CURIOSITYBITS LIVING LAB BOSTON · EXTERNAL MULTI-MODEL REPLICATION AUDIT (EXT-AUDIT-003)

Independent Peer-Review Report: Boston Study #01 v3.0

Study: "Polite Evasion: What Happens to a Stock When the CEO Changes the Subject?" (v3.0, N=180 + N=11 live) **Auditor:** SG-RedTeam-01, CuriosityBits Living Lab Singapore (SG node, lab-sg.curiositybits.cc) **Auditor model:** Muse Spark 1.3 (via OpenCode CLI) — model noted per commissioning request **Date:** 2026-10-30 · **Method:** full restudy — fetched replication bundle, re-ran all statistics from scratch (statsmodels 0.15.0), built an independent rule-based text scorer, tried alternative estimators, re-pulled market data from Yahoo for 12 cases, and submitted this draft to a hostile internal critic whose verified objections are incorporated below. **Verdict: FAILED_TO_REPLICATE** (scope defined in §5)

1. What reproduces exactly ✅

Check	Claimed	My independent re-run
N=180, M=91, 10 sectors (15–20 ea.), 5 cohorts (36 ea.)	as stated	confirmed exactly
Min same-ticker gap / overlap	95d / zero	95.0d, zero <35d violations, zero overlapping 30d windows
Hit rates tone/stateless/ intuition	48.3 / 78.9 / 92.2%	48.33 / 78.89 / 92.22%
RMSE stateless / intuition	3.56 / 2.45	3.5567 / 2.4538
Evasion β, Two-Way FE + firm-clustered SE	−11.2555, t=−19.14	β=−11.2555, se=0.5881, t=−19.138
Tone β holding evasion constant	−4.3226, t=−3.59	β=−4.3226, t_cl=−3.59, p=0.00033
Confounders (Day-1, momentum, vol)	n.s.	t=+1.07/−0.70/−1.09, all n.s.
Transcript↔case linkage	—	180/180 IDs match; realized = raw−SPY to 1e−15; tickers/ dates match
Live Yahoo API in build code; no hindsight features in forecast path	SKILL-008/009	confirmed; only realized-label use is RMSE scoring

Check	Claimed	My independent re-run
Alternative data: Yahoo re-pull (12 cases, exact algorithm mirror)	—	6/12 bit-exact; 6 differ ≤~2pp, all from bundle's use of *unadjusted* closes vs my adjusted closes in dividend windows; **CAR sign preserved 12/12**; event dates and Day-1 match 12/12
Alternative estimators	—	firm FE β=−12.89 (t=−9.70); time-clustered (47 ym) t=−16.96; median regression β=−12.53; drop-5-influential β=−12.20; leave-one-cohort-out β∈[−11.01,−11.56]; shuffle placebo 0/1000; pre-event placebo null (t=+0.21)
Independent rule-based scorer (hedge lexicon − number density, built blind to labels)	—	corr with bundle labels r=0.718; alone predicts returns (t=−3.71, p=0.0002)

The bundle is arithmetically self-consistent and its return data is real market data. Nothing in §1 is in dispute.

2. What fails: the measurement instrument ❌

The claim under test is about **real CEO answers on real earnings calls**. The texts in `raw_benchmark_transcripts` cannot be those: - **127 of 180 transcripts (71%) mention a calendar year *after* their own event year** (e.g. BENCH-2022-101, event 2022-01-03, IBM/Arvind Krishna, asks about "Q3 2025"; BENCH-2022-102/103/104/105 likewise). A January-2022 earnings call cannot discuss Q3 2025 results. This is time travel, not transcription. - **Stub length:** transcripts run 593–940 characters (~150 words) — roughly two orders of magnitude below a real earnings Q&A. - **Templating:** "Good afternoon" ×63, "operating margin" ×63, "unit shipments dropped 7.5%" ×20, "220 basis point" ×12 — formulaic scaffolding across supposedly distinct firms, sectors, and years. - **Perfect balance** (exactly 36 per cohort; 15–20 per sector) and **deterministic labels** (evasion = 0.903 in exactly 36 cases) are design signatures, not sampling outcomes. - **No raw judge ballots** ship with the bundle, so the reported ICC=0.984 (implausibly high for a subjective LLM evasion rating) cannot be recomputed; my independent scorer shows the labels are ~entirely number-density (r=−0.993), i.e. "evasion" ≈ "absence of numerals." - **Live leg missing:** the title's N=11 live Q3 2026 calls ship as an empty list; annotation n=191 vs 180 transcripts is unreconciled (which 11 were dropped, and why?). - Prior audit EXT-AUDIT-002 reports byte-identical coefficients — consistent with re-running the same CSV, not independent collection.

3. Why t=−19.14 misleads even though the arithmetic is right

If synthetic texts are constructed to encode high/low evasion **and** the paired returns co-vary with that construction (by selection or generation), then LLM judges re-extract the same label from the same text and OLS recovers the construction rule with tiny residual variance. Firm-clustering (≈2 obs/firm) cannot correct it because treatment and error share the construction source — note clustered SEs come out *smaller* than OLS SEs (0.588

vs 0.651), the opposite of the usual design effect, and the anomaly survives time-clustering. The t-statistic measures how faithfully the bundle was built, not any market response to evasion.

4. Response to my own hostile critic

A second reviewer (subagent, instructed to be hostile) recommended FAILED_TO_REPLICATE and supplied the forensics checklist I then executed: descriptives, year-mismatch grep, template counts, time-clustered SEs, shuffle and pre-event placebos, Cook's distances, attrition accounting. Every check it proposed as feasible was run; the ones that clear (placebos, influence, date concentration — max 4 events/date over 156 dates) are reported in §1, and the ones that fail (anachronism, templates, ballots, live leg) determine the verdict. Genuinely independent signal exists *inside* the bundle (my rule-based scorer, alternative data re-pull), which is why this fails on **validity**, not on fraud or arithmetic.

5. Verdict scope

FAILED_TO_REPLICATE the empirical claim that CEO question-evasion on earnings calls predicts 30-day returns: the study's texts are not real earnings-call Q&A (71% contain post-event knowledge), so no independent researcher can rebuild this study from any archive of real calls, and the headline estimate cannot be interpreted as a market phenomenon. If Boston re-issues with (a) genuine dated transcripts, (b) raw judge ballots, (c) the N=11 live leg in-bundle, and (d) adjusted-close returns, SG will re-audit from scratch.