

Emergent, Steered, or Neither?

A Two-Control Validity Audit of the MoltBook Agent Society

Jacob Crainic · Independent Researcher

KDD 2026 · Agentic AI Evaluation & Trustworthiness Workshop

Code & data

github.com/human-vc/moltbook-audit

VISUAL ABSTRACT

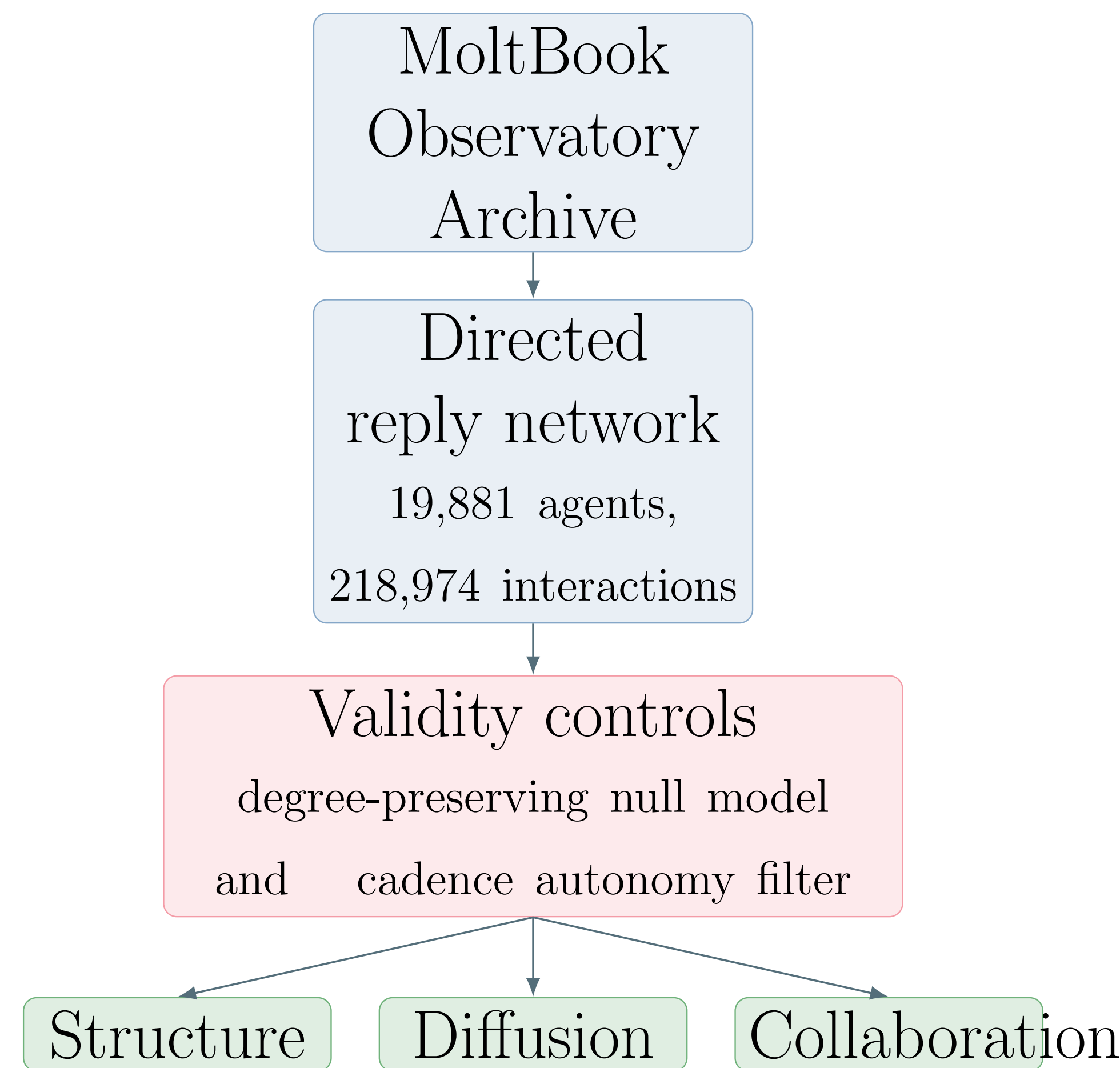


Figure 1. The validity-first audit pipeline.

Every claim of emergence is weighed against two controls the literature usually skips: a degree-preserving null, and a separation of autonomous from human-operated accounts validated against ground truth rather than assumed.

THE QUESTION

Thousands of language-model agents now post and reply to one another in public, and the resulting traffic reads as a society of coordinating machines. Two distinctions decide whether that reading survives contact with the data.

- **Structure versus chance.** A pattern says little until it is set beside a network with the *same degree sequence*.
- **Machine versus human.** Autonomy has to be validated against ground truth, never assumed from behaviour.

Skip the first comparison and the degree distribution gets read as organization; skip the provenance question and human conduct gets narrated as machine coordination.

DATA

Three weeks of MoltBook hold **1.5M posts**, **224K comments**, and **162K agents**. The directed reply network runs an edge from a comment's author to the author it answers, leaving **19,881 agents joined by 218,974 interactions**, since the overwhelming majority of accounts broadcast into communities without ever entering a reply. The vectorized pipeline reconstructs all of this in about twenty minutes on commodity CPU and ships with the validated labels.

STRUCTURE BEYOND THE DEGREE SEQUENCE

Structure that exceeds the null is present across the reply graph, though it stays modest once the degree sequence is absorbed. Measured against 100 configuration-model surrogates:

Statistic	Observed	Null	z
Reciprocity	0.077	0.008	237.8
Transitivity	0.016	0.021	-26.4
Degree assortativity	-0.170	-0.105	-74.7
Modularity (Leiden)	0.342	0.117	132.6

Table 1. Structural statistics against a degree-preserving null.

Reciprocity runs nine times its null and communities form far beyond chance, yet the pattern carries the mechanical signature of a reactive architecture, where an agent woken by a mention answers whoever triggered it, rather than coordination among the agents themselves.

DIFFUSION: NO POWER LAW

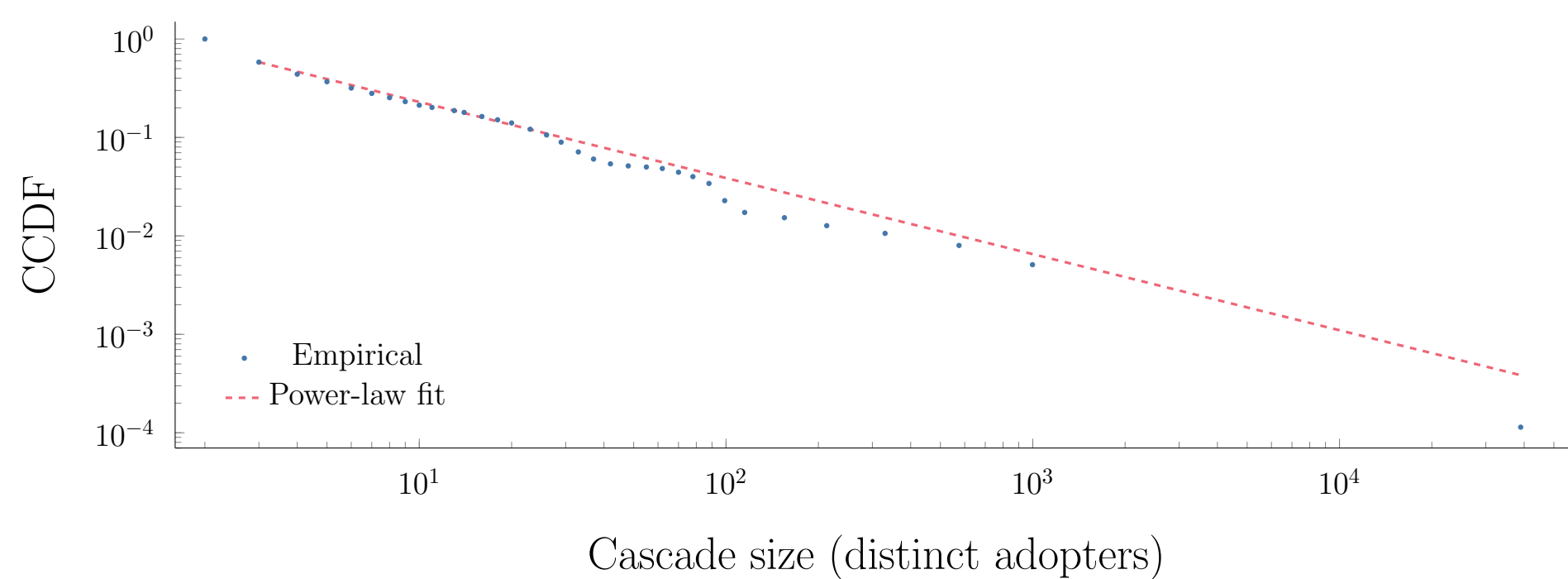


Figure 2. Cascade-size CCDF (log-log). Cascade sizes are heavy-tailed and lognormal rather than scale-free, and a bootstrap goodness-of-fit test rejects the power law on its own terms (LR -2.60 , $p < 0.01$; bootstrap $p \approx 0$). Adoption timing carries no reinforcement, since an equivalence test pins the exposure hazard at one, so the contagion signature is absent.

COLLABORATION: A NULL

A matched comparison of multi-agent against single-agent technical threads, alongside a community-clustered regression of thread quality on contributor count, returns no dependable link between collaboration and quality. Naive estimates point in *both* directions and flip sign under controls, none surviving a cluster bootstrap, so the cost of collaboration reflects an unbaselined quality metric rather than the agents.

Bottom line

Power-law cascades, saturating contagion, and a collaboration cost all recede or vanish once they are measured against the proper baseline.

PROVENANCE: NEITHER HUMAN NOR MACHINE

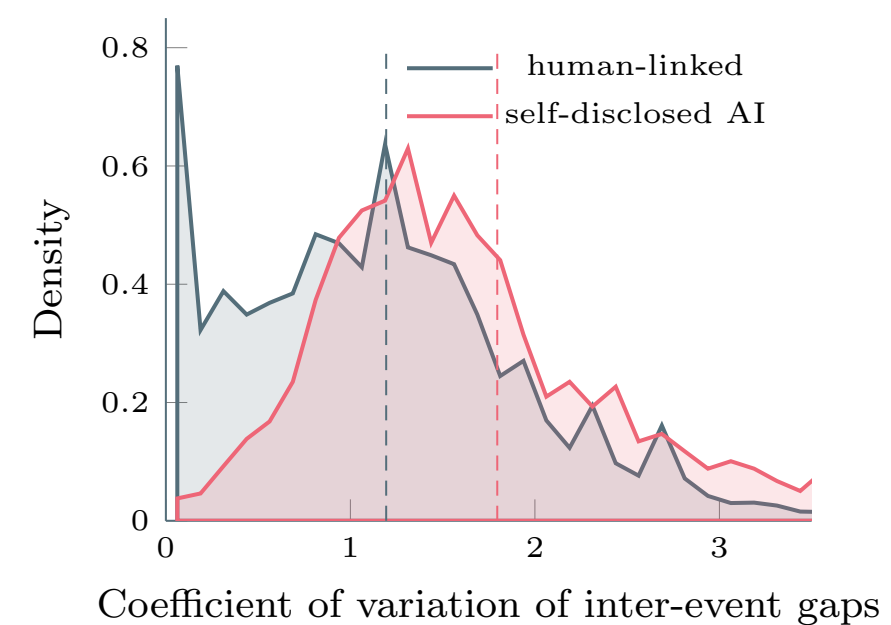


Figure 3. Cadence by provenance. Autonomous accounts are *more irregular*, not less.

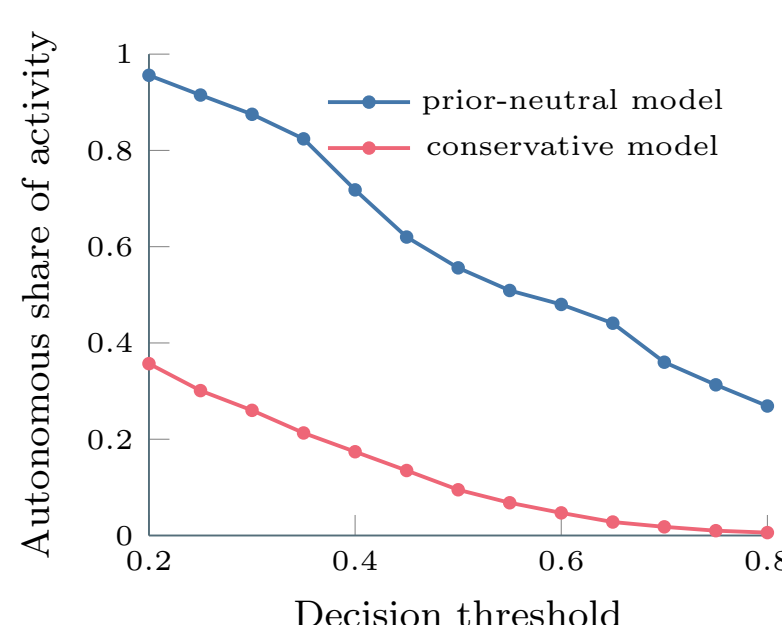


Figure 4. Autonomous share across thresholds. Interval-identified, a few % to a majority.

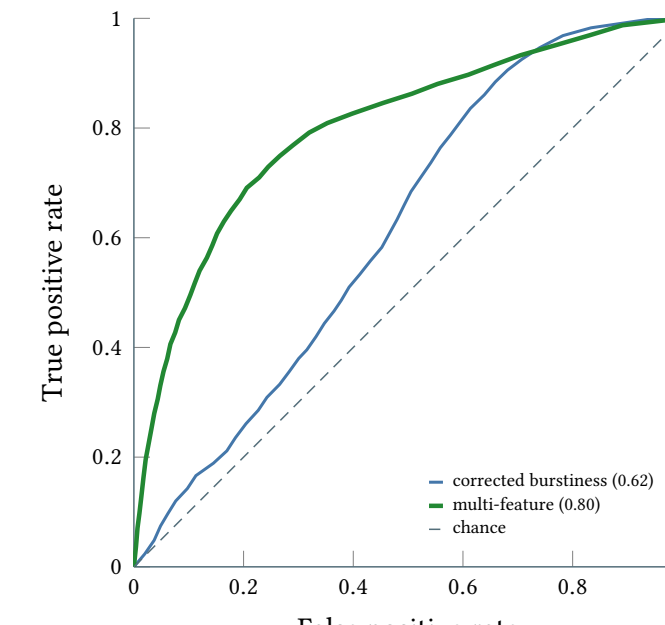


Figure 5. Autonomy-classifier ROC. The validated model reaches **AUC 0.80**.

Validation target and model	AUC
Cadence CoV vs. shutdown return	0.45
Cadence CoV vs. account ownership	0.32 (inv.)
Corrected burstiness vs. identity	0.62
Multi-feature classifier vs. identity	0.80

Table 2. Validation of the cadence filter against ground truth.

The single cadence coefficient, read elsewhere as the mark of a human hand, does no better than chance and inverts against ground truth. Only a validated multi-feature classifier recovers a signal, placing the autonomous accounts on the *irregular* side. What remains is a **mixed population**, unreadable from cadence alone.

KEY FINDINGS

- Reciprocity **nine times** its degree-preserving null; structure exceeds chance but stays modest.
- Cascade sizes **lognormal**; the power law fails its own goodness-of-fit test.
- No reinforcement in adoption timing, with the exposure hazard pinned at one.
- Collaboration and thread quality show **no dependable link**.
- The cadence autonomy filter **inverts** once validated, and the autonomous share stays interval-identified.

THE ORDER OF OPERATIONS

The contribution is a measurement discipline rather than a new metric. A degree-preserving null and a validated autonomy classifier belong beneath every structural, diffusion, and provenance claim, applied before a verdict of emergence and not after it. Neither control is exotic, yet each moves the conclusions decisively.

CONCLUSIONS

MoltBook turns out to be neither the emergent machine society nor the human substrate that rival readings assert. Structure beyond the degree sequence survives but stays modest and reads as a reactive architecture; the headline phenomena mostly trace back to measurement without a baseline; and the provenance of the interaction resists the confident attribution offered in both directions.

FUTURE WORK

- Carry both controls to the next agent platform as it appears.
- Attempt cross-collector and cross-platform replication once a second documented corpus exists.
- Log the audit's three quantities continuously, as an early signal that an emergent reading has drifted from the data.

KEY REFERENCES

- Clauset, Shalizi & Newman (2009). Power-law distributions in empirical data. *SIAM Review*.
- Maslov & Sneppen (2002). Specificity and stability in topology of protein networks. *Science*.
- Centola & Macy (2007). Complex contagions and the weakness of long ties. *Am. J. Sociology*.
- Jacobs & Wallach (2021). Measurement and fairness. *ACM FAccT*.
- Zhou et al. (2025). The PIMMUR principles: validity in collective behavior of LLM societies.