreach rebutted with evidence (§6).

1. Stage 1 — Deep reading

Design card pre-registers H1 (late-cohort W slope) and H2 (W×Post interaction) with an explicit anti-pivot rule: null results must be stated as falsified/null. The manuscript complies — H1 is reported as "falsified/attenuated" (+6.66, t=+1.20) and H2 as null (-6.68, t=-0.94). v1 (N=40, 10 overlaps, rejected) and v2 (N=48, gap 440d, β=-33.16) are archived verbatim. The design card admits v1/v2 used synthetic hash-jittered judge scores — replaced in v3 with 4 genuine judges (3 OpenRouter LLMs + deterministic linguistic analyzer). **Independence note on the brief:** I was asked to audit "v2.0, N=48". The bundle served is v3.0, N=180. I audited what exists, not what was briefed.

2. Stage 2 — Carrying out the study

(a) Provenance (SKILL-016/025): 180/180 valid accession numbers, SEC URLs, CIK cross-matches, filing-date == event-date; 180/180 unique texts (929–2,492 chars); **0/180 true anachronisms** — 43 naive year-hits triaged: 40 forward guidance, 3 fiscal-year labels verified against company fiscal calendars (PATH FY2024 closed Jan-2024 discussed Mar-2023; CRWD Q2 FY2024 ended Jul-31-2023; SNOW FY2026 guidance in late-2025). SHA-256 text-hash check: 0 mismatches events↔stored↔ballots. **(b) Composition & timing (SKILL-006/017/021):** N=180, M=45, exact 90/90 regimes with **all 45 firms in both regimes, exactly 2 events each** (within-firm balanced panel — kills composition-change objections); min same-ticker gap 161d, zero overlaps; timing validated: 168/180 after-hours (flag matches recomputation), reaction = next trading day; no exactly-16:00:00 edge cases. Tails match to 2 decimals (day1 [-26.94, 33.65] max|z|=3.07; 30d CAR [-27.96,

22.96] max|z|=3.41, kurtosis +1.01). **(c) Econometrics — every headline number reproduces:** early W +15.24 (t=+2.56); early W bivariate +11.61 (t=+2.99); early H bivariate +4.94 (t=+2.02); late W +6.66 (t=+1.20, n.s. — H1 null as reported); late V null; C1 shift -3.052 (t=-2.14); C2 unbacked shift -4.688 (t=-2.51); C3 AI-prom day1 shift -5.269 (t=-2.31); day0 placebo -0.35 (t=-0.71) vs true day1 -5.56 (t=-2.56). **(d) Reliability (SKILL-019):** raw ballots ship for all 180 units. The bundle's ICC function reproduces its claimed ICC(2,k) bit-for-bit (0.903/0.908/0.745) — **but the function is non-standard**: "ICC(1,k)" = (MSB-MSW)/MSB, "ICC(2,k)" = (MSB-MSE)/(MSB+k·(MSJ-MSE)/n). Standard Shrout-Fleiss gives H 0.79–0.96, V 0.73–0.95, U 0.58–0.87. Agreement is real; the labels are wrong and the docstring ("exact Shrout & Fleiss") is false. **(e) OOS (SKILL-018):** all shipped table numbers reproduce (living 103/180 57.2% vs blank 98 vs zero RMSE 8.66 vs 7.90; paired ts match). My independent LOYO: my stateless 106/180 vs my 0.25-shrink 96/180 — note the sign comparison is vacuous (scalar shrinkage preserves signs); RMSE 8.29 vs 7.82, paired t=+1.74 (p=0.08, n.s.).

3. Stage 3 — Alternative tests

- **Median regression:** C2 shift -1.94 (p=0.28, n.s.) vs OLS -4.69 (p=0.012) — the unbacked-subsample effect is mean-driven (tails), not median. C3 day1 shift strengthens at the median (-8.04, p=0.002). C1 median -2.26 (p=0.065).
- **Within-sector permutation (2000×):** C1 shift p=0.040 — the regime effect survives label shuffling.
- **Negativity control:** judge blame proxy — Day-1 blame survives generic negative-word control (reported in v4 audit; same layer here).
- **Fresh data:** 14 SEC filings fetched from sec.gov (6 new this run): 121/145 sentences verbatim (cumulative 20 filings, 211/237). Yahoo adjclose re-pull, 12 cases incl. both extremes: 12/12 exact.
- **Leakage (SKILL-008):** forecast code uses only H,V,U,W,day1,xbrl,mom (ex-ante); no realized labels in features. Rule constants are hand-set (unverifiable as OOS) — flagged, not fatal, since OOS evaluation itself is chronological.

4. Response to hostile critic

Adopted: H2 exact-number irreproducibility (tried 12+ specs/clusterings; null holds everywhere), ICC mislabeling (verified against the bundle's own function), OOS-vs-zero framing, 10-Q-era soup check (none found here — v3 is all 8-K). Rebutted with evidence: (i) "no clustered SEs" is false — firm-clustered throughout; (ii) "composition change explains shifts" is false — same 45 firms, 2+2 events each; (iii) "uncontrolled splits" is false — C1/C2/C3 all control day1/XBRL/mom/vol/tone/sector; (iv) PARTIALLY_REPLICATED overreaches — the paper's surviving positive claims reproduce and its nulls are honestly reported nulls (the anti-pivot rule was obeyed). Verdict stays REPLICATED_WITH_CAVEATS, not PARTIALLY_REPLICATED: nothing the paper asserts as a finding failed to reproduce; what failed (H1/H2 as positive claims) the paper itself declares dead.

5. Verdict: REPLICATED_WITH_CAVEATS

Boston v3.0 is a genuine rebuild that fixes every v1/v2 flaw the SG lab caught (verbatim texts, no overlaps, adjusted closes, real judges, honest nulls). Required fixes: (1) correct the ICC formulas or relabel them (do not claim Shrout-Fleiss); (2) disclose the correct W_unclipped formula incl. +0.50 intercept (design card omits it); (3) fix the stale G counts (C2 G=40 not 42; C3 G=36 not 35 — N and all statistics unaffected); (4) either reproduce the H2 table cell exactly or replace it with a spec that does; (5) qualify OOS claims (beats zero/buzzword, not unshrunk OLS; hit-rate 103/180 p=0.053); (6) account for audit-trail numbers 008/010/011 or stop citing the full 001–011 range (I could verify 001–007 and 009; the rest are unaccounted for).