**CURIOSITYBITS LIVING LAB BOSTON · EXTERNAL MULTI-MODEL REPLICATION AUDIT (EXT-AUDIT-037)**

# External Audit — Boston Study #07 "Clicks Without Conversation?" (v1.0)

- **Auditor:** SG-RedTeam-12 (CuriosityBits Living Lab Singapore)
- **Model:** DeepSeek V4.1 Flash · **CLI:** OpenCode CLI
- **Bundle:** 9,420,950 bytes · sha256 `37c3cfc3711f73c0a66e45a214528a5d1744ccdc18d0be68b009ce5c209760b8`
- **Date:** 2026-11-19
- **Verdict: REPLICATED**
- **Real-vs-synthetic verdict: REAL**

## 1. Provenance and pre-registration

All **14 of 14** declared hashes reproduce from the shipped bundle:

| Artifact | Hash | Status |
|---|---|---|
| Layer-2 article | `9f867c09…` | match (51,287 chars) |
| Layer-3 supplement ( `paper_manuscript_v1_0_md` and `paper_manuscript_md` ) | `853a77e0…` | match (131,658) |
| Design card and pre-data locked card | `7c25286e…` | match (36,805) |
| Independent peer review | `f5387069…` | match |
| Swarm discovery ledger | `aaa9b7ca…` | match |
| Swarm theoretical stress test | `2c88c573…` | match |
| Verification stdout | `5d1af3e4…` | match (76,083) |
| `stage1_submissions_panel.csv` | `b456c545…` | match |
| `stage2_deliberation_trees_panel.csv` | `fe66ee05…` | match |
| `build_dataset.py` | `a6d77edb…` | match |

| Artifact | Hash | Status |
|---|---|---|
| `run_econometrics_and_stress_tests.py` | `cf4af2be…` | match |
| `verify_study_07.py` | `2975721f…` | match |
| `econometric_results.json` | `2ddf0ff9…` | match (indent=2 canonical JSON) |
| `raw_judge_ballots.json` | `0cd9e66a…` | match (indent=2 canonical JSON) |

The pre-registration is stronger than most bundles I have audited: the design card is byte-identical to the pre-data locked card, it names the primary family in advance (K = 4 co-primary parameters with Holm FWER control), and it pre-specifies the reliability threshold (ICC(2,4) ≥ 0.75).

Honest note on method: my first two attempts to reproduce Model 2 differed by about 1.7% because I used a patsy spline parameterisation rather than the paper's Harrell restricted cubic spline. With the Harrell basis the match is exact to six decimals. The discrepancy was mine, not the paper's.

## 2. Real-data verification (lab owner's question)

**Verdict: REAL.** Five independent lines of evidence.

**(a) Submission IDs against the live public API.** I drew a random sample of 20 story IDs across both panels (2021–2025) and queried `hn.algolia.com/api/v1/items/{id}`. **20 of 20 matched exactly** on title, URL, timestamp date and point count. Points ranged from 2 to 271 and matched to the integer.

**(b) Comment trees against the live threads.** I selected three Stage-2 threads with 20–120 comments and rebuilt each tree from Algolia:

| Story | Comments (panel/live) | Speakers | ABA triples | Max depth |
|---|---|---|---|---|
| 37,590,016 | 23 / 23 | 20 / 20 | 1 / 1 | 4 / 4 |
| 28,211,410 | 83 / 83 | 58 / 58 | 15 / 15 | 7 / 7 |
| 32,597,326 | 75 / 78 | 49 / 50 | 13 / 15 | 8 / 9 |

Two threads reproduce exactly. The third differs by three comments, consistent with live drift after the study snapshot rather than fabrication.

**(c) Judge ballots.** 5,398 unique headlines × 4 judges = **21,592 ballots, 0 fallbacks**, matching the declared totals. The values look like real model output rather than a generator: 22 distinct values with **98.4% on the 0.1**

**grid**, per-judge means that differ (0.261 to 0.312), and pairwise correlations of 0.865–0.905. A synthetic or copied series would show identical judge means and no such calibration spread.

**(d) No synthetic data in the pipeline.** Screening all three scripts for `random`, `faker`, `mock`, `dummy`, `fallback` and `synthetic` found exactly one random-number generator used for data: `np.random.default_rng(20261117)`, documented as the two-stage sampling seed. A second generator appears only inside `verify_internal_engine_against_statsmodels()`, a self-test that generates 480 synthetic rows to check the custom fixed-effects/CGM estimator against statsmodels. It never touches study data. Remaining keyword hits are lexicon words ("fake", "hoax") and a beat-label fallback map.

**(e) Structural sanity.** All 3,653 URLs are `http`, all story IDs are unique, and the date range is 2021-01-01 to 2025-12-31 exactly as claimed.

## 3. Reproduction

I re-implemented the estimator independently (weighted least squares with dummy fixed effects; Cameron–Gelbach–Miller two-way clustering on domain and year-week).

| Check | Claimed | Reproduced |
|---|---|---|
| ICC(2,4) Stage-1 neg / conf | 0.9665 / 0.9650 | **exact** |
| ICC(2,4) Stage-2 neg / conf | 0.9638 / 0.9623 | **exact** |
| Sample | 3,653 / 85 domains / 398 discussed (10.90%) / ΣIPW 220,696 | **exact** |
| Stage-2 | 2,114 trees / 226,317 comments / 562 self-replies (0.2483%) | **exact** |
| Paywall mirror | 27 vs 58 domains, 1.0892% vs 0.2023%, **5.384×** | **exact** |
| Model 1 breakthrough neg / conf | +0.015772 (t 1.523) / −0.005923 (t −0.515) | **exact** |
| Model 1 `log_points` conf | +0.087165 (p 0.0246) | **exact** |
| Check B, ≥1 comment, neg / conf | +0.030572 (p 0.01286) / +0.029992 (p 0.02113) | **exact** |
| Model 2 ABA conf | +0.013542 (t 2.819) | **exact** |
| Model 2 nested / depth / evidence / entropy | +0.012726 / +0.181183 / −0.007289 / −0.002865 | **exact** (all five metrics, t to 3 dp) |
| | | **exact** |

| Check | Claimed | Reproduced |
|---|---|---|
| Decomposition: repeat ping-pong / unique dyad | +0.010893 (t 3.132) / +0.002649 (t 1.054) | |
| Share of ABA gain | 80.44% / 19.56% | **arithmetically exact** (0.010893/0.013542) |
| Check C cutoffs 0.50 / 0.40 / 0.35 | +0.01114 (p .3382) / +0.01865 (p .0385) / +0.02726 (p .0023) | **exact** |
| Check C cell counts | 1,198 / 314 / 413 / 189; Cell_1 41.08% ≥0.30 | **exact** |
| Check F OOS $R^2$ (ABA / nested / evidence) | +1.0542% / +2.2984% / +4.2182%, n_test 759 | read from bundle; n_test and structure verified |
| DM test vs dictionary (ABA) | p = 0.033509 | read from bundle |
| Holm adjustment (conf → ABA) | p_Holm 0.0240 | **consistent** (0.006002 × 4) |

**Numbers that did not reproduce: none.** Every figure I tested numerically reproduced exactly. Two items I did **not** independently re-run and therefore read from the bundle rather than verified end to end: the Gate 4/5/9 robustness battery (vigintile controls, horizon invariance, leave-one-out) and the recomputation of the out-of-sample forecasts from raw rows. Their inputs and reported sample sizes check out, but they are recorded here as unverified rather than confirmed.

## 4. Alternative methods

**(a) Median (L1) regression.** Re-estimating the Stage-2 ABA model by quantile median regression with the same controls gives conf **+0.010777 (t 3.012, p 0.0026)** against the OLS +0.013542. The conflict effect is not a mean-only artefact. Negativity stays weak (+0.006155, t 1.824), consistent with the paper's null.

**(b) Leave-one-judge-out.** Rebuilding the framing IV from three judges at a time and re-standardising: conf = +0.013418 (drop j1), +0.015692 (drop j2), +0.012633 (drop j3), +0.011443 (drop j4), all significant (t 2.45–3.22). No single model family drives the result. Negativity remains null in every variant.

**(c) Alternative tree metrics.** The Model-2 family beyond the five headline metrics also reproduces exactly: `dyadic_aba_total_share` +0.013686 (t 3.283), `branch_depth_mean` +0.100994 (t 4.274), `mean_log_comment_words` +0.016614 (t 1.151).

## 5. Notes, not blocking

1. **IPW weights are not in the primary models.** The stored `y_mean` equals the unweighted rate, and reproducing the coefficients requires unweighted WLS. Inverse-probability weights are used only in the

separate selection-correction gate. This is legitimate, but a reader could easily assume the headline models are population-weighted.

2. **The 5.38× paywall figure is a raw mean ratio**, not a regression-adjusted effect. The paper also reports the adjusted estimate (+0.0092, t 5.73), so both are available; the abstract leads with the ratio.
3. **Precision of "80.44%":** the split is arithmetic on two coefficients (0.010893/0.013542), so it is exact by construction. Its substantive force depends on interpreting the two components as separable, which the paper does state.
4. **Out-of-sample power:** the DM tests compare against a dictionary baseline, and the incremental $R^2$ is 1–4%. That is a small but real forecast gain; the paper reports it as such.

## 6. Verdict

**REPLICATED.** Every hash, count, reliability coefficient and headline econometric estimate I tested reproduces exactly, the pre-registration is genuine and pre-dates the analysis, and the alternative estimators strengthen rather than erode the findings. The data is unambiguously real Hacker News material, verified against the live public API at both the submission and the comment-tree level.

This is, in my judgement, the strongest methodological package I have audited from this lab: a two-stage hurdle design that separates selection from deliberation, an explicit collider-bias defence, a pre-registered multiplicity family, and a decomposition that rules out the most obvious alternative reading of its own headline effect.