

PHANTOM Supplementary Appendix

Nischit Girish Kumar · NEXUS Research Initiative — Independent Research

Companion to PHANTOM: Benchmark Validity and Causal Limits in Coordinated Listening

Scope and provenance

This appendix presents the scientific chronology and descriptive stratification of existing outputs. Preparing it involved no final population construction, detector execution, null generation, tuning or new prediction. Sections A-H explain the methods, ordering and material verification qualifications; statements about release status within that chronology describe the stage at the time. Exact execution timestamps, original receipt details and the unabridged procedural record remain in the internal provenance archive. Publication history is recorded in Section G.

I. Source snapshot, event windows and evaluation chronology

Phase 4.12 external detector validation completed before the future September population-detector holdout protocol was fixed and sealed. Phases 5–11 then developed separate reconstruction, investigation and reporting components while the population detector remained fixed and the September histories remained unaccessed and unmaterialized. Phase 11 froze the PRE-FINAL manuscript. During pre-access verification, Phase 12A clarified an underspecified randomization mapping; the completed execution implementation was then locked before the final population was accessed. Phase 12 subsequently materialized the September population, scored the fixed detector and recorded its outputs. Verification followed the scientific run; the original receipt failure and subsequent checks are summarized in Section F. The September holdout was not used to tune the detector, and no scientific tuning followed observation of its outcomes. The original protocol applied to the population-detector evaluation, not the later graph or investigation components.

Both May external validation and September final use the official ListenBrainz snapshot 2723 dated 15 October 2026. A source snapshot is an export of prior events: its publication date does not require the events to have occurred after that date. Acquisition is a later retrieval operation; population materialization selects the experimental histories; scoring is later still.

Scientific boundary	May external validation	September pre-access holdout
Snapshot publication/date label	15 October 2026; snapshot 2723	Same snapshot 2723
Event window; exclusive end (UTC)	1 May–29 May 2022	1 September–29 September 2026
Protocol fixed	Source selection and 42-group design fixed before acquisition	After May external validation; amendment sealed before final history access
Population access	After source selection and sampling design were fixed	After the pre-final manuscript freeze, pre-access amendment and completed execution lock
Population fixed before outcomes	Compatibility checks and population freeze preceded scoring	Population freeze and binding preceded scoring
Scoring and verification	Fixed detector scored; results recorded before the September commitment	Fixed detector scored after population binding; outputs recorded before replay and post-run verification

The local compatibility, freeze, binding and verification records support this ordering; they are not third-party attestations. September event-column access began after the superseding execution lock. Earlier requests were limited to archive headers, not final listener histories (Section F3). Exact clock records and the original timing table remain in the internal provenance archive.

The September event histories already existed in the committed snapshot. Protocol, exclusions, seeds, metrics and gates were fixed before the final population was materialized or scored. The appropriate description is “prospectively committed, pre-access holdout evaluation,” not a claim that events were collected after model freeze. Both samples contain 42 groups × 96 listeners = 4,032 because the final reused the predeclared sampling design. September excludes all 16,783 May-window identities and uses a separate time window/population selection. The exclusion audit records zero remaining identity overlap. The observed-event counts are 4,004,802 for May and 3,714,102 for September.

I1. May acquisition scope

Phase 4.12 did not download the full ListenBrainz archive. HTTP byte-range access retained user_id, recording_msid and listened_at from nine Parquet members, 824–832, covering the May window. All rows in those selected members were retained in the local projection, including boundary rows outside the exact window; the strict 1 May inclusive–29 May exclusive filter was applied afterward. Thus the projection contains 19,024,403 partition rows, while the May window contains 16,006,621 events. The complete archive remained remote, and its publisher checksum was recorded without claiming full-archive verification. Archive-header, schema and timestamp-boundary probes located the required members; these probes were not complete listener-history acquisition. The retained boundary probes extend only through October 2022. September 2026 final listener histories were neither materialized nor inspected in Phase 4.12; the later final commitment bound the remote snapshot and selection rules, not downloaded September histories.

J. Frozen final score stratification

This is descriptive stratification of the already-predeclared 32-, 24- and 16-synchronized-slot levels. Each arm/level contains 126 worlds: 42 groups × 3 construction seeds. Shared groups and paired backgrounds remain dependent. The score threshold is $Z \geq 3$. Signed margin is $Z - 3$. Quantiles use the linear sample-quantile convention and are not uncertainty intervals.

Slots	Arm	Alerts/count	Median Z	Min Z	Max Z	Zero scales
32	treatment	126/126	248.212599	75.446609	323.034557	0
32	control	0/126	-0.284084	-2.505251	2.922793	0
24	treatment	126/126	185.966738	55.790021	242.956223	0
24	control	0/126	-0.096356	-2.280421	2.248303	0
16	treatment	126/126	123.977825	36.518857	164.785070	0
16	control	0/126	-0.094695	-2.860771	2.119828	0

Slots	Arm	Q1 Z	Q3 Z	Min signed margin	Max signed margin
32	treatment	248.212599	248.212599	72.446609	320.034557
32	control	-0.891291	0.521927	-5.505251	-0.077207
24	treatment	185.549196	186.954385	52.790021	239.956223
24	control	-0.634342	0.674491	-5.280421	-0.751697
16	treatment	123.142742	124.893206	33.518857	161.785070
16	control	-0.730698	0.603055	-5.860771	-0.880172

All 756 null scales are positive. The frozen $\pm 10^{12}$ zero-scale convention remains part of the method, but no final score is a sentinel. The ECDF figure shows full, uncapped score ranges in separate control and treatment panels;

their horizontal scales differ explicitly. The nearest control has score 2.9227932910652497 (clearance 0.07720670893475035); the weakest treatment has score 36.51885683451851 (margin 33.51885683451851). The very strong planted separation should be read alongside the declared support restriction. These distributions characterize this benchmark, not universal task difficulty.

K. Reconstruction outcome breakdown

The Phase 5/5.1 reconstruction benchmark uses observed histories from the Deezer listening-events dataset, the release distributed as `recsys24_deezer_dataset.tar.gz` (DOI 10.5281/zenodo.13890194), restricted to 1 May 2022 inclusive through 29 May exclusive. The retained source chain runs from the Deezer development sample and additional observed stress-cohort histories, through the permitted evaluator background and Phase 4.7/4.9 original-event records, to the Phase 5 construction. Its six development blocks are 0, 2, 3, 5, 6 and 7, containing 759 observed users and 50,523 original events. Blocks 8 and 9 come from the same historical Deezer source pool; Phase 5.1 opened them only after development passed and the reconstruction pipeline was frozen. They form an internal reconstruction holdout, not the external ListenBrainz population. The Phase 4.12 ListenBrainz population data and numerical result arrays were excluded from Phase 5/5.1 tuning; only its historical reports and public final-commitment metadata were consulted.

The accepted historical holdout has 84 treatment worlds, 84 controls, two original-only worlds and 168 original planted parents. The stored evaluation table has per-world means for candidate and comparator methods. Inspection matches existing candidate member/target sets to existing protected truth by exact set identity, without invoking inference or re-running the detector. Every matched candidate is an exact member and target match to one parent, and every missing parent has no overlap with an existing candidate. This makes the matching unambiguous and reproduces all four frozen world-level averages exactly.

Outcome	Count	Definition
Fully recovered member and target sets	154	Exact member AND target set identity; not exact period or cause
Partially recovered member/target sets	0	No partial overlap among frozen accepted objects
Missed original parents	14	No matched structural object; score zero
Unmatched original parents	14	Same 14 misses, not an additional category
Unmatched predicted candidates	0	Every predicted object matches a parent
Structural ambiguity abstentions	not separately recorded	Payload lacks a reason flag; do not impute causes
Causal attribution withheld	154	All recovered objects withhold fraud attribution

The 84 treatments comprise 76 with two recovered parents, two with one, and six with neither. Member precision, recall, F1 and target F1 each equal $154/168 = 11/12 = 0.9166666666666666$ because every original parent's corresponding value is either one or zero. Holdout block 8 recovers 78/84 parents; block 9 recovers 76/84. No original parent was omitted from the denominator.

Regime	Level	Recovered / parents	Partial	Missed	Member/target F1
broader targets	1	8/8	0	0	1.000
broader targets	2	8/8	0	0	1.000
broader targets	3	8/8	0	0	1.000
distributed	1	8/8	0	0	1.000
distributed	2	8/8	0	0	1.000

Regime	Level	Recovered / parents	Partial	Missed	Member/target F1
distributed	3	8/8	0	0	1.000
fragmented	1	8/8	0	0	1.000
fragmented	2	8/8	0	0	1.000
fragmented	3	8/8	0	0	1.000
gradual evolution	1	8/8	0	0	1.000
gradual evolution	2	8/8	0	0	1.000
gradual evolution	3	1/8	0	7	0.125
low signal	1	8/8	0	0	1.000
low signal	2	8/8	0	0	1.000
low signal	3	8/8	0	0	1.000
reduced participation	1	8/8	0	0	1.000
reduced participation	2	8/8	0	0	1.000
reduced participation	3	1/8	0	7	0.125
weak temporal	1	8/8	0	0	1.000
weak temporal	2	8/8	0	0	1.000
weak temporal	3	8/8	0	0	1.000

Designated-fragment recovery evaluates whether both protected halves of an original fragmented parent are represented by at least half their members in a single matched candidate. All 24/24 parents in the designated fragmented regime satisfy it; this specialized denominator is not the whole holdout.

K1. Why 30-second jitter also rounds to 0.9167

The stored jitter panel has 252 original parents from six reused development blocks (0, 2, 3, 5, 6 and 7). Exactly 231 have member F1 and target F1 one, and 21 have both zero: $231/252 = 11/12$. The historical holdout uses blocks 8/9 and 154/168. Thus the reduced fractions coincide, but the parent populations and perturbation designs differ; these are not identical recovery outcomes. The revision retains original-parent truth rather than switching to surviving-truth denominators.

L. Raw/global/local comparison

Method	Population	AUC	Control alerts	Pooled rate	Worst block
Raw relational fitted-score threshold	Historical validation; four source blocks	1.000	22/180	12.22%	not recorded
Global pooled95	Eight development blocks; nested fit/calibration	1.000	23/180	12.78%	41.67%
Global worst_block95	Eight development blocks; nested fit/calibration	1.000	22/180	12.22%	38.89%
Global calibration_max	Eight development blocks; nested fit/calibration	1.000	12/180	6.67%	29.63%
Local empirical_tail	Same eight development blocks; fixed normalization	1.000	6/180	3.33%	11.11%
Local robust_excess	Same eight development blocks; fixed normalization	1.000	0/180	0.00%	0.00%
Frozen local robust excess	May 2022 external; 42 groups	1.000	0/378	0.00%	0.00%
Frozen local robust excess	September 2026 pre-access holdout; 42 groups	1.000	0/378	0.00%	0.00%

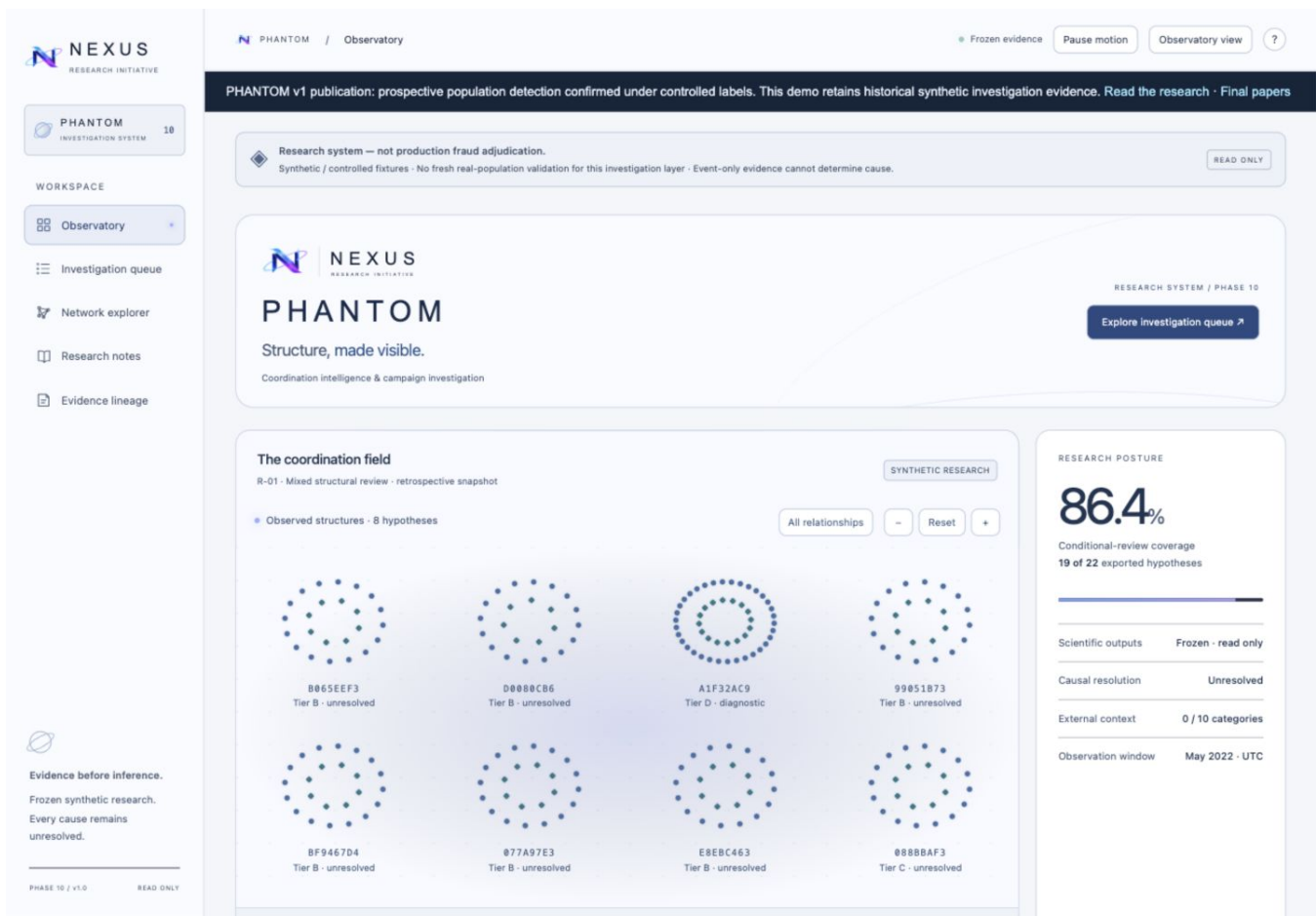
The global calibration procedures use nested held-out source blocks with separate fitting and calibration subsets. Local normalization uses the same eight development backgrounds with fixed formulas and thresholds and no analogous model fitting. The raw validation and later May/September populations differ. Rows therefore do not estimate a common head-to-head method effect. Full and deduplicated standalone rows agree. The main point is that perfect ranking coexisted with failing operating thresholds; a 6.67% pooled control alert rate concealed a 29.63% worst-block rate. CopyCatch, CatchSync, FRAUDAR and other published methods were not run as matched empirical baselines, and none was run now on September.

M. Plain-language metric and terminology guide

Term	Meaning
World	One detector-visible input containing background and either a treatment or matched control assignment.
Source group / block	Background listeners whose histories recur across several constructed worlds; the grouping unit for reporting and historical resampling.
Parent	An original planted campaign in evaluator truth; a reconstruction target, not a real-world organizer.
Support / out of support	Whether an input satisfies the method's declared structural requirements; an unsupported input is not a negative classification.
Semantic presence / atom	One unique user–item–timestamp event for reconstruction, or user–item–minute presence for population coactivity; repeated identical records do not add presence.
Candidate admission	A fixed catalog component becomes a reported structural candidate only after passing the frozen conditional comparison.
Control alert rate / burden	Fraction of control worlds containing an alert; not a real-world fraud false-positive rate.
Background absorption	Fraction of background actors included in predicted structural objects.
Stress absorption	The analogous inclusion fraction within a specified stress cohort; cohort and denominator must be named.
Post-veto member F1	Historical post-veto member F1, with parents removed by veto decisions scored zero; it does not measure analyst usefulness.
Designated-fragment recovery	Both protected halves of a fragmented parent substantially represented in a single matched object; evaluated on the fragmented regime.
Surviving-truth	Secondary evaluation against the part of original truth remaining visible after perturbation; primary metrics retain original truth.
Causal twins	Different hidden accounts constructed to supply identical permitted inputs and the same randomness law; not an estimate of their prevalence.
Representation-relative equivalence	Equal priority for candidates indistinguishable under the retained evidence signature; not proof of common cause or exhaustive explanations.
RQ2	Removed undefined shorthand; the main text directly asks whether the fixed population detector transfers under the controlled mechanism.

N. Figure and interface archive

The current main papers show the frozen score ECDF (asset 11). The external count graphic (asset 05) and September aggregate-count/42-block-rate graphic (asset 10) are archived, superseded figures and are absent from the revised main papers. The Research Manuscript also omits the separate causal-twin diagram; its causal argument retains the necessary assumptions in text. No scientific figure was redrawn from simulated values. The historical synthetic dashboard screenshot (asset 09) remains in this supplement and on the original public website. Section C3 identifies the current figure assets and their provenance.



Historical synthetic dashboard. Its internal Phase 10 version is preserved and is not the manuscript version. Engineering check counts in the historical chronology document software fidelity, not detection accuracy, independent evidence or analyst utility.

A. Preserved historical research chronology

Deployment and publication statements below describe the historical status at each phase. The project website and public repository became live later, and publication/indexing occurred subsequently. These later events do not change the preserved phase decisions.

Specification

Objective: Define an event-level research system. Hypothesis: Relational evidence may improve over individual anomaly heuristics.

Method: Architecture, contracts and staged gates. Result: Aspirational specification, not empirical validation.

Gate: Require evidence before stronger claims. Decision and resulting change: Original scope superseded where later falsification narrows it.

1 / data audit

Objective: Establish the source representation. Hypothesis: The release can support reproducible event research.

Method: Disk-backed parquet audit; duplicate, schema and session checks. Result: 704,728,520 events; all source sessions ten records.

Gate: Data use must respect representation limits. Decision and resulting change: Preserve source scale while rejecting unvalidated natural-session assumptions.

2 / behavior

Objective: Characterize bounded listening histories. Hypothesis: Behavioral features can describe variation.

Method: Release-relative user sampling and descriptive analysis. Result: 671,410 sampled rows from 3,252 users.

Gate: Description is not a verified legitimacy label. Decision and resulting change: Use observed histories as a foundation, not labelled negatives.

3 / simulation

Objective: Create controlled coordination tasks. Hypothesis: Synthetic campaigns permit measured recovery.

Method: Six families, three levels, four popularity regimes. Result: Origin and catalog differences were possible.

Gate: Audit generation shortcuts. Decision and resulting change: Introduce neutral and origin probes.

4 / baseline

Objective: Evaluate behavioral and relational candidates. Hypothesis: High performance may represent coordination.

Method: Logistic baselines and graph comparisons. Result: Apparently useful detection coexisted with session and origin problems.

Gate: Benchmark interpretation rejected. Decision and resulting change: Do not promote initial scores into final headline results.

4.5 / remediation

Objective: Remove known session leakage. Hypothesis: Session-free behavior might remove the mismatch.

Method: Actor-weighted session-free experiments and permutation probes. Result: Neutral alerts 12.08% versus observed-background 4.79%.

Gate: NO-GO: residual benchmark shift. Decision and resulting change: Demand stronger source and catalog controls.

4.6 / realism

Objective: Test calendar/catalog origin effects. Hypothesis: More realistic generation could admit relational comparison.

Method: Augmented-origin gates for 72 candidates. Result: All 72 failed admissibility gates.

Gate: No accepted detection evaluation. Decision and resulting change: Move toward within-history controlled intervention.

4.7 / counterfactual

Objective: Pair intervention and sham on observed histories. Hypothesis: Shared source origin might remove shortcuts.

Method: Additive treatment/control construction. Result: Actor marginals still changed.

Gate: NO-GO under required invariance. Decision and resulting change: Preserve marginals rather than merely adding matched rows.

4.8 / structure preservation

Objective: Hold item/time marginals fixed. Hypothesis: Reassignment can isolate cross-actor structure.

Method: Within-actor assignment changes and duplicate audits. Result: 20/180 pairs changed semantic duplicate counts.

Gate: Processing gate failed. Decision and resulting change: Repair collisions before interpreting relational scores.

4.9 / relational calibration

Objective: Use collision-safe controlled pairs. Hypothesis: A strong relational ranking may transfer a threshold.

Method: Fixed schedules and a frozen raw threshold. Result: AUC 1.000 but 22/180 validation controls alerted.

Gate: Standalone burden gate failed. Decision and resulting change: Separate ranking from calibrated threshold behavior.

4.10 / calibration

Objective: Find transferable global calibration. Hypothesis: A global development threshold might satisfy all blocks.

Method: Three declared calibration procedures. Result: All failed per-block ceiling requirements.

Gate: NO-GO despite useful pooled results. Decision and resulting change: Normalize against each eligible world's conditional null.

4.11 / local surprise

Objective: Freeze self-normalized relational detection. Hypothesis: Local baseline and dispersion can improve transfer.

Method: 4095 permutations; robust median/MAD; threshold three. Result: Development separation with no observed control alerts.

Gate: Accepted as frozen candidate; external evidence still needed. Decision and resulting change: Carry identical method to an independent source population.

4.12 / external validation

Objective: Evaluate frozen detector in ListenBrainz. Hypothesis: Controlled relational signal transfers across populations.

Method: 42 groups, three construction seeds and three difficulties. Result: 378/378 treatment alerts; 0/378 controls; AUC 1.000.

Gate: Bounded external-population test passed. Decision and resulting change: Qualify transfer by synthetic intervention and narrow selector. The September population-detector holdout was then committed and sealed before Phase 5 began; its histories remained unmaterialized. The later reconstruction/investigation phases did not reopen that detector or holdout. Section I gives the scientific ordering; exact timestamps remain in the internal provenance archive.

5 / graph intelligence

Objective: Recover member/target objects. Hypothesis: Temporal graph evidence can reconstruct campaigns.

Method: Pair-level nulls, components and variants. Result: Duplicate sensitivity and unacceptable control burden.

Gate: Graph-null remediation required. Decision and resulting change: Use semantic presence and stronger conditional candidate pooling.

5.1 / graph-null remediation

Objective: Fix representation and pooling. Hypothesis: Fixed catalog pools can aggregate weak relational evidence.

Method: Burst-conditioned null, 60/300 consensus, fragment links. Result: Historical internal member/target F1 154/168 = 11/12; 0/84 controls.

Gate: Accepted frozen core within support. Decision and resulting change: Retain source identity and test hard boundaries.

6 / hybrid

Objective: Improve actionability with behavior. Hypothesis: Behavioral veto may reduce unwanted alerts.

Method: Hybrid fusion around accepted core. Result: Post-veto member F1 fell to 0.5357.

Gate: Rejected degradation. Decision and resulting change: Do not replace core with a weaker veto system.

6.1 / non-veto

Objective: Add information without suppressing core. Hypothesis: Non-veto context can improve ranking.

Method: Within-block ranking, stability and cost tests. Result: Zero added gain over 210 comparisons.

Gate: Rejected augmentation; retain core only. Decision and resulting change: Prioritize failure characterization over unsupported complexity.

7 / robustness

Objective: Map perturbation and hard-negative boundaries. Hypothesis: Moderate missingness and negative contexts can be tolerated.

Method: Frozen core on reused development; separate perturbations. Result: Several moderate settings pass; shocks and shifted overlap fail.

Gate: Overall operating region failed. Decision and resulting change: Seek common-cause/overlap explanation and keep failures visible.

7.1 / common cause

Objective: Repair shocks and overlap. Hypothesis: Conditioning or parent association can resolve the failures.

Method: Alternative conditioned nulls; combined gates; causal twins. Result: Partial fixes conflict; identical-input causes remain indistinguishable.

Gate: Combined requirements not met. Decision and resulting change: Recognize missing causal information instead of tuning it away.

7.2 / reframing

Objective: Define a scientifically defensible task. Hypothesis: Structure can aid investigation without intent labels.

Method: Operating, terminology and provenance contracts. Result: Event-only attribution explicitly withheld.

Gate: Govern later work by the narrower claim. Decision and resulting change: Separate support, alternatives, evidence and external provenance.

8 / red team

Objective: Challenge investigative output. Hypothesis: Structural candidates can support stable review.

Method: Synthetic bridge, overlap, ranking and packet attacks. Result: Parent recovery, identifier ranking and packet completeness fail.

Gate: Investigation remediation required. Decision and resulting change: Specify complete evidence and invariant review semantics.

9 / investigation remediation

Objective: Make outputs auditable. Hypothesis: Complete packets and support checks improve interpretability.

Method: 17-section evidence contract and ranking tests. Result: Packets improve; shared-response ranking and support disagreements remain.

Gate: No unqualified supported claim. Decision and resulting change: Separate review tiers from causal conclusions.

9.1 / triage

Objective: Route ambiguous evidence. Hypothesis: Tiered routing can preserve valid review cases.

Method: Queue and hard-response gates. Result: 18/19 gates pass; shared-response handling fails.

Gate: Remediation remains required. Decision and resulting change: Replace unjustified scalar preference with complete equivalence sets.

9.2 / equivalence

Objective: Preserve unresolved alternatives equally. Hypothesis: Representation-relative sets avoid false preferences.

Method: Equal priority, invariant display and complete filtering. Result: 480 candidates; 327/477 original parents reviewable; OOS 40/44.

Gate: Accepted implementation scope; unresolved support disagreements retained. Decision and resulting change: Freeze cautious evidence packets; no fresh population claim.

10 / dashboard

Objective: Expose frozen evidence faithfully. Hypothesis: Read-only interface can preserve the research contract.

Method: Five synthetic runs, 22 hypotheses, academic exports. Result: 3943 engineering checks recorded as passing.

Gate: Engineering freeze, not scientific or analyst validation. Decision and resulting change: Use the frozen interface as an inspectable research demonstration.

11 / Pre-final freeze

The scientific system and PRE-FINAL manuscript were frozen before the final evaluation. The original versions and their provenance snapshots remain byte-identical. Comparative baselines, broader support coverage and analyst studies remained unresolved; a future final result was not permission to retune the method.

12 / Pre-access holdout and verification lineage

The original final commitment was invalidated before scoring because final412 lacked executable semantics. Phase 12A added one deterministic mapping and resealed before final histories were accessed. A first execution lock was superseded, after toy/non-final equivalence checks and before final event access, to bind subsequently completed execution adapters. The canonical scientific run produced perfect controlled separation. The original verifier completed replay but failed while serializing a NumPy boolean in its JSON receipt. The invalid result and failed attempt remain permanent historical records.

Phase 12B independently checked immutable original/replay artifacts. It reproduced every provisional final metric exactly and confirmed all 756 worlds and 18 applicable scientific gates. It did not rerun scientific producers. Section F summarizes the verification checks; the internal provenance archive retains the two separate tooling incidents in full.

13 / Final synthesis and bounded release preparation

Final publications incorporate the confirmed result and the complete procedural qualification. Figures use frozen aggregates; arithmetic and byte-identity checks create no new scientific evaluation. A sanitized documentation/synthetic-demo candidate is separate from the private review archive. Historical status at Phase 13 completion: no publication, deployment, submission or indexing had occurred within that phase. The website and public repository went live later. PHANTOM v1 science is closed.

B. Methodological appendix

B1. Read the two detectors separately

The Phase 4.11 population detector and Phase 5.1 reconstruction core are different methods. The former selects one exact recurring actor set, computes population minute-bin coactivity and uses 4095 null draws. The latter selects singleton-user burst support, constructs a fixed catalog graph and pools component evidence over 511 draws. External ListenBrainz success belongs to the former. Historical internal graph reconstruction belongs to the latter. Their evidence must not be merged into a claim of externally validated end-to-end investigation.

B2. Population detector pseudocode

1. Read only the detector-visible canonical event array. Group exact timestamps by recurring actor set. Require at least eight distinct users occurring once, exactly 32 recurring timestamps in distinct minute bins, span at least 27 days, and exactly one qualifying group. Otherwise return unsupported.
2. Reduce background to unique (user, item, minute) presences for sufficient counts. Keep background, actor identities and selected timestamps fixed. Independently permute each selected actor's sorted item multiset over its slots using 4095 draws and seed 41101.
3. Compute the same ordered cross-actor coactivity numerator and opportunity denominator for observation and every draw. Center by null median and scale by 1.4826 times MAD. Zero scale maps equal/above/below median to 0/positive sentinel/negative sentinel. Alert at robust excess at least three.
4. Return score and support information. This does not return a Gaussian probability, a fraud probability, a campaign organizer or an estimate of natural-world false positives.

B3. Reconstruction pseudocode

1. Restrict the input to START=1651363200 and END=1653782400, with END exclusive. Canonicalize unique (user,item,timestamp) atoms; retain multiplicities as description. Ambiguous actor fingerprints produce abstention.
2. Form bursts by consecutive gaps no greater than 600 seconds. Retain bursts having at least eight distinct users and eight singleton-user atoms. Keep actors present in at least eight qualifying bursts, apply that filter once, then keep groups of at least two selected atoms.
3. Build an untimed selected-catalog graph using shared-item count at least three and Jaccard at least 0.10. Connected components of size at least three are a null-invariant candidate family.
4. Independently permute selected items within each actor with 511 draws and seed 55101. Sum same-item same-bin evidence over shared slots, use the minimum 60/300 count per pair, and average over every unordered pair within each candidate, including zeros.

5. Normalize observed and null component scores with their common median/MAD. Compare observed scores to the per-draw maximum across fixed components with the add-one denominator 512. Require robust excess at least three and comparison value at most 0.01.

6. Link admitted components with target Jaccard at least 0.25 and overlap at least 0.5 relative to the shorter period. Preserve constituent scores. Targets shared in two member catalogs need not have direct temporal support. Return evidence and withheld causal attribution.

B4. Metrics and uncertainty

For a matched predicted member set P and original parent T , precision is intersection size divided by $|P|$, recall divides by $|T|$, and F1 is twice the intersection divided by $|P|+|T|$. The historical evaluator uses one-to-one Hungarian matching by member F1; unmatched parents receive zero. Mean F1 is the average of per-parent scores, not a recomputation from two unrelated means. Target and fragment metrics have their own denominators. The fragmented-regime recovery claim does not describe every campaign type.

World alert burden counts any structural alert in a control world. It is not a real-world false-positive rate because observed users lack verified legitimacy labels and controls encode a research construction. AUC separates score ranking from the frozen threshold. Primary robustness scores preserve original truth under dropout; surviving-truth analyses answer a different question.

Whole-source resampling respects one layer of clustering but does not establish source independence or mechanism coverage. Perfect observed group outcomes produce degenerate percentile intervals. No population zero-error statement is warranted. The illustrative 42-independent-group zero-event upper bound in the report is explicitly hypothetical, not a replacement confidence interval.

B5. Causal-twin assumptions

The equality-of-output argument requires the same detector-visible input and, for a randomized method, the same randomness law independent of the hidden cause. It does not require the two causal stories to be equally likely. It does not establish posterior odds, or deny that external provenance or well-justified assumptions could identify causes. The six constructed causal-twin fixtures are evaluator cases, not observed real-world causal examples; they demonstrate a limitation of this information setting. Any future provenance extension changes the task and needs new validation.

C. Reproducibility appendix

C1. Historical runtime and source identity

The research relies on Python, NumPy, pandas/DuckDB/parquet workflows, scikit-learn in early behavioral evaluation, SciPy matching and NetworkX in bounded reconstruction, plus browser JavaScript/CSS/HTML for the frozen dashboard. Exact historical versions are recorded in the phase-specific environment receipts; the linked source map includes their identities. They are not replaced with the document-building runtime.

The accepted pooling code SHA-256 is 502267217e01ed904ec1ed7e12088d6b6a062d555819f3fd4cc8c5929a764979. It imports the frozen detector module. A matching file digest does not prove methodological correctness, and running newer dependencies is not automatically a faithful scientific reproduction.

C2. Seed discipline

Population local null: 4095 draws, seed 41101. ListenBrainz sampling: 41201. External construction: 41211, 41212 and 41213. Historical group bootstrap: 2000 replicates, seed 41102. Graph conditional null: 511 draws, seed 55101. Each number belongs to a specific layer; do not substitute them or use them to claim 4095 independent source populations.

C3. Permitted reproduction

This package supports rebuilding figures from frozen aggregate tables, compiling manuscripts, checking citations/claim references, and running a wholly synthetic toy illustration. It does not grant access to licensed raw data or protected evaluator truth. Exact full scientific reruns require their historical environment and separately authorized data. The public package excludes final raw records and protected evaluator truth. Phase 13 inspects frozen aggregate results and hashes preserved inputs, without scientific reruns.

Asset identifiers below refer to preserved source filenames, not the papers' sequential figure numbers. The Technical Report uses assets 02 (matched construction), 03 (local null), 04 (reconstruction), 06 (robustness), 07 (causal twins), 08 (investigation) and 11 (final score distributions). The Research Manuscript uses assets 02, 06 and 11. Assets 02, 03, 04, 07 and 08 are conceptual vector diagrams of verified operations. Asset 06 transcribes the frozen Phase 7 moderate-gate table. Asset 11 plots empirical cumulative distributions of the existing frozen September scores by predeclared difficulty, with separate control/treatment scales; it creates no predictions or inferential estimates. Asset 09 is the archived synthetic-demo Observatory capture reproduced in Section N. Assets 01 (architecture), 05 (external counts) and 10 (September aggregate counts and 42 block rates) are retained as archived source assets and do not appear in the revised main papers. No generative imagery serves as scientific evidence.

C4. Build and integrity

The Technical Report and Research Manuscript have editable LaTeX sources and share a verified BibTeX file. The figure script requires NumPy and Matplotlib; the supplement/brief renderer uses ReportLab and pypdf for inspection. The build receipt records the actual document environment and compiler/class versions. The Phase 13 manifest hashes final payload files, excluding itself and temporary previews. A separate before/after inventory checks preserved Phase 11 and Phase 12 lineage bytes. Historical inputs remain unchanged.

D. Claim and citation navigation

Claims C01-C05 describe the data audit; C06-C13 the controlled external validation; C14-C21 internal graph evidence; C22 causal twins; C23-C27 rejected augmentations and robustness; C28-C30 the investigation layer; C31-C37 the dashboard; C38-C40 earlier benchmark failures. Additional entries transcribe supporting report rows and implementation parameters. Every entry supplies exact wording, source locator, denominator, population, status and scope restrictions.

Citation verification covers 18 primary-source identities and their limited supporting statements, including BAG (AAAI 2026) and Graph Evidential Learning (KDD 2025). RecSys 2026 is used only as an eight-page format reference; that conference has already occurred. SAGE is in WSDM Companion 2026, not the main WSDM research track. The temporal/multiplex work is labelled a preprint. No PHANTOM DOI, peer-review status, institutional sponsor or publication venue is invented.

E. Public-release boundary

The internal evidence directory contains local source pointers and source excerpts. It is not the public repository. Release only the explicit allowlisted subset after rights, privacy and author sign-off. Do not upload the workspace wholesale. Historical status at Phase 13 completion: the offline portfolio remained noindex, had no live canonical domain and sent no telemetry; deployment was deferred. The project website and public repository subsequently went live, with publication/indexing occurring later.

F. Pre-access final: definitions, custody and interpretation

F1. Population and treatment/control denominator

The committed source is ListenBrainz snapshot 2723 (October 15, 2026). The final observation window is [2026-09-01 00:00:00, 2026-09-29 00:00:00) UTC. These midnight timestamps define the inclusive start and exclusive end of the data filter, not project execution times. Selection yields 4,032 listeners in 42 groups of 96 with 3,714,102 original observed events. Sampling seed is 41290; construction seeds are 41291, 41292 and 41293. Three levels per seed give 9 pairs per group, 378 pairs and 756 worlds. The May external test instead uses 4,004,802 original events; these counts must not be interchanged.

The null remains 4,095 draws with seed 41101; alert threshold remains robust excess at least 3.0. No post-final detector or null generation was performed in Phase 12B or Phase 13. “Single scientific execution” means one canonical prospective evaluation. The original verification process deliberately reconstructed worlds and replayed nulls as its committed determinism check; those computations are not concealed or counted as independent scientific trials.

F2. AUC and intervals

Pooled AUC is the fraction of all treatment/control score comparisons with treatment greater than control, plus half the tied fraction. Mean block AUC averages the same statistic over 42 groups with equal group weights. Each group has 9 worlds per arm, so its AUC uses 81 comparisons; the pooled AUC uses 142,884 comparisons. These are arithmetic comparison counts, not independent observations.

The fixed threshold yields 378/378 treatment alerts and 0/378 control alerts. Both AUCs are 1.000. Worst-block control burden is 0.0%. The committed 2,000 whole-block resamples, seed 41102, give the conditional AUC and detection intervals [1.000, 1.000], control-mean interval [0.000, 0.000], and equal-block upper 95% conditional bootstrap summary 0.0% (degenerate; not a population bound). Every observed block is perfect, so these degenerate empirical intervals are expected. They do not bound unobserved mechanisms, establish independence across groups or imply a zero natural-population alert probability. No new interval is substituted for the committed procedure.

F3. Ordered procedural record

1. The original v1 commitment fixed the final protocol, but final412 was not executable from the inherited constructor. The original disposition is permanently invalid before scoring.
2. During pre-access verification, Phase 12A defined explicit ordered hash/RNG inputs for enrollment, catalog and timestamps. The amended protocol was sealed before final history access. Other domains, scientific thresholds and null draws stayed unchanged.
3. The first execution lock preceded completed adapters. Header indexing began, but no final event columns were accessed. After toy/non-final equivalence checks, a superseding lock bound the completed execution files before final event access. The later population manifest refers to that lock; the lock was not rewritten to include outcomes.
4. The selected population was fixed before scoring. The original detector outputs were recorded after the scientific run, followed by the completed independent replay outputs. The corresponding manifest and freeze records remain in the internal provenance archive. Section F5 lists key lock, output and verification identifiers.
5. After replay, the original verifier failed while serializing its receipt. Its partial receipt, traceback and invalid-result record remain intact. Nine true checks were serialized; later unsaved values are not asserted to have been successfully recorded.
6. Post-final forensic verification used immutable original/replay artifacts and independent metric arithmetic. It reproduced every provisional metric exactly, including 42 block records, 126 block/level cells and three level summaries, and confirmed the original 18 scientific gates. The additional reproduction requirement is satisfied by post-final verification; the original status remains NOT EVALUATED in its historical ledger.

F4. Verification qualification

The original verifier failed because a NumPy boolean could not be serialized into its JSON receipt. The locked verifier was not patched. A separate writer converts only NumPy booleans to native booleans and retains errors for other unsupported types. The internal provenance archive retains the exact failing field, software version, traceback and original receipts.

Post-run verification encountered two tooling issues, documented in full in the internal provenance archive. Resolving them left seeded metric arithmetic and scientific equality criteria unchanged. Receipt completion checked 378 truth-file hashes and 1,512 exact cache comparisons across original/replay conditions, reusing completed metric computations. The discrepancy receipts remain intact.

All 756 original/replay world identifiers and expected artifacts were confirmed. The separate verifier generated no new assignments, predictions or null draws. Protected truth arrays were hashed, not deserialized. Independent means separate calculation/process, not an external institution. Hashes establish bytes, not the absence of all possible unrecorded activity.

F5. Key immutable identifiers

Governing execution lock: fda63407156d072ed2cfdcfcdcb46940134a021d953713a883a614f9399dc5c5.

Original detector-output freeze: 4244dfccc3905ef1a71f5850d90eee0c234a72e92a1cdfe3c7a01af545ff8703.

Completed replay freeze: 9bb1f491937798627c35ffc85a38375f66e4350e4b17552bf4bcf886994e5f70.

Original invalid-result record: 585f91baf7ecfa6d2ec4a8b916b8a95e4933beb1f74a67e957304043ac8418d2.

Corrected forensic receipt: 55de97d8c477b4f81d4744f037d297e8333d77b4347c7c2943391cd15f06611d.

The publication source inventory maps these identities to retained local files. Only sanitized hash identifiers and aggregate descriptions belong in the public package; private paths, raw histories, exclusions identifying real users and protected truth do not.

F6. What remains unestablished

Final labels describe a controlled intervention. Real users are not classified as fraudulent. The final does not establish real fraud accuracy, production false-positive rates, manipulative intent, causal attribution, bot identity, unrestricted coverage, analyst effectiveness, enforcement suitability or production readiness. Causal twins remain counterexamples because successful structural separation cannot add missing causal provenance.

G. Version and reproduction boundary

PHANTOM v1 is the scientific system frozen before the final and evaluated in Phase 12. The original public v1.0 release (FINAL v1) remains preserved. Technical Report v1.1 records the revised archival account, and Research Manuscript v1.1 presents the shorter account of the same frozen scientific system. The original website and public repository were subsequently published; the revised manuscripts remain unpublished. The dashboard is the unchanged Phase 10 evidence system. A scientific method change requires PHANTOM v2 and fresh, separately authorized evaluation. The September final population is consumed forever for unseen-evidence claims.

Publicly reproducible: equations, figures from aggregate tables, manuscripts and the synthetic teaching pipeline. Authorized-access reproduction: historical population selection, full experiments and raw evaluation, requiring source permissions and exact historical dependencies. Not released: protected final truth, raw listening histories and private orchestration logs. The public teaching script is not the frozen detector. No general open-source license is selected on the author's behalf.

H. Construction, notation and implementation audit

H1. Controlled-pair construction pseudocode

1. Select a source group with the committed eligibility and exclusion rules. Retain the original event window as the common background. The final selection seed is 41290; this algorithm description does not execute selection.
2. For each committed construction seed and level, use the sealed final412 domain mapping to determine enrollment, catalog and timestamp assignments. Construct the declared recurring schedule with 32 slots and eight items repeated four times. Levels specify 32, 24 or 16 synchronized slots.
3. Produce aligned treatment assignments and the collision-safe matched-control reassignment. Preserve actor identity, row counts, per-actor item and timestamp multisets, and the declared duplicate invariants. Reject a failed invariant; a favorable detector score cannot repair it.
4. Keep arm labels and protected parent information outside the detector-visible array. Score the frozen detector on canonical visible inputs only. The labels describe the construction, not real user intent. Matched arms and seed/level replicas retain shared-source dependence.

This is descriptive pseudocode for the frozen construction, not replacement code. Exact ordering, hash domains and random-number calls are committed in the preserved executable mapping.

H2. Notation checklist

E is the visible event collection; X is the complete permitted input in the causal argument. u and v index actors, i items, t timestamps, b minute bins, and r randomization draws. n_{ib} counts unique actor presences for one item/bin; N_b counts actor-item presences in the bin; d_{ub} is the number of distinct items for one actor/bin. The denominator removes same-actor ordered opportunities.

C is the population coactivity statistic. m is its local null median; the scale is 1.4826 times the median absolute deviation. Z is robust excess, not a normal z-test probability. In reconstruction, G is a fixed candidate component, w_{uv}^h counts shared-slot same-item coincidences at width h, and $T(G)$ averages the 60/300 minimum across all unordered actor pairs, including zeros. M^r is the maximum normalized candidate score for draw r.

Every threshold and exception in the main methods is descriptive of frozen code. The zero-scale ordering sentinels are not measured effect sizes. The maximum-family comparison is not asserted to prove organic-listening familywise error control. A merged object has constituent evidence, not a newly calibrated merged p-value.

H3. Computational scope

The source audit uses disk-backed scans; its 704.7 million-event scope is not a detector throughput result. The population detector scores fixed-background sufficient counts without copying every complete null world. Graph reconstruction stores draw-by-actor-by-slot assignments plus draw-by-component statistics. Catalog comparisons can still be quadratic in selected actors. No production latency, memory budget or unrestricted stream coverage was measured.

H4. Failure-to-inference map

Origin/neutral/catalog/session failures: generated source or representation features could stand in for coordination; reject the benchmark before accepting detection scores. Duplicate mismatch: preprocessing violated the intended controlled comparison; repair collisions before interpreting performance. Raw/global threshold failure: ranking and alert burden are different claims; local normalization is evaluated separately. Bad graph null: catalog connectivity and incorrect reference conditioning can alert on controls; retain fixed candidates and explicit temporal nulls.

Hybrid/non-veto failures: more features need measured benefit and must not erase useful evidence.

Shock/shifted-overlap failures: natural common stimuli and overlapping structure defeat a broad operating claim.

Causal/ranking ambiguity: absent information cannot be repaired by a stronger scalar score; preserve alternatives and equal-priority sets.

H5. Full final gate definitions

The following criteria are transcribed from the preserved gate ledger, not chosen after results. Gates 1-18 are scientific requirements: their original and post-final ledger outcomes are PASS. Gate 19 is the separate reproduction requirement: its original outcome remains NOT EVALUATED and its post-final outcome is PASS. This does not change the original verifier status, FAILED DURING RECEIPT SERIALIZATION.

1. INTEGRITY: v2 seal/code/config/environment match; original provenance and Phase 11 unchanged.
2. SOURCE: Committed snapshot and final window; ≥ 10 events, ≥ 2 tracks; 42 groups x 96 listeners; exact clones kept together; no incompatible substitution.
3. EXCLUSION: Exclude all 16783 committed May 2022 users and known development clone conflicts before scoring.
4. A: Exact paired behavior; both frozen models mean oriented $AUC \leq 0.55$ and $worst \leq 0.60$; $KS \leq 0.10$ and $|SMD| \leq 0.20$.
5. B: Exact original rows/schema/marginals/duplicates/layout; formatting positive-control $AUC \geq 0.95$.
6. C0: Each level mean injected excess ≥ 0.10 ; positive unit fraction ≥ 0.90 ; every block/seed mean positive.
7. D: Strict ladder fraction ≥ 0.80 and level mean excess strictly decreasing.
8. E: Disjoint source groups; all paired units/levels kept together; exact ladder marginals.
9. each_block_controls: Each block control alert rate ≤ 0.07 .
10. block_mean_upper95: Equal-block bootstrap mean upper95 ≤ 0.07 ; 2000 draws, seed41102.
11. each_block_level_detection: Every block/level treatment detection ≥ 0.80 .
12. mean_auc: Mean block $AUC \geq 0.70$.
13. each_block_level_auc: Every block/level $AUC \geq 0.60$.
14. behavior: Frozen behavior gates pass in both duplicate views.
15. block_dependence_reduced: Block eta-squared reduced $\geq 25\%$ from fixed Phase4.10 reference.
16. composition_dependence_reduced: Maximum absolute Pearson reduced $\geq 25\%$ from fixed Phase4.10 reference.
17. science_and_views: All frozen mechanics and full/deduplicated views agree.
18. detector_ladder: Strict detector-feature ladder fraction ≥ 0.80 and decreasing level means.
19. REPRODUCTION: Every world hash matches; independent fresh-process null replay and independent metric verification match.

The scientific verification basis is frozen evidence and unchanged predeclared arithmetic. Reproduction relies on completed immutable reconstruction/replay plus post-final comparisons. The supporting records are PHASE12_CORRECTED_VERIFICATION_RECEIPT.json and evidence/scientific_hash_comparison.json. Their identities appear in the source inventory. The exact machine-readable 19-row ledger, including original and post-final outcomes, is included as tables/final_gate_definitions.csv. Neither the original failed receipt nor its historical outcome is overwritten.

Authorship and AI assistance

Nischit Girish Kumar is the sole human author. NEXUS Research Initiative denotes independent research, not institutional sponsorship.

AI-assisted tools were used extensively for implementation, debugging, analysis scripting, and verification, with minor assistance in manuscript editing. Nischit Girish Kumar originated and directed the research, designed the system architecture and experimental methodology, set the evaluation criteria, interpreted the results, and made the final scientific decisions. The manuscript was written and revised by Nischit Girish Kumar, who assumes responsibility for the work and its conclusions.

Source-based acquisition, schema, preprocessing and eligibility instructions are provided in the repository document docs/SOURCE_REPRODUCTION.md. Researchers must obtain datasets from their official sources; this package contains no raw datasets. Exact experimental reproduction also requires separately authorized scientific code and protected evaluator resources.