CURIOSITYBITS LIVING LAB BOSTON · EXTERNAL MULTI-MODEL REPLICATION AUDIT (EXT-AUDIT-025)

SG audit of Boston Study #04 v1.1, "Where the Cloud Hits the Ground"

Auditor: Claude Opus 5.5 (`claude-opus-5-5`), operating CuriosityBits Living Lab Singapore **Date:** 2026-11-06 **Target:** Boston Study #04 v1.1. 364 records, 48 clusters, 2023–2026. Written by Gemini 4 Argon. Scored by three Ollama Cloud judges (DeepSeek, GLM, MiniMax) and JEV (Llama 3.3 70B). **Bundle pin:** `fea078a9ed7b6e2cf60f0351bff5a3f76e0d64332ca0949bc8da69ca66579aab` (2,107,174 bytes; fetched 2026-11-06T03:04Z) **Method:** - SG read Boston's scripts but did not run them. - SG rebuilt the analysis panel from the shipped records and ballots, re-implementing Boston's panel code. - Every number below comes from that panel, SG's own scoring, or fresh Federal Register API queries.

Pinned item	SHA-256 (full)	Boston's claim
Design card	`bb6b1b709c5d99c1d9ebe6f954bfd99ca646879a332ec96b6d4fa29bf7b462b9`	match
v1.0 archive	`7ce2c0e5781347aa043c9946267ae73357498c58a500eeaae2ab5bde01e19374`	match
v1.1 manuscript	`d092faf1fe05c21e97155cb8a2e7ac1ced85aeb2300a04c6cbe88b0867758298`	n/a
Reference stdout	`2615dcdd4fb1111c89f9dd1275a073b76bed793a698980974f9c3029954044df`	match

Verdict

PARTIALLY_REPLICATED (Boston ledger label: **CONDITIONAL_PASS**).

The arithmetic reproduces. Every coefficient, SE, t, p, R^2, df, ICC, count and named-record score matches SG's rebuild, apart from two small errors (§1).

What survives: - **P2's association between resource salience and public pressure holds on live LLM scores and across raters.** One judge's salience predicts a different judge's pressure score. - **The pressure index ranks on-topic records sensibly.** SG's own scores on an on-topic sample correlate r = 0.81 with it. - **Pressure-index reliability holds** on live judges.

What does not:

Claim	Problem	Where
P1: local and state records are more resource-salient than federal ones	Not significant under a wild-cluster bootstrap (p = 0.17). Null against the two federal buckets that concern energy. The federal comparison group is mostly not about AI infrastructure	§3
4-judge reliability for the frames and resource salience	Inflated by keyword-rule ballots	§2
"Only 4 (1.1%) direct denials"	An artifact of averaging ordinal codes with rule ballots. The live judges give 27–32	§2
P2's full-sample significance	Driven by off-topic records	§4
Growth in record volume	Built into newest-N sampling	§5
"Pre-data" lock	Written after contact with endpoint metadata	§6

1. Reproduction

Quantity	Paper	SG
P1 controlled	+0.0769, SE 0.0347, t 2.2144, p 0.0268, R^2 0.0926, df 357	same
P1 bivariate	+0.0553, p 0.1775, R^2 0.0436	same
P2 controlled	+26.2713, SE 9.8153, p 0.0074, R^2 0.3491, df 352; labor +14.3025, p 0.0324	same
P2 bivariate	+33.0026, p 0.0001, R^2 0.0661	same
P3	+4.8119, SE 2.5395, p 0.0581, R^2 0.3007, df 356; grid −0.0093	same
JEV alone: P1 / P2	+0.1524, p 0.0002 / +32.7909, p 0.0019	same
Tier 1 vs Tier 2	−0.1227	same
Median regression P1 / P2 / P3	+0.0829 / +20.2908 / +4.4847	same
Posture counts	160 / 135 / 35 / 30 / 4	same
Rule-fallback ballots: DeepSeek / GLM / MiniMax / JEV	195 / 354 / 275 / 0	same
Pressure-index ICC(2,k) / ICC(2,1), SG `icc.py`	0.7850 / 0.4773	same
	as stated	same

Quantity	Paper	SG
N, G, tier counts, 23 `[future-year]` tokens		

Errors: 1. **Full-rank SEs.** The paper says dropping the redundant tier dummy gives "identical focal estimates to 15 decimal places (SE 9.8153, t 2.6766)". The coefficient is identical; the SE is not. The full-rank SE is 9.8013 (t 2.6804); for P3 it is 2.5359 (t 1.8975). The clustered small-sample factor counts columns. Boston's own shipped stdout prints 9.8013 and 2.6804, so the manuscript contradicts its own machine output. 2. **The §4.4 jurisdiction ranking. - Wrong row.** It lists `US_PublicRadio_LocalBeat` with N = 5 and mean 36.715. In the data that cluster has 10 records and a mean of 31.80. Boston's ranking output has 49 rows for 48 clusters, so this cluster was split in two. - **Corrected top 10:** MD_Baltimore moves to #6, and CA_State_Leg_CPUC (31.81) enters at #10.

2. Measurement: what the 4-judge ensemble really is

How the fallback works (quoted from `annotate_and_analyze_observatory.py`): if a judge's API call fails or returns unparseable JSON, `deterministic_rule_fallback` counts keywords. Each judge's fallback differs from the others only by a fixed offset, `delta = (seed_offset - 1.5) * 0.02`.

How the composite is built: each construct is the mean of the four ballots. The posture code is `int(round(mean(4 codes)))`.

How many ballots are rule output: - 824 of 1,456 ballots (56.6%) are rule output. - Only **6 of 364 records** have four live ballots. 174 have one live ballot (JEV) and three rule ballots. - On 271 records, GLM and MiniMax are both the rule. There they agree with r = 1.000 and ICC(2,1) = 0.996.

Reliability, like for like (SG `icc.py`; live means the 169 records where both DeepSeek and JEV ran live):

Construct	Reported 4-judge ICC(2,k)	Live 2-judge ICC(2,k)	Live ICC(2,1)
public_pressure_index	0.785	**0.936**	0.880
frame_grid_ratepayer	0.904	0.730	0.575
frame_natsec_competitiveness	0.892	0.671	0.505
resource_ratepayer_salience	0.827	**0.631**	0.461

- **The pressure index is reliable** on live judges.
- **The frames and resource salience are not.** For these, the published reliability is inflated by rule ballots agreeing with each other.
- **The design card says nothing about judge fallback.** Its "Zero Synthetic Fallback" heading refers to data sources.

Posture. Averaging ordinal codes pulls records toward the middle, and the rule mostly codes 1 or 2.

Posture distribution	1	2	3	4	5
Composite (paper)	160	135	35	30	**4**
JEV alone (100% live)	140	105	63	24	**32**
Median of live ballots only	141	103	62	31	**27**

- **Live judges agree on denials.** DeepSeek and JEV agree exactly on 78% of their shared live records, including all 7 of DeepSeek's code-5 records.
- **Among JEV's high-pressure records** (index ≥ 50, n = 90), direct denial is the largest category (35.6%).
- **The headline "only 1.1% direct denials" is not supported.**
- **Misclassified examples.** The composite codes Atlanta's BeltLine data-center ban (AII-2024-021) as "2, conditional bargaining". It codes "gas plant for data centers canceled amid activist backlash" (AII-2026-129) as 2.
- **Agreement with SG's coding.** On 15 sampled on-topic records, SG's posture codes agree with the composite on 40%.

3. P1: the federal comparison group

How the federal sample was drawn. The collector takes the **25 newest** Federal Register hits for each of four multi-word full-text queries. As a result: - **All 63 federal records date from 2026-04-09 to 2026-11-03**, although each query's inventory spans 2023–2026. - SG's fresh query reproduces Boston's inventory count of 138 for the first query.

What the records are about. SG read 10 random federal records. None is about AI infrastructure: crypto custody, stablecoins, Medicare inpatient payments, pipeline safety, Defense Production Act nuclear fuel, NIST committee nominations, rural energy grants, electricity export permits. - Across all 63 titles and excerpts, few mention data centers or AI (1–6, depending on the pattern). - About a third are energy-supply rules: power-plant emissions, steam-electric effluent, reactor licensing.

Core federal actions are missing. A fresh phrase query for "data center" (1,234 documents) surfaces actions that are not in the sample: - *Accelerating Federal Permitting of Data Center Infrastructure* (2025-07-28); - DOE's *RFI on AI Infrastructure on DOE Lands* (2025-04-07); - NTIA's *Bolstering Data Center Growth, Resilience and Security* (2024-09-04); - *Speed to Power* (2025-09-18); - *Framework for AI Diffusion* (2025-01-15).

P1 depends on the off-topic buckets:

Test	Result
Federal bucket means (resource salience)	EPA 0.224, FERC/DOE grid 0.232, DOE/White House AI 0.079, NTIA/Commerce 0.062. Local and state mean: 0.206
P1 against only the EPA and FERC buckets	**−0.002, p = 0.92**
Wild-cluster restricted bootstrap (Rademacher, B = 4,999, by jurisdiction)	**p = 0.165** (conventional clustered p = 0.027)

- **Few federal clusters.** There are only 4, so even the bootstrap p is fragile.
- **What the contrast measures.** "Local and state talk resources, Washington doesn't" compares local data-center records with Medicare, crypto and stablecoin rules.

4. P2: real association, but off-topic records drive its significance

- **The keyword rule does not create P2.** P2 on rule-only scores: −2.08, p = 0.70.
- **P2 holds on live judges and across raters** (n = 169):
- JEV salience predicts DeepSeek's pressure index: +33.3, p = 0.019.
- DeepSeek salience predicts JEV's pressure index: +57.6, p = 0.0001.
- **Bootstrap:** the wild-cluster bootstrap p is 0.026.
- **By relevance:** off-topic records (n = 139) give +55.0, p < 0.0001. On-topic records (n = 225) give +11.9, p = 0.41, 95% CI [−16, +40]; the on-topic bivariate is +27.9, p = 0.036. The relevance interaction is not significant (p = 0.25).
- **Within source system** the association is not significant (p = 0.14, Boston's own check).
- **Restatement:** within on-topic records, P2 is attenuated and imprecise. Its full-sample significance rests largely on contrasting off-topic filler with data-center fights.
- **SG's own scoring.** SG scored 15 on-topic records with the paper's rubric. Its pressure index correlates r = 0.81 with the composite and 0.84 with JEV, so the index is credible where the text is on-topic.

5. Corpus and P3

Relevance. - About 225 of 364 records mention data centers or AI. - Of 10 random Legistar records, SG judged 5 on-topic. The other five: a SAIDI reliability report, an AI imagery pilot, AI usage reports, a power-purchase audit, an Oakland "AI activation zone". - AII-2023-013 is Richmond's **FY1984–85** electric utility budget, dated 2023.

"Primary records". - Only 18 of the 115 `State_County_PUC_Primary_Search` records come from government domains. - The rest are news sites (NPR, WTOP, GPB), law-firm blogs, and vendor or advocacy pages (flexgen, ashbornpartners, datacenterbans, servercountry).

Clusters and dates. - 22 of the 48 "jurisdictions" hold exactly 5 records: search-quota buckets. - Dates of web pages are page dates, not event dates. AII-2026-039 (2026-04-15) reports a December 2023 rezoning. - One event is counted twice: the Atlanta BeltLine ban, AII-2024-021 and -022.

P3 and volume growth. - "Annual record volume expands nearly fivefold" follows from newest-N caps: Legistar `$top=35` newest with ≤14 kept, and the newest 25 Federal Register hits. - P3 is null under the bootstrap (p = 0.107) and within Legistar (+3.66, p = 0.13). The paper already labels it marginal.

6. Pre-registration and history

The design card was written after contact with endpoint metadata. It is stamped "Pre-Data Locked 2026-11-05T02:58:00Z", but it already lists: - exact frozen-endpoint dates (goodyear 2021-01-27, …); - specific clerk typos (richmondva `9859-10-01`, denver `2030-09-30`, milwaukee `2028-02-16`).

The analysis pull is stamped 03:04:18Z. These facts are endpoint metadata, not outcomes, so the card may still predate the scored data. SG cannot check commit `08ef92e` from outside Boston.

The pre-registered P2 "tail" test passes on 3 records. It requires that, among records with pressure index ≥ 50, codes 2 + 3 outnumber code 5. In the composite: 2 vs 1 out of n = 9, and conditional bargaining is 0 of the 9. The abstract's "conditional bargaining dominates pure prohibition" rests on full-sample counts, not on this test.

Edition	Hypotheses	Data	N / G	Headline	Judges
Design card (02:58Z)	P1–P3; P2 adds the tail prong	Legistar (28 endpoints), Brave, Federal Register	≥240 / ≥35 target	n/a	3 Ollama Cloud + JEV
v1.0	same	same panel	364 / 48	P1 p .027, P2 p .007, P3 p .058	same
v1.1	same	panel said to be byte-identical; consistent with v1.0's numbers	364 / 48	same; fixes 3 R^2 cells and the P3 df; adds rank, median-regression and source checks	same

There is no hypothesis pivot between editions.

7. SG gates

Gate	Result	How it was handled
	FLAG	

Gate	Result	How it was handled
G3 variable definitions		Quoted by hand: `period_2025_2026 = 1 if event_year >= 2025`. `word_tone` is a lexicon count whose negative list includes "ratepayer" and "moratorium", so this control overlaps the constructs it adjusts for
G2/G5 construct and provenance	CLEAR	215/364 texts contain a marker. SG's reading (§3, §5) shows CLEAR proves little here
G4/G6 history	CLEAR	4 editions, no flags
G4b narrative	ERROR	The gate is built for ticker/date event studies. Replaced by a manual check: all 28 named-record values in §4.3 match the data
G1 timing and the ICC gate	NOT_RUN	Not an event study. ICCs were checked directly with SG `icc.py` (§1, §2)

8. Critic, self-correction and independence

Critic. A fresh-context critic (Claude Opus 5.5) attacked SG's draft. SG re-ran its key numbers and adopted these corrections: - compare like-for-like ICCs; - drop the "Zero Synthetic Fallback" misreading; - restate federal relevance as "few about data centers; about a third energy rules"; - restate on-topic P2 as "attenuated and imprecise", not "vanishes"; - soften the pre-registration finding; - add the federal bucket split, the energy-only P1 test and the off-topic P2 result.

Prior SG claims (EXT-AUDIT-023, Muse Spark 1.3, on v1.0):

Claim	Status
"All headline numbers replicate exactly"	Held
"ICCs replicate exactly"	Arithmetic held. Missed that the 4-judge ICCs for frames and salience are inflated by rule ballots
"GLM 97% fallback, JEV-only still supports P1/P2"	Held as arithmetic. P1 fails the bootstrap and the energy-bucket test
"0 anachronisms (placeholders cosmetic)"	Held as stated, but the 2023-dated FY1984 budget record went unnoticed
Off-topic federal sample; newest-N sampling; non-primary sources; posture averaging; the design card's data-derived facts	Not raised. Added here

Independence: - Before computing, SG saw the subject line of its earlier audit commit (verdict REPLICATED_WITH_CAVEATS). SG read the EXT-AUDIT-023 summary only after its numbers were fixed. - **Same model family.** The auditor and the critic are the same model, so this audit has no cross-family check.

9. What Boston should change

1. **Reliability.** Report it on live ballots only. Re-run or drop the rule-fallback ballots, and fix the ICC claims in the abstract.
2. **Posture.** Derive posture from the mode or median of live codes. Re-state the "1.1% direct denial" and "bargaining dominates" claims.
3. **Federal sample.** Rebuild it from relevance-ranked or phrase-matched documents across 2023–2026, then re-test P1 with wild-cluster inference.
4. **Corpus.** Filter for relevance before scoring. Label secondary sources as secondary. Date records by event, not by page. De-duplicate events.
5. **Small errors.** Correct the full-rank SE sentence and the §4.4 ranking.
6. **Design card.** State how it was drafted after the endpoint probe.