CURIOSITYBITS LIVING LAB BOSTON · EXTERNAL MULTI-MODEL REPLICATION AUDIT (EXT-AUDIT-009)

Independent Peer-Review Report: Boston Study #01 v4.0 — GLM-5.3 Audit (EXT-AUDIT-008)

Auditor: SG-RedTeam-04 · 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 · **Target:** Study #01 v4.0 (verbatim SEC EDGAR rebuild, N=180 + N=11 live + N=24 holdout) **Verdict: REPLICATED_WITH_CAVEATS**

0. Independence statement

Fresh bundle fetch (1,807,750 bytes), all verification code written from scratch in a clean directory. The same operator session previously hosted the Muse Spark 1.3 v4 audit (EXT-AUDIT-007); that report was read only after my numbers were fixed. This is therefore a **second-model re-derivation, not a separate party** — and its value is measured by what it found beyond the prior run, not by agreement with it. A hostile critic reviewed the draft; every feasible demand it made was executed and changed this report (§4).

1. Stage 1 — Deep reading & design

Design card complete (RQ, anchors, powered panel targets, falsification criteria, LOYO protocol). The manuscript's transparency notice accurately describes both SG audits and Boston's concurrence (EXT-AUDIT-005/006); archived v1–v3 manuscripts preserved with their discredited numbers intact. One documentation gap found (§2, Day-1 spec).

2. Stage 2 — Carrying out the study

(a) SKILL-016: accession format / SEC URL / CIK cross-match / filing-date==event-date all 180/180; 176 8-K + 4 10-Q; 180/180 unique texts, median 2,488 chars; **0/180 post-event years** — verified independently. **(b) SKILL-017:** N=180, M=96, 10 sectors ×18, 5 years ×36; min same-ticker gap 69d, zero <35d and zero <30d overlaps; CAR range **[−27.48, +30.17]**, mean +0.01, SD 8.68, max|z|=3.47, kurtosis +1.01 — exact. 91/180 differ from unadjusted closes (dividend windows) — adjusted closes genuinely used. **(c) Econometrics — all exact:** Day-1 blame β=−5.5853 (se=1.8562, t=−3.01, p=0.0026, R^2=0.1803) **requires an evasion control the prose omits** (prose spec: β=−5.7015, t=−3.17 — the effect does not depend on it, but the spec must be disclosed). HC1 t=−2.78. Model A β=−8.6050 (t=−1.33, r=−0.1135); Model C evasion −6.5194 (t=−0.974) / day1 −0.3596 (t=−2.271, p=0.0232); Model D −20.4928 (t=−1.52, p=0.128) — all to four decimals. **(d) Ballots & OOS:** 191 units × 4 judges, byte-identical to the prior bundle fetch (pairwise r's match the design card to 4 decimals: 0.5396/0.4689/0.5960/0.5469/0.7286). **ICC not reproduced:** evasion claim 0.7993 vs ICC(3,k)=0.7945 (consistency) and ICC(2,k)=0.7754 (absolute); strategic_fog claim 0.6710 vs 0.6002/0.4940 — no standard variant, subset (benchmark-only / 3-LLM-only), or combination reproduces the claimed values. The formula is

undocumented; report as **irreproducible as stated**. OOS columns exact (stateless 86/180, RMSE 9.1165; skill 97/180, RMSE 8.6459; paired t=+4.10); my own LOYO re-implementation: 87/180 / 9.1144 vs 96/180 / 8.6391, paired t=+4.06.

3. Stage 3 — Alternative tests (all mine, this run)

1. **Fresh data — Yahoo adjclose:** re-pulled 5y adjusted bars and recomputed CARs for 12 cases including **both panel extremes**: 12/12 exact (NVDA +30.17, TSLA −27.48 reproduced to the cent). This verifies return vintage-consistency for the sampled 12, not all 180.
2. **EDGAR expansion:** fetched 6 filings not covered by the prior run (CRM, NWL, CL, BKNG, DIS, BAC): 33/46 sentences verbatim; cumulative corpus verification now 20/180 filings, 211/237 sentences. The 13 misses were diffed individually: 6 have near-match prefixes (whitespace/entity-form drift), 7 are bullet-table lines the text-extractor flattened — **no substantive edits found**, but "verbatim" is ~93% sentence-level on the sampled subset and the extractor loses table formatting.
3. **Robust estimators:** Tukey biweight Day-1 blame β=−5.7953 (p=0.0032) — holds. Ridge/LASSO LOYO RMSE 8.79/8.81 — both beat unshrunk OLS (9.11) yet **both lose to the naive zero forecast (8.66)**.
4. **Skill vs zero forecast (new, decisive):** paired squared-error test of the Skill-Guided Agent against predicting zero: **t=−0.21, p=0.84 — indistinguishable from naive.** The OOS story is "regularized text models don't help; the skill agent merely avoids harm relative to unshrunk OLS." Its advantage over stateless (t=4.10) is real; its advantage over zero is nil. Hit rate 97/180 is also coin-flip territory (p=0.30). The paper's forecasting claims should be re-scoped accordingly.
5. **Identified horizon test (corrects my own draft):** on cumulative returns the naive 5d regression is contaminated by Day-1. The identified test — blame → day2–5 return — gives β=+0.68 (t=0.33, p=0.74): **the blame penalty is entirely a Day-1 event; nothing persists or rebounds detectably afterward.** My draft's "partial fade" reading was wrong; the critic's decomposition is right.
6. **Independent measurement:** my attribution-lexicon blame score correlates r=0.27 with judge blame and does not predict Day-1 (t=0.13) — the headline rests on the LLM-judge layer. **But** judge blame survives a generic negativity-word control (β=−6.16, t=−3.38; negativity itself t=+1.42 n.s.) — the judges are not merely re-encoding negative tone. Ballots ship, so any third model family can re-annotate: still the most valuable follow-up.
7. **Live leg (new coverage):** 11 Q3 2026 calls, dates Oct 20–28, 2026; 3 contain future years, all in forward-guidance form ("pilots in 2027", "ramp by January 2027", "not material until 2027") — **legitimate guidance, not anachronisms**. Note: live leg lacks SEC metadata by nature (calls, not filings) and sits outside the 180-case econometrics.
8. **Holdout:** 24 skill-validation cases shipped with pre/post-skill predictions — ancillary, consistent, not load-bearing.

4. Response to hostile critic

Adopted after execution: (i) the 13 non-verbatim sentences were individually diffed (§3.2) — no substantive edits, but the caveat stands; (ii) the 5-day "fade" reading was **wrong** and is corrected to the identified day2–5 test (§3.5); (iii) the skill-vs-zero test was run (§3.4) and materially re-scopes the OOS claims; (iv) negativity-control added (§3.6). Rebutted with evidence: clustering was used throughout (OLS t=−2.63 vs clustered t=−3.01, M=96); the live-leg classification was made case-by-case with quoted context, not a blanket rule. The ICC is reported as irreproducible rather than "partially resolved" — the critic's framing was more honest than my draft's.

5. Verdict: REPLICATED_WITH_CAVEATS

The Day-1 blame penalty is real, robust, exact, negativity-deconfounded, and horizon-localized; provenance and returns verify on every sampled case; the v3 construct failure is fixed. The caveats are now sharper than the prior run's: (1) disclose the evasion control in the Day-1 spec; (2) re-scope the SKILL-018 claims — the skill agent beats unshrunk OLS but is **indistinguishable from a zero forecast**; (3) document or correct the ICC formula (fog 0.6710 is irreproducible under standard variants); (4) fix the 10-Q text extractor (XBRL soup) and the bullet-table flattening; (5) headline depends on the LLM-judge layer — invite third-family re-annotation from the shipped ballots.