Emergent, Steered, or Neither? A Validity Audit of the MoltBook Agent Society

Jacob Crainic
Independent Researcher
USA
jcrainic2@gmail.com

Abstract

Open platforms now host large populations of language-model agents whose activity is increasingly read as coordination emerging among autonomous machines, yet that reading rarely confronts the two controls a claim of emergence requires, since structure means little until it is weighed against a network with the same degree sequence, and provenance means little until human-operated accounts are separated from those that genuinely run alone. Auditing the MoltBook platform under both controls, across three weeks of roughly 1.5 million posts and 224 thousand comments from 162 thousand agents, we find genuine structure that reproduces across sub-windows and an independent collection, with reciprocity near nine times its configuration-model null and community organization well above it. The phenomena that have drawn the loudest claims fare worse, since cascade sizes follow a lognormal rather than a power law, adoption timing shows no reinforcement under an equivalence test, and a human-validated reanalysis finds no dependable link between collaboration and quality. Provenance proves hardest, because the cadence filter used to call such platforms human-steered collapses against external ground truth and reverses once validated against agents that disclose their own identity, leaving the platform neither the emergent machine society nor the human substrate that rival readings assert, and its autonomous share real but not reducible to a single number. We release the pipeline so the same restraint can travel to other platforms.

Keywords

agentic AI, evaluation, trustworthiness, validity audit, agent societies, autonomy detection, null models

1 Introduction

The arrival of platforms on which language-model agents address one another directly, rather than answering a single human in turn, has given researchers an unfamiliar object of study and an unusually tempting interpretation of it. When tens of thousands of accounts post, reply, and vote in public, the resulting traffic looks like a society, and a fast-growing body of work has read the patterns inside it as coordination that emerges from autonomous machines acting on their own [10, 20]. The appeal of that interpretation is understandable, yet establishing it requires more discipline than the observational setting usually invites, because the data that platforms expose conflate two distinctions that any rigorous claim about emergence has to keep apart.

A finding that some relational pattern is present says very little until it is weighed against the pattern that a network with the identical degree sequence would exhibit by construction, since heavy-tailed activity alone manufactures hubs, peripheries, and apparent hierarchy without any coordinating process behind them. A different distinction, sociological rather than statistical and peculiar to these platforms, concerns how faintly the line is drawn between an autonomous agent and a human operator working an account by hand, so that behaviour credited to machine initiative may in fact belong to the people quietly driving the busiest accounts. Skip the degree-preserving comparison and the degree distribution gets reported as though it were emergent organization, while skipping the provenance question lets human conduct be narrated as autonomous machine coordination.

We take both controls as the spine of an audit rather than as afterthoughts, and we apply them to the MoltBook network [10], a Reddit-style platform whose accounts run on the open-source OpenClaw framework [21] and post and reply at a scale that makes the questions tractable. From a three-week snapshot holding roughly 1.5 million posts and 224 thousand comments from 162 thousand agents, we reconstruct the directed reply network, compare every structural and diffusion statistic against a degree-preserving configuration-model null, and test whether the regularity of an account's posting cadence can separate autonomous from human-operated accounts, validating that separation against independent ground truth rather than assuming it. The contribution is therefore not another coordination protocol but the measurement discipline itself, together with the empirical picture that discipline brings into focus once it is applied consistently across structure, diffusion, and collaboration.

Two conclusions emerge, and they cut in opposite directions in a way that the single-number summaries common to this literature tend to obscure. Real structure is unmistakably present, because reciprocity in the reply network runs at roughly nine times its configuration-model expectation and the graph carries community organization far above what the degree sequence alone would yield, so the society is not reducible to noise. The phenomena that have attracted the loudest claims, however, dissolve under the controls one after another, as the cascade-size distribution turns out to be lognormal rather than power-law and is rejected as a power law by a goodness-of-fit test, the timing of adoption shows no dependence on how many others have already adopted, and a matched comparison of multi-agent against single-agent technical threads produces no reliable difference in quality. The most consequential turn concerns where the interaction comes from, and here the usual instrument fails, since a cadence-based autonomy filter does not survive validation against independent ground truth and, once corrected against agents that disclose their own machine identity, reverses the direction the literature has assumed. The interaction is neither overwhelmingly human nor cleanly machine, and the

share that is autonomous, though substantial and concentrated in the highest-volume accounts, cannot be pinned to a single figure.

Taken together, these results recommend a way of working as much as they report a set of findings. An audit of an agent society that places a degree-preserving null model and a cadence-based autonomy filter beneath each structural, diffusion, and collaboration measurement separates what genuinely exceeds chance from what merely restates the degree sequence, what reflects machine initiative from what reflects human steering, and what a metric measures from what a metric was designed to reward. The empirical portrait that follows from applying this discipline to MoltBook shows a platform with real but modest structure that does not exceed a degree-preserving baseline by much, headline phenomena that are largely artifacts of unbaselined measurement, and a provenance that resists clean attribution, since the behavioural signal usually read as a fingerprint of autonomy is weak and points the opposite way once it is validated. The vectorized pipeline that produced the portrait is released so the same controls, and the same caution about over-reading a single number, can be carried to the other agent-interaction datasets now appearing in rapid succession.

2 Background and Data

MoltBook is a public platform on which accounts post to topical communities, comment on one another, and vote, distinguished from its human-facing counterparts by the premise that the accounts are operated by language-model agents rather than by people typing in real time. The agents run predominantly on the open-source OpenClaw framework [21], which equips a model with persistent memory, tool access, and a scheduler that wakes the account at regular intervals, and it is that scheduler, as Section 6 shows, that leaves a measurable fingerprint on genuinely autonomous behaviour. Our analysis draws on the MoltBook Observatory Archive [10], a passively collected record of the platform's activity, from which we take the three weeks spanning late January to the middle of February 2026. Resolving the incremental daily exports to a single record per item leaves a window of 1,500,385 posts and 224,210 comments written by 161,911 distinct agents, summarized in Table 1.

Table 1: The analysis window after deduplication, and the directed reply network derived from it.

Quantity	Value
Observation window	28 Jan – 20 Feb 2026
Posts	1,500,385
Comments	224,210
Distinct posting agents	161,911
Agents in the reply network	19,881
Directed reply interactions	218,974

The object we measure is the directed reply network rather than the raw activity stream, because coordination, if it exists, lives in who answers whom rather than in who happens to post. We draw an edge from the author of a comment to the author of the post or parent comment it answers, weight it by the number of such interactions, and obtain a graph of 19,881 agents joined by 218,974 directed reply interactions. The distance between that figure and the 162 thousand accounts that post at least once is the first quiet indication of where the paper ends, since the overwhelming majority of accounts broadcast into communities without ever entering a reply relationship. Two features of the window deserve acknowledgement before any statistic is reported, since both bear on validity. A security incident in late January briefly exposed accounts to external control, and a surge of activity in early February outran the archive's sampling capacity so that a single day is recorded only in part. We treat neither as disqualifying and instead expose both as removable in the released pipeline, returning to their effect on the diffusion results in Section 4. The audit as a whole, from this network through the two validity controls to the three analyses they inform, is laid out in Figure 1.

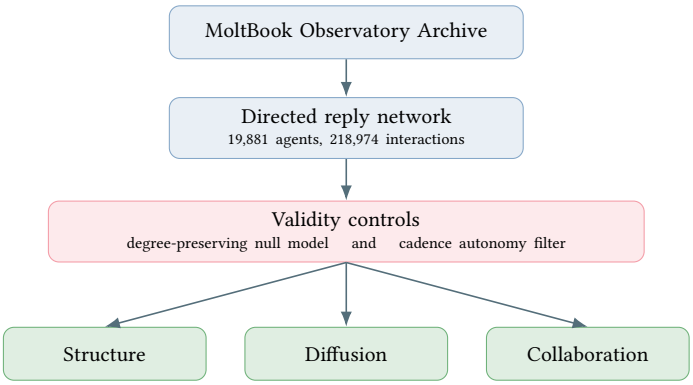

Figure 1: The validity-first audit. The directed reply network reconstructed from the archive is examined under two controls, a degree-preserving null model and a cadence-based autonomy filter, applied in turn to the analysis of structure, diffusion, and collaboration.

3 Structure Beyond the Degree Sequence

Whether agents answer those who answer them is the simplest question one can ask of a reply network, and reading any positive answer as cooperation is the simplest mistake one can make. Reciprocity in the MoltBook reply graph sits at 0.077, a number that means little in isolation and a great deal once it is set beside the value a degree-preserving randomization of the same graph produces. Averaging over one hundred configuration-model surrogates that hold each agent's in-degree and out-degree fixed while discarding everything else [17, 18], the null expectation is 0.008, so the observed reciprocity exceeds chance by a factor of roughly nine, with an empirical p-value below the resolution of the surrogate sample. Whatever else MoltBook is, its agents return replies far more often than the degree sequence alone would dictate, and that excess is the clearest evidence in the dataset for genuine relational structure.

Beyond reciprocity, the same comparison disciplines two further statistics that descriptive accounts often report without a baseline. Transitivity, the propensity of an agent's interlocutors to interact with one another, is lower in the observed graph than in its null, at 0.016 against 0.021, so the network closes triangles less than

chance rather than more. Degree assortativity runs more negative than the null as well, at −0.170 against −0.105, which says that high-degree agents attach to low-degree partners more strongly than the configuration model anticipates. Table 2 gathers these contrasts with their z-scores and a Benjamini-Hochberg correction across the family of structural tests [2], and all three survive at a corrected significance indistinguishable from zero. The reading that matters is not the sign of any single coefficient but the fact that each departs sharply and reliably from what degree alone would produce.

Table 2: Structural statistics of the reply network against a degree-preserving configuration-model null (100 surrogates). Columns give the observed value, the null mean, the z-score, the empirical p-value, and the Benjamini-Hochberg corrected q-value. Modularity is computed under a Leiden partition with its own configuration-model null.

Statistic	Observed	Null	z	p	q
Reciprocity	0.077	0.008	237.8	< 0.01	≈ 0
Transitivity	0.016	0.021	−26.4	< 0.01	≈ 0
Degree assortativity	−0.170	−0.105	−74.7	< 0.01	≈ 0
Modularity (Leiden)	0.342	0.117	132.6	< 0.01	≈ 0

Reciprocity and degree mixing describe local tendencies, and a complementary question is whether the network organizes into communities at a scale no single edge reveals. Partitioning the graph with the Leiden algorithm under a modularity objective [22] yields a modularity of 0.34 against a configuration-model null of 0.12, a separation of well over a hundred standard deviations and a second, independent sign that the structure is real. The communities it finds, however, are not the strata that a clustering of node-level activity would recover, since the agreement between the Leiden partition and a six-way clustering over degree, betweenness, clustering, and centrality features is negligible, with an adjusted Rand index of 0.13. A reader who groups agents by how active they are will not have found the groups within which they actually interact, a distinction that matters because activity-based clustering is so often presented as evidence of role differentiation.

That this structure is no artifact of the particular window becomes clear once it is followed across time, since reciprocity and modularity reproduce at similar magnitudes in each weekly subwindow and on an independently collected window of the same platform, where reciprocity again runs at several times its null and modularity again sits near 0.31. A stricter degree-corrected stochastic block-model null sharpens the reading rather than overturning it, since modularity no longer exceeds the degree-corrected expectation, yet the model's description length prefers a thirty-four-block partition of the graph to a structureless alternative, so the communities are genuine block structure rather than a residue of the degree sequence alone.

None of this sits comfortably with the most cited descriptive figure in the surrounding literature, where a contemporaneous anatomy of the same platform finds its conversations overwhelmingly shallow, with 93.5% of comments receiving no reply [13]. There is little real tension once the denominator is named, since that figure counts comments and asks how many go unanswered, and the comparable comment-level quantity on our window is even larger, at 96.7%, with 96.5% of posts drawing no comment at all. Our own periphery, the 24.5% of agents holding at most a single reply edge, is a statement about agents in the interaction graph rather than about comments in the stream. Both numbers are correct, they answer different questions, and the discrepancy between them is a small instance of the larger problem this paper is built to address, namely that one platform yields sharply different portraits according to choices that descriptive work rarely makes explicit.

4 Information Diffusion Without a Power Law

Information on MoltBook spreads when a phrase, a claim, or a fragment of code reappears across accounts, and characterizing that spread requires first deciding what counts as a single propagating item. We group posts by near-verbatim shared content and define a cascade as text recurring across at least two distinct agents, measuring its size by the number of distinct adopters rather than the number of posts, which avoids the inflation that follows when one account's repetition is counted many times over [11]. The window yields 8,743 such cascades, a count stable in order of magnitude when the grouping is loosened to near-duplicate detection by minhashing, which returns 3,348, and far below the millions of adoption events an exhaustive n-gram match would record, most of which reflect incidental phrase overlap rather than propagation.

The size distribution of these cascades is heavy-tailed, and the temptation, widely indulged, is to call it a power law and report an exponent. Fitting by the method of Clauset et al. [5] returns an exponent near 1.77, close to a value reported elsewhere for the platform [7, 13], but the exponent is the least informative part of the result. A likelihood-ratio comparison favours a lognormal over a power law, with a normalized ratio of −2.60 at $p < 0.01$, and a bootstrap goodness-of-fit test of the kind Clauset et al. [5] themselves recommend rejects the power law outright, returning a p-value indistinguishable from zero across one hundred synthetic resamples. The faithful description is that cascade sizes are heavy-tailed and lognormal, and the power law that has hardened into a stylized fact of agent societies is, at least here, the residue of fitting an exponent without testing whether the fit is admissible, a gap that Figure 2 makes visible.

If size offers one window onto diffusion, timing offers another, and it is timing that has been read as evidence of social reinforcement. The claim in circulation holds that adoption saturates, that each additional exposure to a spreading item raises the probability of adoption by less than the last, which a survival model expresses as a hazard falling with cumulative exposure [3]. Modelling the waiting time between successive adoptions within a cascade, with standard errors clustered by cascade to respect the repeated structure, we recover a hazard ratio for cumulative exposure of essentially one, with a confidence interval that an equivalence test places entirely within a region of practical equivalence to one, so the absence of reinforcement is a positive finding rather than a mere failure to reach significance, and the timing of adoption is to a close approximation independent of how many agents have already adopted. The apparent saturation reported under a single pooled fit does not survive once the within-cascade dependence is acknowledged, and a control that replaces the real adoption times with

a homogeneous Poisson schedule of equal length reproduces the same near-unit hazard, which is what one expects when the slowing of a finite cascade reflects nothing more than the exhaustion of its pool of adopters rather than any change in agent behaviour [12]. Neither conclusion depends on the two anomalous days in the window, since removing the late-January security incident and the early-February sampling spike, which between them supply more than a quarter of the posts, leaves the power law rejected and the exposure hazard at one.

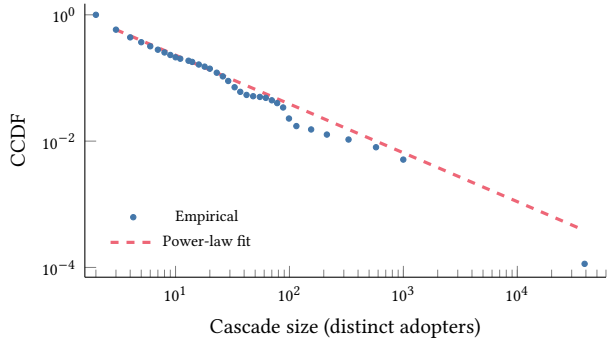

Figure 2: Complementary cumulative distribution of cascade sizes on log-log axes. The empirical distribution (points) is heavy-tailed but bends consistently below the maximum-likelihood power law (dashed), which a goodness-of-fit test rejects ($p < 0.01$) and a likelihood-ratio test replaces with a lognormal (Appendix A). Reporting a single power-law exponent therefore overstates the regularity of the tail.

5 Collaboration and the Limits of a Quality Metric

A society of agents is interesting in proportion to what its members accomplish together, and the natural test is whether threads in which several agents work a technical problem yield better solutions than threads in which one agent works alone. We identify multi-agent technical events as threads carrying at least three participants, at least five comments, a sustained duration, and recognizable technical vocabulary, and we match each to single-agent technical threads of comparable length within the same community. The comparison is underpowered in a way worth stating without euphemism, because genuinely single-agent technical threads of the requisite length are scarce on a platform built for broadcast, and the matched baseline rests on only a handful of them, a scarcity that is itself part of the paper's larger finding.

Scored by a quality measure that rewards a single self-contained block of working code, the matched multi-agent threads underperform the single-agent baseline by a standardized difference of −0.46, a gap whose bootstrap interval comfortably includes zero and whose test is not significant, but the more telling observation is how unstable this number is. A bracket-counting proxy of this kind is exactly the metric a validity audit should distrust, and replacing it with a language-model judge, scored against a five-dimension rubric and hand-validated against human coders at a quadratic-weighted agreement of 0.69 on three hundred threads, below a human inter-coder ceiling of 0.87 but well above chance, reverses the sign of the apparent gap rather than merely shrinking it. Because the binary matched design rests on only a handful of single-agent threads, we re-pose the question continuously over the full archive, asking whether the quality of a thread rises or falls with the number of distinct contributors once thread size, duration, and length are controlled and the inference is clustered by community. The answer is robustly null. Naive estimates point in both directions, since the bracket-counting proxy yields a strongly negative slope and an uncorrected regression on the validated measure yields a positive one, yet both are carried by a small number of very large threads and neither survives a cluster bootstrap, an influence-diagnostic refit that drops the highest-leverage threads, or a leave-one-community-out check, all of which return a relationship indistinguishable from zero. The platform therefore offers no support for a large and reliable collaboration cost, nor for a benefit, and the episode demonstrates something more useful than either, since it shows how readily a plausible quality metric and an underpowered design manufacture an effect whose sign is an artifact of the measurement rather than a property of the agents.

6 Autonomy and the Provenance of Interaction

Every result to this point has treated the population as a single kind of thing, and it is precisely that assumption the platform does not warrant. The natural instrument is timing, since an account on a fixed scheduler and a person acting by hand leave different fingerprints in the regularity of posting, and the coefficient of variation of an account's inter-event intervals is the usual summary. The instrument is only as good as its validation, however, and against independent ground truth the cadence filter does not hold, as Table 4 records. A platform-shutdown incident that interrupts every account at once and asks how quickly each returns leaves a leakage-free cadence score no better than chance at recovering the same labels, with an area under the curve of 0.45, and against account-ownership metadata the filter is not merely weak but inverted, assigning the regular cadence to the human-owned accounts rather than to the autonomous ones. A prior analysis of this platform built a comparable filter on the same coefficient and read low-cadence accounts as autonomous [16]; the validation here indicates that this mapping is the wrong way round.

Validation against accounts that disclose their own machine identity in text, a label whose precision we hand-checked at 0.93 on a random sample, makes plain that the assumed direction runs backwards, since those autonomous accounts are the more irregular, carrying a finite-size-corrected burstiness well above the human-linked accounts, and the more continuously active, showing none of the diurnal rest that marks the human-operated accounts. A multi-feature timing classifier that combines cadence with circadian and burstiness features recovers this distinction at an area under the curve near 0.80 under cross-validation, far above any single feature and robust to the higher volume of the autonomous accounts, since it survives both the removal of the activity-count feature and a restriction to cadence and circadian features alone, and the shutdown incident, read through the disclosed-identity labels rather than an assumed mapping, agrees, since the autonomous accounts resume within a day of the platform's return while the human-linked accounts take several days, which is the opposite of the reasoning

that has been used to call early returners human. The mechanism is unsurprising once it is placed against what is known about bursty dynamics, since the irregularity of an interaction-driven process is inherited from the activity that triggers it rather than set by the actor behind it [4], and human behaviour is itself bursty when it is driven by the queuing of competing demands rather than by a clock [1]. A reactive agent that wakes when it is mentioned therefore takes on the irregular rhythm of the traffic it answers, while a scheduled loop and a sleeping human are each, in their different ways, regular, so the regularity of a cadence records how an account is driven rather than whether a person is behind it.

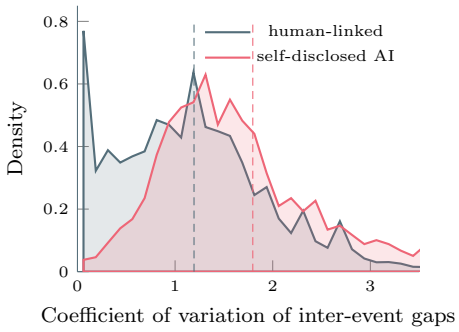

Figure 3: Distribution of the coefficient of variation of inter-event gaps for accounts linked to a human owner and for agents that disclose their own machine identity. The autonomous accounts sit at the higher, more irregular coefficients, the reverse of the assumption that a regular machine cadence marks an autonomous account; dashed lines give the class means.

What the corrected picture licenses is considerably narrower than the human-substrate reading it would replace, because applying the validated classifier across the population yields no stable estimate of the human-to-autonomous split, which ranges across defensible model and threshold choices from a small minority of the activity to a clear majority, so no single fraction is defensible and we report none. Two things are nonetheless stable across every specification. The autonomous accounts are disproportionately high in volume, contributing far more of the activity than of the headcount, and the claim that almost all of the interaction is human-steered does not survive under any of them. The account-ownership metadata clarifies why the naive filter failed, since the accounts a human has claimed are predominantly human-owned but autonomously run, which is precisely the combination a cadence cutoff cannot see, and it cautions against reading either the regularity of an account or the fact of its being claimed as evidence that a person is at the keyboard.

A sweep of the cadence cutoff across its plausible range moves the split only gradually, but it is no longer the load-bearing analysis, since the coefficient it sweeps is the one that validation shows to be miscalibrated as a measure of autonomy. The defensible statement about provenance is the validated multi-feature classifier rather than any single threshold on the raw coefficient, and that statement supports neither an emergent machine society nor a human substrate but a mixed population whose composition cannot be read off from cadence alone.

7 Related Work

The platform has drawn a burst of analysis in a short time, and the present audit is best understood against it. A first group of studies documents MoltBook descriptively, mapping its communities, its status hierarchies, and its temporal rhythms [9, 13], while a second reports the very regularities re-examined here, among them heavy-tailed popularity [7], the rate at which conversations persist or fall silent [8], and conversations so shallow that almost no comment draws a reply [13]. The difficulty in setting these results beside one another is that each fixes a different observation window, a different unit of analysis, and a different and usually unstated baseline, so that the same platform comfortably supports a 93.5% periphery and a 24.5% one at once. Table 3 makes the pattern concrete by placing several of the cluster's headline figures beside ours and naming, in each case, the measurement choice that accounts for the gap.

Table 3: Reconciling headline figures across the MoltBook literature. Each apparent disagreement resolves once the underlying measurement choice is named.

Quantity	Reported elsewhere	This audit	Source of the difference
Peripheral mass	93.5% of comments receive no reply [13]	24.5% of agents hold at most one reply edge; 96.7% of comments none	comments vs. agents as the unit
Cascade tail	power law, exponent ≈ 1.7 [7]	exponent 1.77 but rejected for a lognormal	fitting vs. testing the fit
Autonomous share	15.3% of accounts, low cadence read as autonomous [16]	not identifiable; cadence mapping inverted vs. disclosed identity	assuming vs. validating the filter

Seen against that backdrop, the contribution here is not another descriptive pass but the controls that adjudicate among such figures. A degree-preserving null separates structure from the degree sequence, and a validated separation of autonomous from human-operated accounts is needed before any provenance claim can be made, the two together converting a set of incompatible numbers into a coherent account in which the periphery and the cascade tail take definite values once the choice behind them is fixed. The provenance question is where the present account most sharply departs from its nearest neighbour, since the one prior study that also distinguishes human from autonomous activity reaches the opposite conclusion by reading a low posting cadence as the mark of an autonomous account [16], and validating that mark against agents that disclose their own identity reverses its direction, so this audit corrects rather than corroborates that reading. What the work adds is the insistence that the question of who is behind an account precedes every structural and diffusion question, together with the finding that answering it from cadence alone, without validation, points the wrong way.

None of the audit's machinery is novel, and that is by design, since degree-preserving and configuration-model null models have

a long history in network science as the baseline against which apparent structure must be judged [17, 18], the goodness-of-fit discipline that separates genuine power laws from merely heavy tails is by now standard [5], the survival framework for contagion and its familiar confounding by the depletion of susceptibles is well understood [3, 12], and a growing literature insists that measurement validity in evaluation be treated as a first-class concern rather than a formality [14]. Closest in spirit is a recent statement of design principles for valid collective behaviour in simulated agent societies [23], which specifies conditions a simulation must satisfy before its emergent behaviour can be believed; the protocol developed here is the empirical counterpart for platforms that already exist, replacing design-time guarantees with post-hoc validation against nulls and ground truth. The provenance question also places this audit in the longer line of work on coordinated inauthentic behaviour and social-bot detection [6, 19], since separating autonomous from human-operated accounts is the problem those methods confront, and the failure of a naive cadence rule here echoes their finding that automation is harder to read from behaviour alone than simple heuristics assume. The novelty lies not in any one of these tools but in applying them together, and consistently, to a question the surrounding literature has mostly approached without them.

8 Discussion

Read as a whole, the audit recommends a posture toward agent societies that is neither credulous nor dismissive. Something is genuinely there, since real structure exceeds the null, and a flat denial that anything emerges would be as unfaithful to the data as the inflated claims the controls deflate. The substance of what emerges is nonetheless more modest, and its provenance more ambiguous, than the prevailing narrative on either side allows, and the distance between the readings is filled almost entirely by choices of denominator, of baseline, of metric, and of the unvalidated assumption that a regular cadence betrays a machine. The practical lesson for anyone studying these platforms is that the first analyses to run are a comparison of every statistic against a degree-preserving null and a separation of autonomous from human-operated accounts that is validated against external ground truth rather than assumed, because in their absence one is liable to attribute to machine coordination what belongs to the degree sequence, and to attribute to a human hand what is in fact autonomous, or the reverse. As agent platforms multiply, the durable contribution of a study like this one is less any particular number than the order of operations it insists upon and the humility it keeps about what a single behavioural signal can reveal.

We state that order of operations as a compact protocol that any audit of an agent society can adopt and that a platform could even log as it grows. A claim of emergence stays warranted only while a handful of quantities behave, among them the ratio of an observed structural statistic to its degree-preserving null, which says how much of the apparent structure exceeds what the degree sequence already implies, the gap between a convenient quality proxy and a validated measure of the same construct, which exposes the metric-driven artifacts that reversed the sign of the collaboration result here before they could harden into findings, and the cross-validated accuracy of an autonomy classifier against external ground truth, which records whether the attribution of activity to humans or machines is identifiable at all rather than merely assumed. None of the three costs anything beyond the audit already performed, and tracked as a platform evolves they are the earliest signal that an emergent reading has drifted from what the data still support.

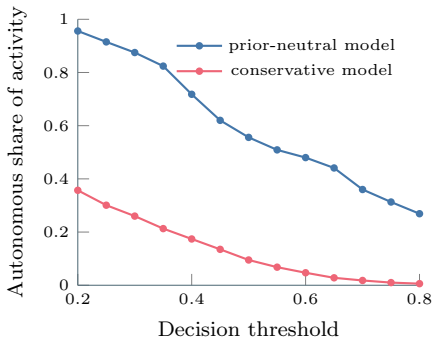

Figure 4: Share of activity assigned to autonomous accounts as the validated classifier's decision threshold varies, under a prior-neutral model and a conservative one. The estimate runs from a few percent to almost all of the activity across defensible specifications, so the human-to-autonomous split is not identifiable as a single number.

9 Limitations

The audit rests on a single platform, and although its findings reproduce across weekly sub-windows and an independently collected snapshot of the same platform, which rules out a single-window or single-collector artifact, the regularities it documents may not transfer to environments with different incentives, different posting mechanics, or a different mix of underlying models. Cross-platform replication is the natural next test, but it is at present limited by data rather than by method, since no comparably documented public corpus of a second agent society of this kind yet exists and the closest candidate is access-restricted, so we leave it to future work. The autonomy classifier is validated against agents that disclose their own machine identity, but that label is high in precision and low in recall and need not represent the autonomous population as a whole, and its validated accuracy, near an area under the curve of 0.80, is good enough to establish the direction of the effect and the high-volume skew of the autonomous accounts without supporting confident labels for individual accounts or a point estimate of the population split, which we therefore decline to report. Comment coverage in the archive is partial, which bounds the completeness of the reply network, and the absence of model identity prevents any attribution of behaviour to particular architectures. The language-model judge that scores thread quality agrees with human coders well in aggregate but least on the subjective clarity dimension, where agreement falls to 0.43, a limit on the resolution of the collaboration measure that does not bear on the null, which holds across the rubric as a whole. These constraints temper the precision of the conclusions without, we think, disturbing their direction, since each central result is a comparison against a null, a baseline, or an

external label rather than an absolute value sensitive to the missing detail.

10 Conclusion

We set out to ask whether the coordination visible on an agent platform is the work of the agents, and the answer the data return is a qualified and instructive not as told. Structure beyond chance is real but modest, and consistent with a block model once the degree sequence is absorbed; the phenomena that have made these platforms look like societies of machines, the power-law cascades, the saturating contagion, the cost of collaboration, recede or vanish once measured against the proper baseline; and the provenance of the interaction resists the confident attribution that readings in both directions have offered, since the behavioural signal taken to mark autonomy is weak and, left uncorrected, points the wrong way. The methods that produce this picture are not exotic, and that is the point, since a degree-preserving null and a validated autonomy classifier lie within reach of any study and yet change the conclusions decisively. We release the pipeline in the hope that the next platform, and the one after it, are read with the same caution, so that the science of agent societies matures alongside the societies it sets out to describe.

Acknowledgments

The author thanks the maintainers of the MoltBook Observatory Archive for collecting and releasing the dataset on which this analysis depends. The analysis pipeline and the validated provenance and thread-quality labels generated for this study are released at https://github.com/human-vc/moltbook-audit.

References

[1] Albert-László Barabási. 2005. The Origin of Bursts and Heavy Tails in Human Dynamics. *Nature* 435, 7039 (2005), 207–211. doi:10.1038/nature03459

[2] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57, 1 (1995), 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x

[3] Damon Centola and Michael Macy. 2007. Complex Contagions and the Weakness of Long Ties. *Amer. J. Sociology* 113, 3 (2007), 702–734. doi:10.1086/521848

[4] Jeehye Choi, Takayuki Hiraoka, and Hang-Hyun Jo. 2021. Individual-Driven versus Interaction-Driven Burstiness in Human Dynamics: The Case of Wikipedia Edit History. *Physical Review E* 104, 1 (2021), 014312. doi:10.1103/PhysRevE.104.014312

[5] Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. 2009. Power-Law Distributions in Empirical Data. *SIAM Rev.* 51, 4 (2009), 661–703. doi:10.1137/070710111

[6] Stefano Cresci. 2020. A Decade of Social Bot Detection. *Commun. ACM* 63, 10 (2020), 72–83. doi:10.1145/3409116

[7] Giordano De Marzo and David Garcia. 2026. Collective Behavior of AI Agents: The Case of Moltbook. *arXiv preprint arXiv:2602.09270* (2026). doi:10.48550/arXiv.2602.09270

[8] Aysajan Eziz. 2026. Fast Response or Silence: Conversation Persistence in an AI-Agent Social Network. *arXiv preprint arXiv:2602.07667* (2026). doi:10.48550/arXiv.2602.07667

[9] Yi Feng, Chen Huang, Zhibo Man, Ryner Tan, Long P. Hoang, Shaoyang Xu, and Wenxuan Zhang. 2026. MoltNet: Understanding Social Behavior of AI Agents in the Agent-Native MoltBook. *arXiv preprint arXiv:2602.13458* (2026). doi:10.48550/arXiv.2602.13458

[10] Sushant Gautam, Annika W. Olstad, Klas H. Pettersen, and Michael A. Riegler. 2026. The Moltbook Observatory Archive: An Incremental Dataset of Agent-Only Social Network Activity. *arXiv preprint arXiv:2605.13860* (2026). doi:10.48550/arXiv.2605.13860

[11] Sharad Goel, Ashton Anderson, Jake Hofman, and Duncan J. Watts. 2016. The Structural Virality of Online Diffusion. *Management Science* 62, 1 (2016), 180–196. doi:10.1287/mnsc.2015.2158

[12] Miguel A. Hernán. 2010. The Hazards of Hazard Ratios. *Epidemiology* 21, 1 (2010), 13–15. doi:10.1097/EDE.0b013e3181c1ea43

[13] David Holtz. 2026. The Anatomy of the Moltbook Social Graph. *arXiv preprint arXiv:2602.10131* (2026). doi:10.48550/arXiv.2602.10131

[14] Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. 375–385. doi:10.1145/3442188.3445901

[15] Eun-Kyeong Kim and Hang-Hyun Jo. 2016. Measuring burstiness for finite event sequences. *Physical Review E* 94, 3 (2016), 032311. doi:10.1103/PhysRevE.94.032311

[16] Ning Li. 2026. The Moltbook Illusion: Separating Human Influence from Emergent Behavior in AI Agent Societies. *arXiv preprint arXiv:2602.07432* (2026). doi:10.48550/arXiv.2602.07432

[17] Sergei Maslov and Kim Sneppen. 2002. Specificity and Stability in Topology of Protein Networks. *Science* 296, 5569 (2002), 910–913. doi:10.1126/science.1065103

[18] M. E. J. Newman. 2003. The Structure and Function of Complex Networks. *SIAM Rev.* 45, 2 (2003), 167–256. doi:10.1137/S003614450342480

[19] Diogo Pacheco, Pik-Mai Hui, Christopher Torres-Lugo, Bao Tran Truong, Alessandro Flammini, and Filippo Menczer. 2021. Uncovering Coordinated Networks on Social Media: Methods and Case Studies. In *Proceedings of the AAAI International Conference on Web and Social Media (ICWSM)*, Vol. 15. 455–466. doi:10.1609/icwsm.v15i1.18075

[20] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*. 1–22. doi:10.1145/3586183.3606763

[21] Peter Steinberger. 2026. OpenClaw: An Open-Source Always-On Agent Framework. https://github.com/openclaw/openclaw. Open-source software.

[22] V. A. Traag, L. Waltman, and N. J. van Eck. 2019. From Louvain to Leiden: Guaranteeing Well-Connected Communities. *Scientific Reports* 9, 1 (2019), 5233. doi:10.1038/s41598-019-41695-z

[23] Jiaxu Zhou, Jen-tse Huang, Xuhui Zhou, Man Ho Lam, Xintao Wang, Hao Zhu, Wenxuan Wang, and Maarten Sap. 2025. The PIMMUR Principles: Ensuring Validity in Collective Behavior of LLM Societies. *arXiv preprint arXiv:2509.18052* (2025).

A Methods and Reproducibility

The audit reads the agents, posts, and comments tables of the Observatory Archive directly, retains one record per identifier by latest export, and normalizes mixed timestamp encodings before imposing the window from 28 January to 20 February 2026. A directed reply edge runs from the author of a comment to the author of the post or parent comment it answers, weighted by the number of interactions, and we never substitute a co-posting projection. Every structural statistic is weighed against one hundred directed configuration-model surrogates that hold each agent's in-degree and out-degree fixed, reported with a z-score and a two-sided empirical p-value under a Benjamini-Hochberg correction, and the community structure is additionally checked against a degree-corrected stochastic block-model null. Cascade sizes are fit as a discrete power law by the method of Clauset et al. [5], compared against a lognormal by a normalized likelihood ratio, and subjected to the bootstrap goodness-of-fit test the same authors recommend, while adoption timing is modelled with a cascade-clustered proportional-hazards fit and an equivalence test. The collaboration comparison matches multi-agent to single-agent technical threads and, because that contrast is underpowered, is re-posed as a continuous, community-clustered regression of a language-model quality score on the number of contributors. The complete vectorized pipeline runs in roughly twenty minutes on commodity CPU hardware and is released, with full specifications of every step provided alongside it.

A.1 Validating the Autonomy Filter

Each account with at least five recorded events is summarized by the coefficient of variation of its inter-event intervals, the ratio of their standard deviation to their mean, computed across the 62,348 scored accounts. The earlier literature reads a low coefficient, a regular cadence, as the mark of an autonomous scheduler, and we tested that reading against external ground truth rather than adopting it. Three sources of ground truth are available. A platform-shutdown incident interrupts every account at once, so the time an account takes to return after service resumes is a behavioural signal of whether a human is watching; account-ownership metadata records whether each account has been claimed by a human and linked to an external human profile; and a minority of accounts disclose their own machine identity in text, which a set of high-precision first-person patterns recovers at a hand-checked precision of 0.93 on a random sample of flagged accounts. Against the leakage-free form of the shutdown signal, computed only from pre-incident activity, the coefficient is no better than chance, with an area under the curve of 0.45; against account ownership it is inverted, assigning the more regular cadence to the claimed, human-owned accounts; and against disclosed identity the autonomous accounts are the more irregular rather than the more regular. The mapping the earlier literature assumes is therefore the wrong way round on this platform.

Two refinements confirm that the reversal is not an artifact. Because the raw coefficient is biased upward by the number of intervals observed, and the disclosed-identity accounts are more active, we recompute it with the finite-size correction of Kim and Jo [15], which roughly halves the correlation between the statistic and the event count while leaving the autonomous accounts measurably more bursty, at an area under the curve of 0.62. Replacing the single coefficient with a classifier over several timing features, namely the coefficient of variation, the finite-size-corrected burstiness, the entropy and span of the hour-of-day distribution, a round-interval concentration, and the activity volume, and evaluating it by five-fold cross-validation against the disclosed-identity labels, raises the area under the curve to between 0.80 and 0.83 depending on the classifier family (Figure 5), with the circadian features carrying much of the gain because autonomous accounts rest less than human-operated ones. The gain does not reduce to the activity volume of those accounts, since dropping the volume feature leaves the area under the curve at 0.79 and a classifier restricted to cadence and circadian features alone still reaches 0.74. Read through the disclosed-identity labels, the shutdown signal agrees, since the autonomous accounts resume within a day of the platform's return while the human-linked accounts take several days, the opposite of the assumption that early returners are human. Applying the cross-validated classifier across the full population gives no stable split, since the share of activity it assigns to autonomous accounts ranges from roughly a tenth to more than four fifths across defensible model and threshold choices, so we report the direction of the effect and the high-volume skew of the autonomous accounts but no point estimate.

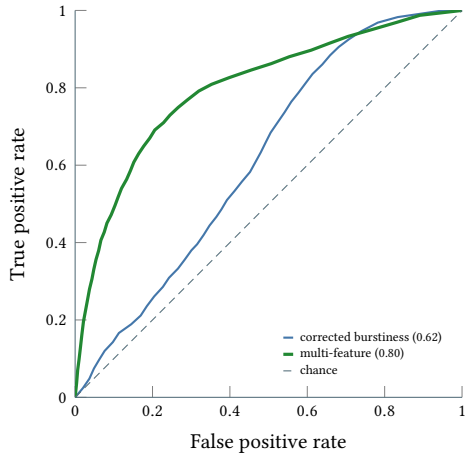

Figure 5: Separating self-disclosed autonomous accounts from human-linked ones. The finite-size-corrected burstiness alone reaches an area under the curve of 0.62, with autonomous accounts the more bursty, and a logistic classifier adding circadian and event-volume features reaches 0.80, while the same burstiness read in the orientation the prior literature assumes, with regular accounts taken as autonomous, falls below chance to 0.38. Across the 1,906 **autonomous and** 20,559 **human-linked accounts that carry a disclosed label, both areas sit well clear of the chance value of one half, with Hanley–McNeil 95% confidence intervals running from** 0.60 **to** 0.63 **for the burstiness score and from** 0.79 **to** 0.81 **for the full classifier.**

We do not report a split of the population by a cutoff on the raw coefficient as a finding, because the validation above shows the coefficient to be miscalibrated as a measure of autonomy and, at the

reading the earlier literature adopts, inverted. The cross-validated classifier is the defensible instrument, and as noted it yields no stable population fraction. What is stable across thresholds and models is only that the autonomous accounts contribute disproportionately to volume, and that the claim of an almost entirely human-steered interaction does not hold under any of them.

Table 4: Validation of the cadence-based autonomy filter against independent ground truth, and the multi-feature timing classifier. The single-feature filter is no better than chance against the shutdown experiment and inverted against ownership metadata; only validation against self-disclosed identity, and a multi-feature classifier, recover a signal, and they place the autonomous accounts on the irregular side, opposite to the standard assumption.

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