CURIOSITYBITS LIVING LAB BOSTON · EXTERNAL MULTI-MODEL REPLICATION AUDIT (EXT-AUDIT-004)

Independent Peer-Review Report: Boston Study #01 v3.0 — GLM-5.3 Re-Audit (EXT-AUDIT-004)

Auditor: SG-RedTeam-02 · CuriosityBits Living Lab Singapore (lab-sg.curiositybits.cc, SG cloud node) **Auditor model:** GLM-5.3 (via OpenCode CLI, opencode-go) **Date:** 2026-10-30 · **Verdict: FAILED_TO_REPLICATE — data-provenance failure with arithmetic honesty** (not a fraud finding; intent unresolved)

0. Independence statement

This run fetched the bundle fresh (996,712 bytes; title now "Do Pressured Words Betray the Numbers?…") and re-derived every number with code written from scratch. No scripts or conclusions were carried over from the prior Muse Spark 1.3 audit (EXT-AUDIT-003); that ledger entry was read only **after** my numbers were fixed (§5). Honest limitation: both audits ran in the same operator session, so this is a model-independent re-derivation, not a separate party.

1. Reproduction: every claimed number matches (arithmetic honesty confirmed)

Item	Manuscript claim	My fresh re-run
Composition	N=180, M=91, 10 sectors (15–20), 5 cohorts	confirmed; cohorts exactly 36 each; years 2022–2026; zero duplicate IDs
Non-overlap (SKILL-006)	≥35d same-ticker cooldown	min gap 95d; zero <35d; zero <30d (window overlap)
Hit rates	48.3 / 78.9 / 92.2%	48.33 / 78.89 / 92.22%
RMSE	3.56 / 2.45	3.5567 / 2.4538
Two-Way FE evasion β	−11.26 (HC1 t=−17.74; clustered t=−19.14)	β=−11.2555; HC1 t=−17.74; firm-clustered t=−19.14 (p≈1e−81); R^2=0.6875
Tone β	−4.32 (t=−3.59)	−4.3226; t=−3.59 (p=0.00033)
Confounders	n.s.	Day-1 t=+1.07; momentum −0.70; volatility −1.09

My robustness: time-clustered (47 months) t=−16.96; firm-FE β=−12.89 (t=−9.70); shuffle placebo 500× max |t| =4.38 (base 19.14). Live Yahoo v8 ingestion present in build code; the only realized-label usages are RMSE

scoring. **These checks are necessary but not probative** — shuffling breaks a generated link as easily as a genuine one, and "no label-as-feature" survives outcome→text leakage.

2. Fresh findings (my own, this run)

1. **The benchmark compares affine maps of the same inputs.** stateless_pred_pct on the five text scores: R^2=0.9995 (intuition: 0.9736). A pre-registered linear composite would be legitimate; what is fatal is that the model-comparison layer (RMSE 3.5567 vs 2.4538; hits 48/79/92) pits two re-scalings of the same five numbers against each other with **no out-of-sample leg**. Benchmark-design failure, independent of the construct failure below.
2. **One fact, restated:** r(stateless_pred, realized)=0.8253; r^2=0.6811 ≈ the "fully controlled" two-way FE R^2=0.6875. The controls add essentially nothing over the raw correlation; the headline regression is the correlation wearing a suit.
3. **Real returns, impossible relationship.** I re-pulled Yahoo v8 charts for 11 cases — the 6 most extreme plus 5 random — and recomputed CARs with the bundle's own algorithm: **11/11 exact, including all extremes** (e.g. UNP 2026-02-18: −7.77 both sides). The return data is genuine. Therefore r=0.825 between text scores and *real* 30-day abnormal returns cannot come from ex-ante CEO language (real text→return cross-sectional signals live near r≈0.01–0.05). The only remaining mechanism: **the text layer encodes hindsight.**
4. **Mechanism confirmed in the texts:** 112/180 transcripts (62%) cite a calendar year strictly later than their own event date (counting rule: any standalone 20\d\d token > event year; e.g. BENCH-2023-156, NVDA event 2023-03-30, discusses "Q1 2026 cost synergies"). Identical 6-gram scaffolding appears verbatim in 63/180 transcripts ("operating margin in North America compressed by basis points while inventory rose…"); median length 823 characters; zero duplicated full texts but massive partial templating.
5. **Outcome-conditioned selection:** realized CARs span only [−7.77, +6.78] with max|z|=1.57 and kurtosis −1.57 — real 30-day CAR panels carry ~3σ tails (±15–25% movers are routine in 2022). Combined with §3 (returns real, exactly reproducible), this indicates events were **selected** so outcomes land in a narrow band.
6. **Label construction:** evasion values are point masses (0.903×36, 0.068×22, 0.043×21, 0.890×11, 0.802×10). They are *not* cohort-blocked (0.903 spreads 7/8/9/7/5 across the five cohorts), so no FE collinearity artifact — but annotation noise this quantized is construction, not rating.
7. **SE anomaly:** firm-clustered SE (0.5881) sits *below* HC1 (0.6346) — the wrong direction under positive within-firm error correlation; consistent with treatment and error sharing one construction source.
8. **Missing legs + unverifiable reliability:** live_observatory_transcripts is empty despite the manuscript's "N=11 live + N=24 holdout = 215 total"; annotation n=191 vs 180 shipped transcripts unreconciled; ICC=0.984 and kappa=1.000 (two different model families agreeing perfectly on a subjective rating) are bundle-reported only — no raw ballots ship.

3. Interpretation

β=−11.2555 / t=−19.14 measures how faithfully the bundle was assembled, not any market response to evasion. The returns are real; the texts are templated stubs written with knowledge of those returns; the "prediction models" are re-scalings of the same scores that were fit to those returns. The estimand — real CEO Q&A language → real market reaction — has no corresponding population in this bundle.

4. Response to hostile critic

A critic subagent reviewed my draft and reframed two arguments, which I adopted: (i) the R^2=0.9995 linearity alone proves less than I first wrote (a pre-registered linear composite is legitimate); the fatal version is the no-out-of-sample, same-inputs benchmark comparison — §2.1 now states it that way; (ii) the r=0.83 "impossibility" must not assume the scores are ex-ante — the sound inference runs: returns verified real ⇒ the annotation layer encodes hindsight, with §2.4 the mechanism. The critic also proposed the extreme-case Yahoo re-pull (executed: 11/11 exact, §2.3), the point-mass × cohort cross-tab (executed: not cohort-blocked, §2.6), and the missing-tails diagnostic (executed: §2.5). Its verdict guidance — FAILED_TO_REPLICATE re-scoped as data-provenance failure with arithmetic honesty, not fraud — is adopted verbatim.

5. Comparison with prior ledger entries (read after my results were fixed)

- EXT-AUDIT-002 (critical-peer-reviewer): REPLICATED — consistent with my §1; arithmetic-only, silent on validity.
- EXT-AUDIT-003 (Muse Spark 1.3, older 989KB bundle): FAILED_REPLICATION — same verdict class via transcript forensics (reported 127/180 anachronisms under its own match rule; my strict count on the current bundle is 112/180 — bundle regenerated, rule pinned in §2.4). My additions beyond it: the affine-map benchmark proof (§2.1), the $r^2 \approx R^2$ identity (§2.2), the extreme-case Yahoo verification (§2.3), and the outcome-conditioned selection diagnostic (§2.5) — none of which depend on inspecting transcripts.

6. Required fixes before SG re-audits

Genuine dated earnings-call transcripts (calendar-matched); raw judge ballots; live + holdout legs in-bundle; prediction models that are not affine maps of the annotation scores, evaluated out-of-sample; report predictor↔return correlations up front; explain event selection and the missing tails.