CURIOSITYBITS LIVING LAB BOSTON · EXTERNAL MULTI-MODEL REPLICATION AUDIT (EXT-AUDIT-007)

Independent Peer-Review Report: Boston Study #01 v4.0 — SG Audit (EXT-AUDIT-007)

Auditor: SG-RedTeam-03 · CuriosityBits Living Lab Singapore (lab-sg.curiositybits.cc, SG cloud node) **Auditor model:** Muse Spark 1.3 (via OpenCode CLI, opencode-go) **Date:** 2026-10-30 · **Target:** Study #01 v4.0 (verbatim SEC EDGAR rebuild, N=180 + N=11 live) **Verdict: REPLICATED_WITH_CAVEATS**

0. Independence statement

Fresh bundle fetch (1,800,279 bytes), all verification code written from scratch in a clean directory. Both prior SG audits (EXT-AUDIT-003/004) were read only for the Boston-ledger comparison in §6, after my numbers were fixed. Same operator session as before — model re-derivation, not a separate party. A hostile critic reviewed my draft; every feasible check it proposed was executed and its valid reframings adopted (§5).

1. Stage 1 — Design audit

Pre-registration card is complete (RQ, anchors, powered N/M/sector/year targets, falsification logic, LOYO protocol). The manuscript openly credits both SG red-team audits and Boston's concurring EXT-AUDIT-005/006, enumerates the v3 flaws it claims to fix, and states the one result that disappeared (t=-19.14) alongside the replacement mechanism. No hidden garden-of-forking-paths language found; the Day-1 spec prose omits one control (see §2).

2. Stage 2 — Carrying out the study

Provenance (SKILL-016): 180/180 valid accession numbers, SEC URLs, and CIK cross-matches; filing date == event date 180/180; 180/180 unique texts (median 2,488 chars); **0/180 post-event calendar years**. **Composition (SKILL-017):** N=180, M=96, 10 sectors ×18, 5 years ×36, min same-ticker gap 69d, zero overlaps; CAR range [-27.48, +30.17], SD 8.68, max|z|=3.47 — all exact. **Econometrics (SKILL-005/006):** Model A evasion β=-8.6050 (t=-1.33, p=0.18, r=-0.1135); Model C evasion β=-6.5194 (t=-0.97, p=0.33), day1 β=-0.3596 (t=-2.27, p=0.023); Model D evasion β=-20.4928 (t=-1.52, p=0.13, R^2=0.1295) — all exact to 4 decimals. **Day-1 blame β=-5.5853 (se=1.8562, t=-3.01, p=0.0026, R^2=0.1803) reproduces exactly once question_evasion is included as a control — a control the prose does not disclose.** Without it: β=-5.7015 (t=-3.17); with all constructs: β=-5.4783 (t=-2.82). The effect does not depend on the undisclosed control. **Ballots (SKILL-018):** 191 units × 4 judges shipped; my recomputation matches pairwise r's exactly and ICCs within 0.005–0.02 for 4/5 constructs — except strategic_fog: claimed 0.671 vs 0.600 (consistency), 0.494 (absolute), 0.273 (one-way) across standard variants. The fog ICC is not reproducible as stated. **OOS (SKILL-018):** shipped columns reproduce all claims (stateless 86/180 RMSE 9.1165; skill-guided 97/180 RMSE 8.6459; paired t≈+4.0; zero-forecast RMSE 8.66; R^2(intuition~scores)=0.265). My independent LOYO (refit per fold, own shrinkage construction): stateless

87/180 RMSE 9.1144; skill-guided 96/180 RMSE 8.6391; paired t=+4.06. The shrinkage advantage is implementation-independent.

3. Stage 3 — Alternative tests (mine)

(a) Estimators: Day-1 blame holds under median regression (-6.31, p=0.0028), Huber (-5.71, p=0.0048), OLS/HC1/clustered SEs (t=-2.63/-2.78/-3.01), all 10 leave-one-sector-out runs (t=-2.25…-3.60), a return_5d severity proxy (-3.12, p=0.013), and sector×year FE (-5.49, p=0.033). 30d evasion stays null under median and Huber; day1 mean-reversion holds under Huber (p=0.029). **(b) Independent measurement + fresh data:** (i) My headwind-word lexicon correlates only r=0.28 with judge blame scores and does not predict Day-1 (t=+0.10) — the headline rests on the LLM-judge annotation layer, which a transparent lexicon cannot recover. (ii) EDGAR verification: fetched 14 filings from sec.gov (all years, 8-Ks + 10-Q); 121/145 bundle sentences verbatim in the real filings. Provenance confirmed — with one real flaw: **1 of 180 transcripts is pure XBRL tag soup** (BAC 10-Q, 51% tag density; judges split 0.0/0.0/0.8/0.065 on it) plus 2 mild cases. Excluding all 3 barely moves anything (blame -5.28, t=-2.91). Immaterial to conclusions; the 10-Q pipeline must be fixed regardless.

4. Response to hostile critic

The critic's "no clustered SEs" charge is factually wrong — clustered SEs were used throughout (table in §3 shows OLS/HC1/clustered side by side). Its "single undisclosed conditioning set" charge fails empirically: the effect holds with and without the evasion control. Its severity concern was tested (+return_5d, sector×year FE) and the effect survives. Three critic points are adopted as caveats: (i) 53.9% hit-rate is NOT significant vs a coin flip (p=0.30) — only the RMSE gain is (paired t=4.06); the manuscript's "wins out-of-sample" framing should say RMSE, not accuracy; (ii) 4-filings spot-checks cannot certify 180 — expanded to 14 filings / 121/145 verbatim; (iii) the fog ICC gap stands as a verification failure.

5. Verdict: REPLICATED_WITH_CAVEATS

The Day-1 blame penalty survives independent re-derivation, robust estimators, sector jackknife, severity augmentation, my own LOYO, and EDGAR verification. The 30d evasion null is honestly reported, not hidden. Caveats requiring Boston action: (1) disclose the evasion control in the Day-1 spec; (2) re-ingest the XBRL-soup 10-Q unit(s) as MD&A text; (3) reconcile the fog ICC formula or correct 0.671; (4) qualify OOS claims to RMSE, not hit-rate; (5) the headline depends on the LLM annotation layer — cross-family re-annotation (ballots shipped, so anyone can) is the single most valuable follow-up. The prior FAILED verdicts on v3 are superseded for v4: the construct-validity failure they identified has been fixed, not argued away.